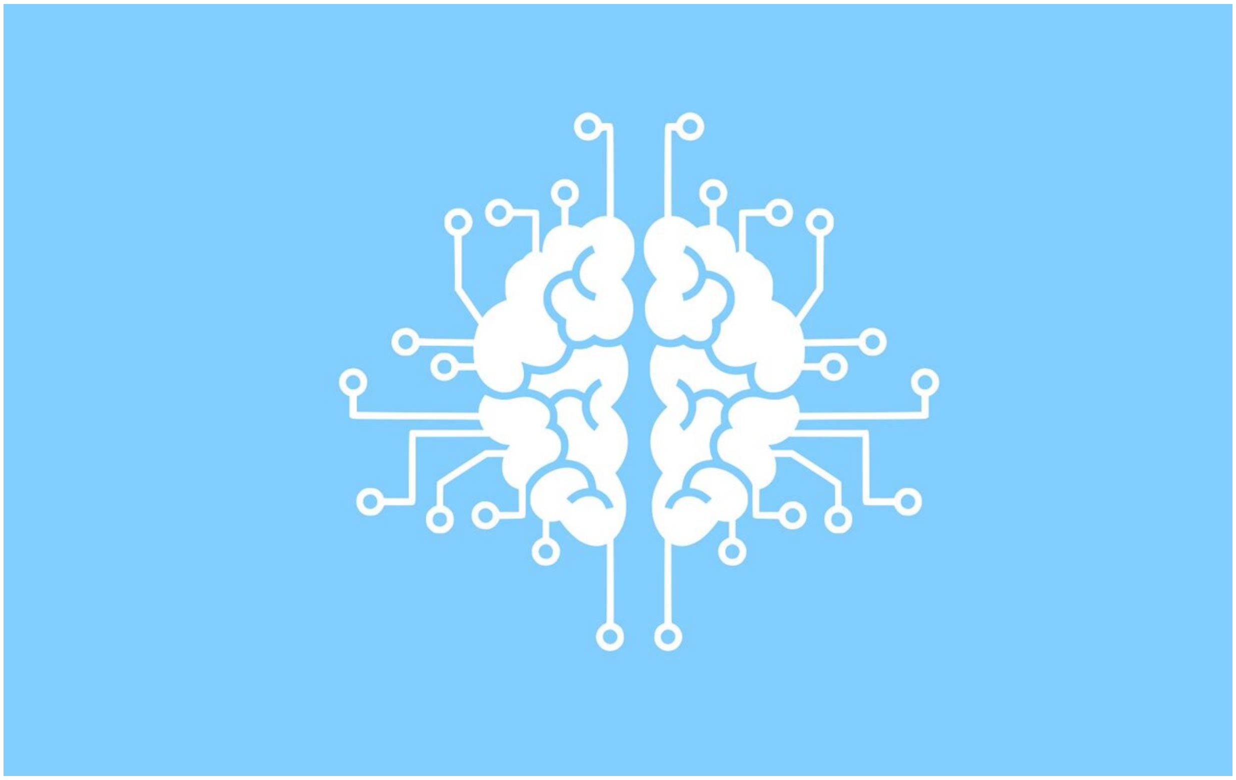
Can the bias in algorithms help us see our own?

April 9 2024, by Molly Callahan



Credit: Pixabay/CC0 Public Domain

Algorithms were supposed to make our lives easier and fairer: help us find the best job applicants, help judges impartially assess the risks of bail and bond decisions, and ensure that health care is delivered to the patients with the greatest need. By now, though, we know that algorithms can be just as biased as the human decision-makers they inform and replace.

What if that weren't a bad thing?

New research by Carey Morewedge, a Boston University Questrom School of Business professor of marketing and Everett W. Lord Distinguished Faculty Scholar, found that people recognize more of their biases in algorithms' decisions than they do in their own—even when those decisions are the same. The research, published in the *Proceedings of the National Academy of Sciences*, suggests ways that awareness might help human decision-makers recognize and correct for their biases.

"A social problem is that algorithms learn and, at scale, roll out biases in the human decisions on which they were trained," says Morewedge, who also chairs Questrom's marketing department. For example: In 2015, Amazon tested (and soon scrapped) an algorithm to help its hiring managers filter through job applicants. They found that the program boosted résumés it perceived to come from male applicants, and downgraded those from female applicants, a clear case of gender bias.

But that same year, just 39 percent of Amazon's workforce were women. If the algorithm had been trained on Amazon's existing hiring data, it's no wonder it prioritized male applicants—Amazon already was. If its algorithm had a gender bias, "it's because Amazon's managers were biased in their hiring decisions," Morewedge says.

