The New York Times | https://nyti.ms/38brSto

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

By Sendhil Mullainathan

Dec. 6, 2019

In one study published 15 years ago, two people applied for a job. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In a study published this year, two patients sought medical care. Both were grappling with diabetes and high blood pressure. One patient was black, the other was white.

Both studies documented racial injustice: In the first, the applicant with a black-sounding name got fewer job interviews. In the second, the black patient received worse care.

But they differed in one crucial respect. In the first, hiring managers made biased decisions. In the second, the culprit was a computer program.

As a co-author of both studies, I see them as a lesson in contrasts. Side by side, they show the stark differences between two types of bias: human and algorithmic.

Marianne Bertrand, an economist at the University of Chicago, and I conducted the first study: We responded to actual job listings with fictitious résumés, half of which were randomly assigned a distinctively black name.

The study was: "Are Emily and Greg more employable than Lakisha and Jamal?"

The answer: Yes, and by a lot. Simply having a white name increased callbacks for job interviews by 50 percent.

I published the other study in the journal “Science” in late October with my co-authors: Ziad Obermeyer, a professor of health policy at University of California at Berkeley; Brian Powers, a clinical fellow at Brigham and Women’s Hospital; and Christine Vogeli, a professor of medicine at Harvard Medical School. We focused on an algorithm that is widely used in allocating health care services, and has affected roughly a hundred million people in the United States.

To better target care and provide help, health care systems are turning to voluminous data and elaborately constructed algorithms to identify the sickest patients.

We found these algorithms have a built-in racial bias. At similar levels of sickness, black patients were deemed to be at lower risk than white patients. The magnitude of the distortion was immense: Eliminating the algorithmic bias would more than double the number of black patients who would receive extra help. The problem lay in a subtle engineering choice: to measure “sickness,” they used the most readily available data, health care expenditures. But because society spends less on black patients than equally sick white ones, the algorithm understated the black patients’ true needs.

One difference between these studies is the work needed to uncover bias.

Our 2004 résumé study resembled a complex covert operation more than traditional academic research. We created a large bank of fictitious résumés and scraped help wanted ads every day. We faxed (yes, the study was that long ago) résumés for each job offer, and established phone numbers with voice mail. Then we waited for prospective employers to call back.

This went on for months — all before we had even one data point to analyze. Pinpointing discriminatory behavior by a particular group of people — in this case, hiring managers — is often very hard.

By contrast, uncovering algorithmic discrimination was far more straightforward. This was a statistical exercise — the equivalent of asking the algorithm “what would you do with this patient?” hundreds of thousands of times, and mapping out the racial differences. The work was technical and rote, requiring neither stealth nor resourcefulness.

Humans are inscrutable in a way that algorithms are not. Our explanations for our behavior are shifting and constructed after the fact. To measure racial discrimination by people, we must create controlled circumstances in the real world where only race differs. For an algorithm, we can create equally controlled just by feeding it the right data and observing its behavior.

Algorithms and humans also differ on what can be done about bias once it is found.

With our résumé study, fixing the problem has proved to be extremely difficult. For one, having found bias on average didn't tell us that any one firm was at fault, though recent research is finding clever ways to detect discrimination.

Another problem is more fundamental. Changing people's hearts and minds is no simple matter. For example, implicit bias training appears to have a modest impact at best.

By contrast, we've already built a prototype that would fix the algorithmic bias we found — as did the original manufacturer, who, we concluded, had no intention of producing biased results in the first place. We offered a free service to health systems using these algorithms to help build a new one that was not racially biased. There were many takers.

Changing algorithms is easier than changing people: software on computers can be updated; the "wetware" in our brains has so far proven much less pliable.

None of this is meant to diminish the pitfalls and care needed in fixing algorithmic bias. But compared with the intransigence of human bias, it does look a great deal simpler.

Discrimination by algorithm can be more readily discovered and more easily fixed. In a 2018 paper with Cass Sunstein, Jon Kleinberg and Jens Ludwig, I took a cautiously optimistic perspective and argued that with proper regulation, algorithms can help to reduce discrimination.

But the key phrase here is "proper regulation," which we do not currently have.

We must ensure all the necessary inputs to the algorithm, including the data used to test and create it, are carefully stored. Something quite similar is already required in financial markets, where copious records are preserved and reported, while preserving the commercial secrecy of the firms involved. We will need a well-funded regulatory agency with highly trained auditors to process this data.

Once proper regulation is in place, better algorithms can help to ensure equitable treatment in our society, though they won't resolve the deep, structural bias that continues to plague the United States. Fixing the biases of society is no easier than fixing the biases of people.

After reading our report on the bias in health algorithms, my father reminded me an episode from my childhood.

When I was eight or nine, we were going to Sears to have a family photo taken. We had just come to Los Angeles from India, where I had grown up in a world in which photographs were rare and almost magical. So this was a special trip. I remember that my mother donned a beautiful sari for the occasion.

A photographer took our photos and, days later, at home, we expectantly opened the envelope, only to find disappointment inside. Our faces were barely visible. Only the whites of our teeth and eyes came through. We learned, much later, that the equipment had been calibrated for white skin, an experience shared by many people with darker skin.

My father had made an astute connection. The photo developers and the algorithm builders had made the same error: failing to appreciate the diversity of people their equipment might be used on.

The analogy can go further: Our résumé study was akin to finding that photographers themselves were biased.

And that would have been a different problem. It is much easier to fix a camera that does not register dark skin than to fix a photographer who fails to see dark skinned people.

Sendhil Mullainathan is a professor of behavioral and computational science at the University of Chicago. Follow him on Twitter: @m_sendhil.