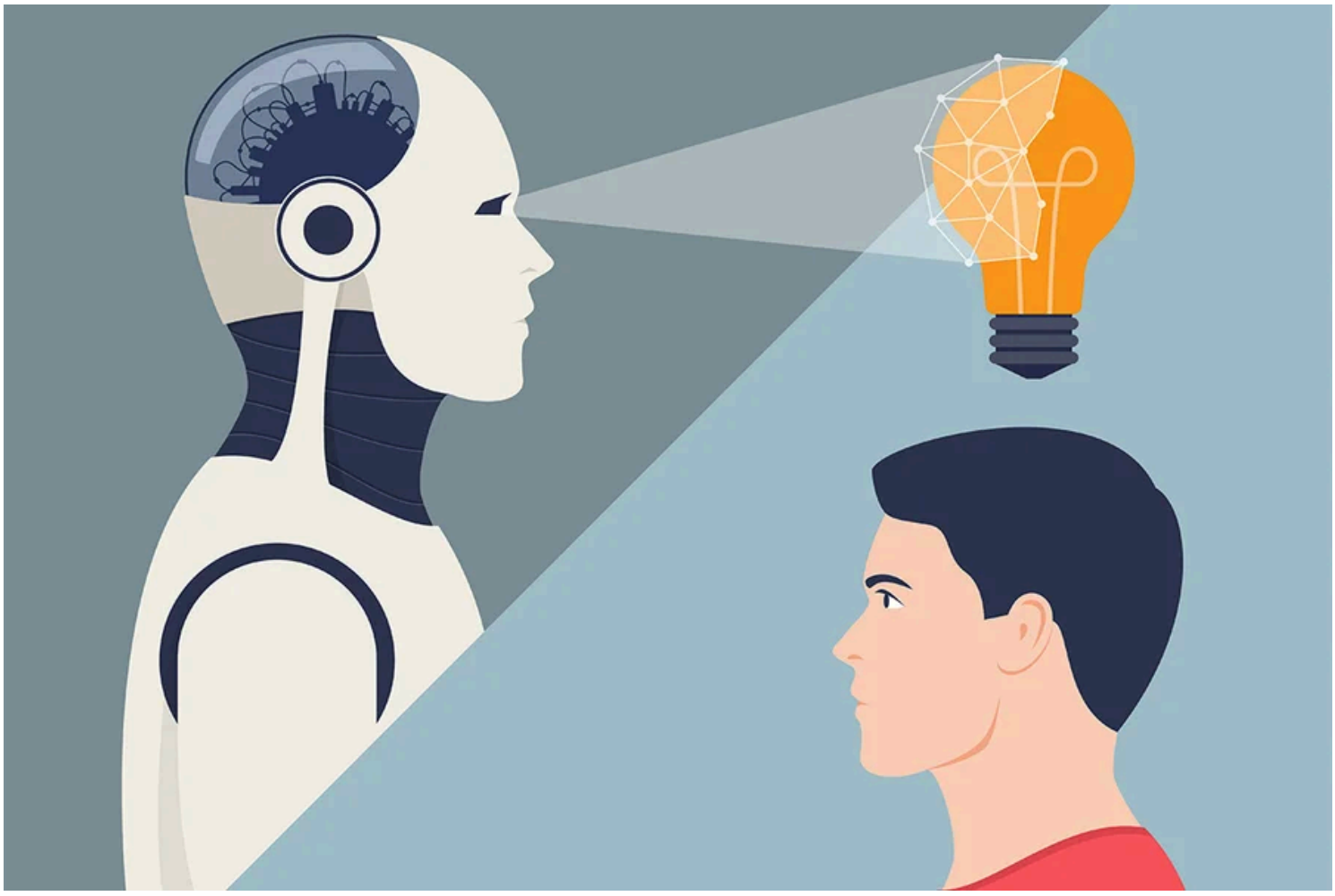
OCTOBER 26, 2023 | 6 MIN READ

Humans Absorb Bias from AI—And Keep It after They Stop Using the Algorithm

People may learn from and replicate the skewed perspective of an artificial intelligence algorithm, and they carry this bias beyond their interactions with the AI

BY LAUREN LEFFER



Maksim Akhramenka/Getty Images

Artificial Intelligence

Artificial intelligence programs, like the humans who develop and train them, are far from perfect. Whether it's machine-learning software that analyzes medical images or a generative chatbot, such as ChatGPT, that holds a seemingly organic conversation, algorithm-based technology can make errors and even "hallucinate," or provide inaccurate information. Perhaps more insidiously, AI can also display biases that get introduced through the massive data troves that these programs are trained on—and that are indetectable to many users. Now new research suggests human users may unconsciously absorb these automated biases.

