vox.com

Why algorithms can be racist and sexist

Rebecca Heilweil

14–18 minutes

Humans are error-prone and biased, but that doesn't mean that algorithms are necessarily better. Still, the tech is already making important decisions about your life and potentially ruling over which political advertisements you see, how your application to your dream job is screened, how police officers are deployed in your neighborhood, and even predicting your home's risk of fire.

But these systems can be biased based on who builds them, how they're developed, and how they're ultimately used. This is commonly known as algorithmic bias. It's tough to figure out exactly how systems might be susceptible to algorithmic bias, especially since this technology often operates in a corporate black box. We frequently don't know how a particular artificial intelligence or algorithm was designed, what data helped build it, or how it works.

Typically, you only know the end result: how it has affected you, if you're even aware that AI or an algorithm was used in the first place. *Did you get the job? Did you see that Donald Trump ad on your Facebook timeline? Did a facial recognition system identify you?* That makes addressing the biases of artificial intelligence tricky, but even more important to understand.

Machine learning-based systems are trained on data. Lots of it.

When thinking about "machine learning" tools (machine learning is a type of artificial intelligence), it's better to think about the idea of "training." This involves exposing a computer to a bunch of data — any kind of data — and then that computer learns to make judgments, or predictions, about the information it processes based on the patterns it notices.

For instance, in a very simplified example, let's say you wanted to train your computer system to recognize whether an object is a book, based on a few factors, like its texture, weight, and dimensions. A human might be able to do this, but a computer could do it more quickly.

To train the system, you show the computer metrics attributed to a lot of different objects. You give the computer system the metrics for every object, and tell the computer when the objects are books and when they're not. After continuously testing and refining, the system is supposed to learn what indicates a book and, hopefully, be able to predict in the future whether an object is a book, depending on those metrics, without human assistance.

That sounds relatively straightforward. And it might be, if your first batch of data was classified correctly and included a good range of metrics featuring lots of different types of books. However, these systems are often applied to situations that have much more serious consequences than this task, and in scenarios where there isn't necessarily an "objective" answer. Often, the data on which many of these decision-making systems are trained or checked are often not complete, balanced, or selected appropriately, and that can be a major source of — although certainly not the only source — of algorithmic bias.

Nicol Turner-Lee, a Center for Technology Innovation fellow at the Brookings Institution think tank, explains that we can think about algorithmic bias in two primary ways: accuracy and impact. An AI can have different accuracy rates for different demographic groups. Similarly, an algorithm can make vastly different decisions when applied to different populations.

Importantly, when you think of data, you might think of formal studies in which demographics and representation are carefully considered, limitations are weighed, and then the results are peer-reviewed. That's not necessarily the case with the AI-based systems that might be used to make a decision about you. Let's take one source of data everyone has access to: the internet. One study found that, by teaching an artificial intelligence to crawl through the internet — and just reading what humans have already written — the system would produce prejudices against black people and women.

Another example of how training data can produce sexism in an algorithm occurred a few years ago, when Amazon tried to use AI to build a résumé-screening tool. According to Reuters, the company's hope was that technology could make the process of sorting through job applications more efficient. It built a screening algorithm using résumés the company had collected for a decade, but those résumés tended to come from men. That meant the system, in the end, learned to discriminate against women. It also ended up factoring in proxies for gender, like whether an applicant went to a women's college. (Amazon says the tool was never used and that it was nonfunctional for several reasons.)

Amid discussions of algorithmic biases, companies using AI might say they're taking precautions, taking steps to use more representative training data and regularly auditing their systems for unintended bias and disparate impact against certain groups. But Lily Hu, a doctoral candidate at Harvard in applied mathematics and philosophy who studies AI fairness, says those aren't assurances that your system will perform fairly in the future.

"You don't have any guarantees because your algorithm performs 'fairly' on your old dataset," Hu told Recode. "That's just a fundamental problem of machine learning. Machine learning works on old data [and] on training data. And it doesn't work on new data, because we haven't collected that data yet."

Still, shouldn't we just make more representative datasets? That might be part of the solution, though it's worth noting that not all efforts aimed at building better data sets are ethical. And it's not just about the data. As Karen Hao of the MIT Tech Review explains, AI could also be designed to frame a problem in a fundamentally problematic way. For instance, an algorithm designed to determine "creditworthiness" that's programmed to maximize profit could ultimately decide to give out predatory, subprime loans.

Here's another thing to keep in mind: Just because a tool is tested for bias — which assumes that engineers who are checking for bias actually understand how bias manifests and operates — against one group doesn't mean it is tested for bias against another type of group. This is also true when an algorithm is considering several types of identity factors at the same time: A tool may deemed fairly accurate on white women, for instance, but that doesn't necessarily mean it works with black women.

In some cases, it might be impossible to find training data free of bias. Take historical data produced by the United States criminal justice system. It's hard to imagine that data produced by an institution rife with systemic racism could be used to build out an effective and fair tool. As researchers at New York University and the AI Now Institute outline, predictive policing tools can be fed "dirty data," including policing patterns that reflect police departments' conscious and implicit biases, as well as police corruption.