"Algorithms can codify and amplify human bias, but algorithms also reveal structural biases in our society," he says. "Many biases cannot be observed at an individual level. It's hard to prove bias, for instance, in a single hiring decision. But when we add up decisions within and across persons, as we do when building algorithms, it can reveal structural biases in our systems and organizations."

Morewedge and his collaborators—Begüm Çeliktutan and Romain Cadario, both at Erasmus University in the Netherlands—devised a series of experiments designed to tease out people's social biases (including racism, sexism, and ageism).

The team then compared research participants' recognition of how those biases colored their own decisions versus decisions made by an algorithm. In the experiments, participants sometimes saw the decisions of real algorithms. But there was a catch: other times, the decisions attributed to algorithms were actually the participants' choices, in disguise.

Across the board, participants were more likely to see bias in the decisions they thought came from algorithms than in their own decisions. Participants also saw as much bias in the decisions of algorithms as they did in the decisions of other people. (People generally better recognize bias in others than in themselves, a phenomenon called the bias blind spot.) Participants were also more likely to correct for bias in those decisions after the fact, a crucial step for minimizing bias in the future.

Algorithms remove the bias blind spot

The researchers ran sets of participants, more than 6,000 in total, through nine experiments. In the first, participants rated a set of Airbnb listings, which included a few pieces of information about each listing: its average star rating (on a scale of 1 to 5) and the host's name. The researchers assigned these fictional listings to hosts with names that were "distinctively African American or white," based on previous research identifying racial bias, according to the paper. The participants rated how likely they were to rent each listing.

In the second half of the experiment, participants were told about a research finding that explained how the host's race might bias the ratings. Then, the researchers showed participants a set of ratings and asked them to assess (on a scale of 1 to 7) how likely it was that bias had influenced the ratings.

Participants saw either their own rating reflected back to them, their own rating under the guise of an algorithm's, their own rating under the guise of someone else's, or an actual algorithm rating based on their preferences.

The researchers repeated this setup several times, testing for race, gender, age, and attractiveness bias in the profiles of Lyft drivers and Airbnb hosts. Each time, the results were consistent. Participants who thought they saw an algorithm's ratings or someone else's ratings (whether or not they actually were) were more likely to perceive bias in the results.

Morewedge attributes this to the different evidence we use to assess bias in others and bias in ourselves. Since we have insight into our own thought process, he says, we're more likely to trace back through our thinking and decide that it wasn't biased, perhaps driven by some other factor that went into our decisions. When analyzing the decisions of other people, however, all we have to judge is the outcome.

"Let's say you're organizing a panel of speakers for an event," Morewedge says. "If all those speakers are men, you might say that the outcome wasn't the result of gender bias because you weren't even thinking about gender when you invited these speakers. But if you were attending this event and saw a panel of all-male speakers, you're more likely to conclude that there was gender bias in the selection."

Indeed, in one of their experiments, the researchers found that participants who were more prone to this bias blind spot were also more likely to see bias in decisions attributed to algorithms or others than in their own decisions. In another experiment, they discovered that people more easily saw their own decisions influenced by factors that were fairly neutral or reasonable, such as an Airbnb host's star rating, compared to a prejudicial bias, such as race—perhaps because admitting to preferring a five-star rental isn't as threatening to one's sense of self or how others might view us, Morewedge suggests.

Algorithms as mirrors: Seeing and correcting human bias

In the researchers' final experiment, they gave participants a chance to correct bias in either their ratings or the ratings of an algorithm (real or not). People were more likely to correct the algorithm's decisions, which reduced the actual bias in its ratings.

This is the crucial step for Morewedge and his colleagues, he says. For anyone motivated to reduce bias, being able to see it is the first step. Their research presents evidence that algorithms can be used as mirrors—a way to identify bias even when people can't see it in themselves.

"Right now, I think the literature on algorithmic bias is bleak," Morewedge says. "A lot of it says that we need to develop statistical methods to reduce prejudice in algorithms. But part of the problem is that prejudice comes from people. We should work to make algorithms better, but we should also work to make ourselves less biased.

"What's exciting about this work is that it shows that algorithms can codify or amplify human bias, but algorithms can also be tools to help people better see their own biases and correct them," he says. "Algorithms are a double-edged sword. They can be a tool that amplifies our worst tendencies. And algorithms can be a tool that can help better ourselves."

More information: Carey K. Morewedge et al, People see more of their biases in algorithms, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2317602121. doi.org/10.1073/pnas.2317602121

Provided by Boston University

Citation: Can the bias in algorithms help us see our own? (2024, April 9) retrieved 21 May 2024 from https://phys.org/news/2024-04-bias-algorithms.html