Past studies have demonstrated that biased AI can harm people in already marginalized groups. Some impacts are subtle, such as speech recognition software's inability to understand non-American accents, which might inconvenience people using smartphones or voice-operated home assistants. Then there are scarier examples—including health care algorithms that make errors because they're only trained on a subset of people (such as white people, those of a specific age range or even people with a certain stage of a disease), as well as racially biased police facial recognition software that could increase wrongful arrests of Black people.

Yet solving the problem may not be as simple as retroactively adjusting algorithms. Once an AI model is out there, influencing people with its bias, the damage is, in a sense, already done. That's because people who interact with these automated systems could be unconsciously incorporating the skew they encounter into their own future decision-making, as suggested by a recent psychology study published in *Scientific Reports*. Crucially, the study demonstrates that bias introduced to a user by an AI model can persist in a person's behavior—even after they stop using the AI program.

“We already know that artificial intelligence inherits biases from humans,” says the new study’s senior researcher Helena Matute, an experimental psychologist at the University of Deusto in Spain. For example, when the technology publication *Rest of World* recently analyzed popular AI image generators, it found that these programs tended toward ethnic and national stereotypes. But Matute seeks to understand AI-human interactions in the other direction. “The question that we are asking in our laboratory is how artificial intelligence can influence human decisions,” she says.

Over the course of three experiments, each involving about 200 unique participants, Matute and her co-researcher, Lucía Vicente of the University of Deusto, simulated a simplified medical diagnostic task: they asked the nonexpert participants to categorize images as indicating the presence or absence of a fictional disease. The images were composed of dots of two different colors, and participants were told that these dot arrays represented tissue samples. According to the task parameters, more dots of one color meant a positive result for the illness, whereas more dots of the other color meant that it was negative.

Throughout the different experiments and trials, Matute and Vicente offered subsets of the participants purposefully skewed suggestions that, if followed, would lead them to classify images incorrectly. The scientists described these suggestions as originating from a “diagnostic assistance system based on an artificial intelligence (AI) algorithm,” they explained in an email. The control group received a series of unlabeled dot images to assess. In contrast, the experimental groups received a series of dot images labeled with “positive” or “negative” assessments from the fake AI. In most instances, the label was correct, but in cases where the number of dots of each color was similar, the researchers introduced intentional skew with incorrect answers. In one

experimental group, the AI labels tended toward offering false negatives. In a second experimental group, the slant was reversed toward false positives.

The researchers found that the participants who received the fake AI suggestions went on to incorporate the same bias into their future decisions, even after the guidance was no longer offered. For example, if a participant interacted with the false positive suggestions, they tended to continue to make false positive errors when given new images to assess. This observation held true despite the fact that the control groups demonstrated the task was easy to complete correctly without the AI guidance—and despite 80 percent of participants in one of the experiments noticing that the fictional "AI" made mistakes.

A big caveat is that the study did not involve trained medical professionals or assess any approved diagnostic software, says Joseph Kvedar, a professor of dermatology at Harvard Medical School and editor in chief of *npj Digital Medicine*. Therefore, Kvedar notes, the study has very limited implications for physicians and the actual AI tools that they use. Keith Dreyer, chief science officer of the American College of Radiology Data Science Institute, agrees and adds that "the premise is not consistent with medical imaging."

Though not a true medical study, the research offers insight into how people might learn from the biased patterns inadvertently baked into many machine-learning algorithms—and it suggests that AI could influence human behavior for the worse. Ignoring the diagnostic aspect of the fake AI in the study, Kvedar says, the "design of the experiments was almost flawless" from a psychological point of view. Both Dreyer and Kvedar, neither of whom were involved in the study, describe the work as interesting, albeit not surprising.

There's "real novelty" in the finding that humans might continue to enact an AI's bias by replicating it beyond the scope of their interactions with a machine-learning model, says Lisa Fazio, an associate professor of psychology and human development at Vanderbilt University, who was not involved in the recent study. To her, it suggests that even time-limited interactions with problematic AI models or AI-generated outputs can have lasting effects.

Consider, for example, the predictive policing software that Santa Cruz, Calif., banned in 2020. Though the city's police department no longer uses the algorithmic tool to determine where to deploy officers, it's possible that—after years of use—department officials internalized the software's likely bias, says Celeste Kidd, an assistant professor of psychology at the University of California, Berkeley, who was also not involved in the new study.

It's widely understood that people learn bias from human sources of information as well. The consequences when inaccurate content or guidance originate from artificial intelligence could be even more severe, however, Kidd says. She has previously studied and written about the unique ways that AI can shift human beliefs. For one, Kidd points out that AI models can easily become even more skewed than humans are. She cites a recent assessment published by Bloomberg that determined that generative AI may display stronger racial and gender biases than people do.

