

POLICY

This is how AI bias really happens—and why it's so hard to fix

Bias can creep in at many stages of the deep-learning process, and the standard practices in computer science aren't designed to detect it.

By Karen Hao

February 4, 2019



MS. TECH; PHOTO: PIXOLOGICSTUDIO/SCIENCE PHOTO LIBRARY

Over the past few months, we've documented how the vast majority of AI's applications today are based on the category of algorithms known as deep learning, and how deep-learning algorithms find patterns in data. We've also covered how these technologies affect people's lives: how they can perpetuate injustice in hiring, retail, and security and may already be doing so in the criminal legal system.

But it's not enough just to know that this bias exists. If we want to be able to fix it, we need to understand the mechanics of how it arises in the first place.

How AI bias happens

We often shorthand our explanation of AI bias by blaming it on biased training data. The reality is more nuanced: bias can creep in long before the data is collected as well as at many other stages of the deep-learning process. For the purposes of this discussion, we'll focus on three key stages.

🔥 Meet the 35 Innovators Under 35, plus save 25% and get a free gift when you subscribe today.

Framing the problem. The first thing computer scientists do when they create a deep-learning model is decide what they actually want it to achieve. A credit card company, for example, might want to predict a customer's creditworthiness, but "creditworthiness" is a rather nebulous concept. In order to translate it into something that can be computed, the company must decide whether it wants to, say, maximize its profit margins or maximize the number of loans that get repaid. It could then define creditworthiness within the context of that goal. The problem is that "those decisions are made for

various business reasons other than fairness or discrimination,” explains Solon Barocas, an assistant professor at Cornell University who specializes in fairness in machine learning. If the algorithm discovered that giving out subprime loans was an effective way to maximize profit, it would end up engaging in predatory behavior even if that wasn’t the company’s intention.

Collecting the data. There are two main ways that bias shows up in training data: either the data you collect is unrepresentative of reality, or it reflects existing prejudices. The first case might occur, for example, if a deep-learning algorithm is fed more photos of light-skinned faces than dark-skinned faces. The resulting face recognition system would inevitably be worse at recognizing darker-skinned faces. The second case is precisely what happened when Amazon discovered that its internal recruiting tool was dismissing female candidates. Because it was trained on historical hiring decisions, which favored men over women, it learned to do the same.

Preparing the data. Finally, it is possible to introduce bias during the data preparation stage, which involves selecting which attributes you want the algorithm to consider. (This is not to be confused with the problem-framing stage. You can use the same attributes to train a model for very different goals or use very different attributes to train a model for the same goal.) In the case of modeling creditworthiness, an “attribute” could be the customer’s age, income, or number of paid-off loans. In the case of Amazon’s recruiting tool, an “attribute” could be the candidate’s gender, education level, or years of experience. This is what people often call the “art” of deep learning: choosing which attributes to consider or ignore can significantly influence your model’s prediction accuracy. But while its impact on accuracy is easy to measure, its impact on the model’s bias is not.

Why AI bias is hard to fix

Given that context, some of the challenges of mitigating bias may already be apparent to you. Here we highlight four main ones.

Unknown unknowns. The introduction of bias isn’t always obvious during a model’s construction because you may not realize the downstream impacts of your data and choices until much later. Once you do, it’s hard to retroactively identify where that bias came from and then figure out how to get rid of it. In Amazon’s case, when the engineers initially discovered that its tool was penalizing female candidates, they reprogrammed it to ignore explicitly gendered words like “women’s.” They soon discovered that the revised system was still picking up on implicitly gendered words—verbs that were highly correlated with men over women, such as “executed” and “captured”—and using that to make its decisions.

Imperfect processes. First, many of the standard practices in deep learning are not designed with bias detection in mind. Deep-learning models are tested for performance before they are deployed, creating what would seem to be a perfect opportunity for catching bias. But in practice, testing usually looks like this: computer scientists randomly split their data *before* training into one group that’s actually used for training and another that’s reserved for validation once training is done. That means the data you use to test the performance of your model has the same biases as the data you used to train

it. Thus, it will fail to flag skewed or prejudiced results.

Lack of social context. Similarly, the way in which computer scientists are taught to frame problems often isn't compatible with the best way to think about social problems. For example, in [a new paper](#), Andrew Selbst, a postdoc at the Data & Society Research Institute, identifies what he calls the "portability trap." Within computer science, it is considered good practice to design a system that can be used for different tasks in different contexts. "But what that does is ignore a lot of social context," says Selbst. "You can't have a system designed in Utah and then applied in Kentucky directly because different communities have different versions of fairness. Or you can't have a system that you apply for 'fair' criminal justice results then applied to employment. How we think about fairness in those contexts is just totally different."

The definitions of fairness. It's also not clear what the absence of bias should look like. This isn't true just in computer science—this question has a long history of debate in philosophy, social science, and law. What's different about computer science is that the concept of fairness has to be defined in mathematical terms, like balancing the false positive and false negative rates of a prediction system. But as researchers have discovered, there are many different mathematical definitions of fairness that are also mutually exclusive. Does fairness mean, for example, that the [same proportion](#) of black and white individuals should get high risk assessment scores? Or that the [same level of risk](#) should result in the same score regardless of race? It's impossible to fulfill both definitions at the same time ([here's](#) a more in-depth look at why), so at some point you have to pick one. But whereas in other fields this decision is understood to be something that can change over time, the computer science field has a notion that it should be fixed. "By fixing the answer, you're solving a problem that looks very different than how society tends to think about these issues," says Selbst.



35 Innovators Under 35: Save 25%

Meet the exceptional honorees that are redefining what's possible across AI, biotech, climate and energy, and more. Plus, get a free gift.

CLAIM OFFER

If you're reeling from our whirlwind tour of the full scope of the AI bias problem, so am I. But fortunately a strong contingent of AI researchers are working hard to address the problem. They've taken a variety of approaches: algorithms that help [detect](#) and [mitigate](#) hidden biases within training data or that [mitigate](#) the [biases](#) learned by the model regardless of the data

POPULAR

Happy birthday, baby! What th
for those born today


Kara Platoni

This researcher wants to repla
little by little

Antonio Regalado

quality; processes that hold companies accountable to the fairer outcomes and discussions that hash out the different definitions of fairness.

“‘Fixing’ discrimination in algorithmic systems is not something that can be solved easily,” says Selbst. “It’s a process ongoing, just like discrimination in any other aspect of society.”

*This originally appeared in our AI newsletter *The Algorithm*. To have it directly delivered to your inbox, [sign up here](#) for free. *

by Karen Hao



35 Innovators Under 35: Save 25%

Meet the exceptional honorees that are redefining what's possible across AI, biotech, climate and energy, and more. Plus, get a free gift.

CLAIM OFFER

DEEP DIVE

POLICY



What Japan's "megaquake" warning really tells us

Fears that a quake last week was a foreshock