The Algorithms Aren't Biased, We Are

Rahul Bhargava · Follow

Published in MIT MEDIA LAB

5 min read · Jan 3, 2018

Excited about using AI to improve your organization's operations? Curious about the promise of insights and predictions from computer models? I want to warn you about bias and how it can appear in those types of projects, share some illustrative examples, and translate the latest academic research on "algorithmic bias."

First off — language matters. What we call things shapes our understanding of them. That's why **I try to avoid the hype-driven term "artificial intelligence."** Most projects called that are more usefully described as "machine learning." Machine learning can be described as the process of training a computer to make decisions that you want help making. This post describes why you need to worry about the data in your machine learning problem.

This matters in a lot of ways. "Algorithmic bias" is showing up all over the press right now. What does that term mean? Algorithms are doling out discriminatory sentence recommendations for judges to use. Algorithms are baking in gender stereotypes to translation services. Algorithms are pushing viewers towards extremist videos on YouTube. Most folks I know agree this is not the world we want. Let's dig into why that is happening, and put the blame where it should be.

Your machine is learning, but who is teaching it?

Physics is hard for me. Even worse — I don't think I'll *ever* be good at physics. I attribute a lot of this to a poor high school physics teacher, who was condescending to me and the other students. On the other hand, while I'm not great at complicated math, I like trying to learn it better. I trace this continued enthusiasm to my junior high school math teacher, who introduced us to the topic with excitement and playfulness (including donut rewards for solving bonus problems!).

My point in sharing this story? Teachers matter. This is even more true in machine learning — machines don't bring prior experience, contextual beliefs, and all the other things that make it important to meet human learners where they are and provide many paths into content. Machines only learn from only what you show them.

So in machine learning, the questions that matter are "what is the textbook" and "who is the teacher." The textbook in machine learning is the "training data" that you show to your software to teach it how to make decisions. This usually is some data you've examined and labeled with the answer you want. Often it is data you've gathered from lots of other sources that did that work already (we often call this a "corpus"). If you're trying to predict how likely someone receiving a micro-loan is to repay it, then you might pick training data that includes previous payment histories of current loan recipients.

The second part is about who the teacher is. The teacher decides what questions to ask, and tells learners what matters. In machine learning, the teacher is responsible for "feature selection" — deciding what pieces of the data the machine is allowed to use to make its decisions. Sometimes this feature selection is done for you by what is and isn't included in the training sets you have. More often you use some statistics to have the computer pick the features most likely to be useful. Returning to our micro-loan example: some candidate features could be loan duration, total amount, whether the recipient has a cellphone, marital status, or their race.

These two questions — training data and training features — are central to any machine learning project.

Algorithms are mirrors

Let's return to this question of language with this in mind.. perhaps a more useful term for "machine learning" would be "machine teaching." This would put the responsibility where it lies, on the teacher. If you're doing "machine learning." you're most interested in what it is learning to do. **With "machine teaching," you're most interested in what you are teaching a machine to do**. That's a subtle difference in language, but a big difference in understanding.

Putting the responsibility on the teacher helps us realize how tricky this process is. Remember this list of biases examples I started with? That sentencing algorithm is discriminatory because it was taught with sentencing data for the US court system, which data shows is very forgiving to everyone except black men. That translation algorithm that bakes in gender stereotypes was probably taught with data from the news or literature, which we known bakes in out-of-date gender roles and norms (ie. Doctors are "he," while nurses are "she"). That algorithm that surfaces fake stories on your feed is taught to share what lots of other people share, irrespective of accuracy.

All that data is about us.

Those algorithms aren't biased, we are! Algorithms are mirrors.

Algorithmic mirrors don’t fully reflect the world around us, nor the world we want

They reflect the biases in our questions and our data. These biases get baked into machine learning projects in both feature selection and training data. This is on us, not the computers.

Corrective lenses

So how do we detect and correct this? Teachers feel a responsibility for, and pride in, their students’ learning. Developers of machine learning models should feel a similar responsibility, and perhaps should be allowed to feel a similar pride.

I’m heartened by examples like Microsoft’s efforts to undo gender bias in publicly available language models (trying to solve the “doctors are men” problem). I love my colleague Joy Buolamwini’s efforts to r[illegible] is [illegible] of “jus[illegible]” [illegible] th[illegible] te[illegible] [illegible]n she calls the “Algorithmic Justice League” (video). ProPublica’s investigative reporting is holding companies accountable for their discriminatory sentencing predictions. The amazing Zeynep Tufekci is leading the way in speaking and writing about the danger this poses to society at large. Cathy O’Neil’s Weapons of Math Destruction documents the myriad of implications for this, raising a warning flag for society at large. Fields like law are debating the implications of algorithm-driven decision making in public policy settings. City ordinances are starting to tackle the question of how to legislate against some of the effects I’ve described.

Research Machine Learning Ethics Algorithms Artificial Intelligence

These efforts can hopefully serve as “corrective lenses” for these algorithmic mirrors — addressing the troubling aspects we see in our own reflections. The key here is to remember that it is up to us to do something about this. **Determining a decision with an algorithm doesn’t automatically make it reliable**