The foundational assumptions of engineers can also be biased

So you might have the data to build an algorithm. But who designs it, and who decides how it's deployed? Who gets to decide what level of accuracy and inaccuracy for different groups is acceptable? Who gets to decide which applications of AI are ethical and which aren't?

While there isn't a wide range of studies on the demographics of the artificial intelligence field, we do know that AI tends to be dominated by men. And the "high tech" sector, more broadly, tends to overrepresent white people and underrepresent black and Latinx people, according to the Equal Employment Opportunity Commission.

Turner-Lee emphasizes that we need to think about who gets a seat at the table when these systems are proposed, since those people ultimately shape the discussion about ethical deployments of their technology.

But there's also a broader question of what questions artificial intelligence can help us answer. Hu, the Harvard researcher, argues that for many systems, the question of building a "fair" system is essentially nonsensical, because those systems try to answer social questions that don't necessarily have an objective answer. For instance, Hu says algorithms that claim to predict a person's recidivism don't ultimately address the ethical question of whether someone deserves parole.

"There's not an objective way to answer that question," Hu says. "When you then insert an AI system, an algorithmic system, [or] a computer, that doesn't change the fundamental context of the problem, which is that the problem has no objective answer. It's fundamentally a question of what our values are, and what the purpose of the criminal justice system is."

That in mind, some algorithms probably shouldn't exist, or at least they shouldn't come with such a high risk of abuse. Just because a technology is accurate doesn't make it fair or ethical. For instance, the Chinese government has used artificial intelligence to track and racially profile its largely Muslim Uighur minority, about 1 million of whom are believed to be living in internment camps.

Transparency is a first step for accountability

One of the reasons algorithmic bias can seem so opaque is because, on our own, we usually can't tell when it's happening (or if an algorithm is even in the mix). That was one of the reasons why the controversy over a husband and wife who both applied for an Apple Card — and got widely different credit limits — attracted so much attention, Turner-Lee says. It was a rare instance in which two people, who at least appeared to be exposed to the same algorithm and could easily compare notes. The details of this case still aren't clear, though the company's credit card is now being investigated by regulators.

But consumers being able to make apples-to-apples comparisons of algorithmic results are rare, and that's part of why advocates are demanding more transparency about how systems work and their accuracy. Ultimately, it's probably not a problem we can solve on the individual level. Even if we do understand that algorithms can be biased, that doesn't mean companies will be forthright in allowing outsiders to study their artificial intelligence. That's created a challenge for those pushing for more equitable, technological systems. How can you critique an algorithm — a sort of black box — if you don't have true access to its inner workings or the capacity to test a good number of its decisions?

Companies will claim to be accurate, overall, but won't always reveal their training data (remember, that's the data that the artificial intelligence trains on before evaluating new data, like, say, your job application). Many don't appear to be subjecting themselves to audit by a third-party evaluator or publicly sharing how their systems fare when applied to different demographic groups. Some researchers, such as Joy Buolamwini and Timnit Gebru, say that sharing this demographic information about both the data used to train and the data used to check artificial intelligence should be a baseline definition of transparency.

Artificial intelligence is new, but that doesn't mean existing laws don't apply

We will likely need new laws to regulate artificial intelligence, and some lawmakers are catching up on the issue. There's a bill that would force companies to check their AI systems for bias through the Federal Trade Commission (FTC). And legislation has also been proposed to regulate facial recognition, and even to ban the technology from federally assisted public housing.

But Turner-Lee emphasizes that new legislation doesn't mean existing laws or agencies don't have the power to look over these tools, even if there's some uncertainty. For instance, the FTC oversees deceptive acts and practices, which could give the agency authority over some AI-based tools.

The Equal Employment Opportunity Commission, which investigates employment discrimination, is reportedly looking into at least two cases involving algorithmic discrimination. At the same time, the White House is encouraging federal agencies that are figuring out how to regulate artificial intelligence to keep technological innovation in mind. That raises the challenge of whether the government is prepared to study and govern this technology, and figure out how existing laws apply.

“You have a group of people that really understand it very well, and that would be technologists,” Turner-Lee cautions, “and a group of people who don’t really understand it at all, or have minimal understanding, and that would be policymakers.”

That’s not to say there aren’t technical efforts to “de-bias” flawed artificial intelligence, but it’s important to keep in mind that the technology won’t be a solution to fundamental challenges of fairness and discrimination. And, as the examples we’ve gone through indicate, there’s no guarantee companies building or using this tech will make sure it’s not discriminatory, especially without a legal mandate to do so. It would seem it’s up to us, collectively, to push the government to rein in the tech and to make sure it helps us more than it might already be harming us.

Open Sourced is made possible by Omidyar Network. All Open Sourced content is editorially independent and produced by our journalists.

contributions to the Vox Contributions program before the end of the year, which in turn helps us keep this work free. We need to add 2,500 contributions this month to hit that goal. Will you make a contribution today to help us hit this goal and support our policy coverage? Any amount helps.

$5/month

$10/month

$25/month

$50/month

Other

Yes, I'll give $5/month

Yes, I'll give $5/month

We accept credit card, Apple Pay, and Google Pay. You can also contribute via

PayPal