There's also the risk that humans might ascribe more objectivity to machine-learning tools than to other sources. "The degree to which you are influenced by an information source is related to how intelligent you assess it to be," Kidd says. People may attribute more authority to AI, she explains, in part because algorithms are often marketed as drawing on the sum of all human knowledge. The new study seems to back this idea up in a secondary finding: Matute and

Vicente noted that that participants who self-reported higher levels of trust in automation tended to make more mistakes that mimicked the fake AI's bias.

Plus, unlike humans, algorithms deliver all outputs—whether correct or not—with seeming "confidence," Kidd says. In direct human communication, subtle cues of uncertainty are important for how we understand and contextualize information. A long pause, an "um," a hand gesture or a shift of the eyes might signal a person isn't quite positive about what they're saying. Machines offer no such indicators. "This is a huge problem," Kidd says. She notes that some AI developers are attempting to retroactively address the issue by adding in uncertainty signals, but it's difficult to engineer a substitute for the real thing.

Kidd and Matute both claim that a lack of transparency from AI developers on how their tools are trained and built makes it additionally difficult to weed out AI bias. Dreyer agrees, noting that transparency is a problem, even among approved medical AI tools. Though the Food and Drug Administration regulates diagnostic machine-learning programs, there is no uniform federal requirement for data disclosures. The American College of Radiology has been advocating for increased transparency for years and says more work is still necessary. "We need physicians to understand at a high level how these tools work, how they were developed, the characteristics of the training data, how they perform, how they should be used, when they should not be used, and the limitations of the tool," reads a 2021 article posted on the radiology society's website.

And it's not just doctors. In order to minimize the impacts of AI bias, everyone "needs to have a lot more knowledge of how these AI systems work," Matute says. Otherwise we run the risk of letting algorithmic "black boxes" propel us into a self-defeating cycle in which AI leads to more-biased humans, who in

turn create increasingly biased algorithms. “I’m very worried,” Matute adds, “that we are starting a loop, which will be very difficult to get out of.”

RIGHTS & PERMISSIONS

LAUREN LEFFER is a contributing writer and former tech reporting fellow at *Scientific American*. She covers many subjects, including artificial intelligence, climate and weird biology, because she's curious to a fault. Follow her on X @lauren_leffer and on Bluesky @laurenleffer.bsky.social

More by Lauren Leffer

Popular Stories

OPINION | MAY 25, 2024

How the Guinness Brewery Invented the Most Important Statistical Method in Science

The most common test of statistical significance originated from the Guinness brewery. Here’s how it works

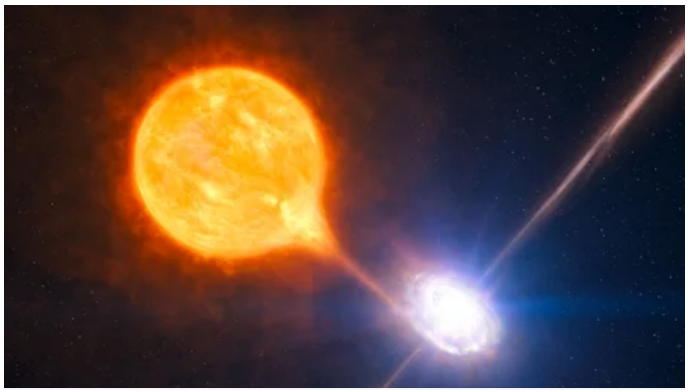
JACK MURTAGH

PLANETARY SCIENCE | MAY 22, 2024

The Great Solar Storm of 2024 May Have Made the Strongest Auroras in Centuries

Northern and southern lights produced by a recent bout of severe space weather may rival the most intense auroras of the past 500 years

SHARMILA KUTHUNUR, SPACE.COM



BLACK HOLES | MAY 23, 2024

We've Finally Seen Matter Plunge into a Black Hole

For the first time, scientists observed matter's free fall into a black hole's "plunging region"

ALLISON PARSHALL

ENERGY | JULY 14, 2016

New Concentrating Solar Tower Is Worth Its Salt with 24/7 Power

A California firm is converting sunlight to heat and storing it in molten salt so it can supply electricity when the wind is calm or the sun isn't shining

KNVUL SHEIKH

WEATHER | MAY 24, 2024

The 2024 Hurricane Season Could Be a Dangerous One

The National Hurricane Center's hurricane season outlook for the Atlantic Ocean forecasts 17 to 25 named storms in 2024 because of an expected combination of warm ocean temperatures and a La Niña climate pattern

JHORDANNE JONES, THE CONVERSATION US

OPINION | MAY 23, 2024

The Emerging Artificial Intelligence Era Faces a Growing Threat from Directed Energy Weapons

Autonomous and AI-enabled systems increasingly rely on optical and radio frequency sensors and significant computer power. They face growing vulnerabilities from directed-energy laser and microwave weapons

DAVID C. STOUDT