BROOKINGS

Report

To stop algorithmic bias, we first have to define it

Emily Bembeneck, Rebecca Nissan, and Ziad Obermeyer Thursday, October 21, 2021

In sectors as diverse as health care, criminal justice, and finance, algorithms are increasingly used to help make complex decisions that are otherwise troubled by human biases. Imagine criminal justice decisions made without race as a factor or hiring decisions made without gender preference. The upside of AI is clear: human decisionmakers are far from perfect, and algorithms hold great promise for improving the quality of decisions. But disturbing examples of algorithmic bias have come to light. Our own work has shown, for example, that a widely-used algorithm recommended less health care to Black patients despite greater health needs. In this case, a deeply biased algorithm reached massive scale without anyone catching it—not the makers of the algorithm, not the purchasers, not those affected, and not regulators.[1]

Since both human and algorithmic decisionmakers introduce the possibility of bias, removing algorithms entirely isn't always the best approach. In fact, in some cases, biased algorithms may be easier to fix than biased humans. However, it falls on policymakers to ensure that the algorithms helping make complex decisions are doing so in a just and equitable way. To do so, we believe three key steps are required. First, regulators must define bias practically, with respect to its real-world consequences. Second, once the goalposts are clear, regulators must use them to provide much-needed guidance for industry and to define targets for objective, hard-hitting investigations into biased algorithms. Third, as in other fields, regulators should insist on specific internal accountability structures, documentation protocols, and other preventative measures that can stop bias before it happens.

Rather than prescriptive rules, which would quickly become obsolete in a rapidly evolving field, we believe these practices will both set a foundation for mitigating bias and provide clear protocols for investigating bad actors.

Defining the goalposts

We believe the key reason AI is still mostly unregulated is that regulators don't currently have a vocabulary to articulate the benefits and harms of algorithms and hold them accountable. When we regulate a toaster oven, we know it should not catch on fire. When we regulate a pharmaceutical, we know the benefits should outweigh any side effects. But what do we measure when we regulate algorithms?

Algorithms provide decisionmakers with information—a forecast, a probability, or some other key unknown—in order to improve the quality of decisions. Building on this simple observation, we propose that regulators should ask two questions to hold AI accountable: first, what is the ideal information that an algorithm should be providing? And second, is it doing so accurately, both overall and for protected groups?

Consider the example we examined in Obermeyer et al (2019) of a biased health algorithm that caused untold harm. We used data from one health system to study the way a large family of algorithms works. Population health management algorithms like this one are used around the world by health systems and insurers alike to decide who gets access to so-called "extra help" programs. These 'care management' programs provide additional up-front care to patients with chronic conditions with the aim of reducing flare-ups and complications in the future. Patients avoid health problems, and the health care system saves the money it would have spent on ER visits and hospitalizations—a win-win.

What is the ideal information the algorithm should be providing in this instance? It was supposed to identify patients who were going to get sick tomorrow so hospitals could enroll them in the extra help program today. We call that goal of the algorithm its **ideal target**. But what was the algorithm actually doing? In fact, it was doing something subtly but importantly different: it was predicting not who was going to get sick but who was going to generate high costs for the health care system. This is the algorithm's **actual target**. The wedge between these two is a key driver of bias.

The algorithm made its predictions based on cost because an assumption was made that health care costs are a fitting proxy for health care need. While that seems reasonable, not everyone who needs health care gets health care—which means some patients end up having lower costs than others, even though they need the same care. This wedge—between what the algorithm was supposed to be doing and what it was actually doing—led to Black patients being deprioritized for the program, resulting in unmeasured harm against many patients. The ideal target, the decision we care about, was *need for care*, but the algorithm's actual target was *cost of care*. We wanted the algorithm to answer one question, but it was answering something else.

Algorithms are like genies. A man asks a genie for his wish, "I want to be rich!" The genie replies, "Ok, Rich, what's your next wish?" Algorithms are concrete and literal to a fault. When we investigate the decisions an algorithm is informing, we can understand the gap between what we want it to do and what we actually told it to do. This gap between the ideal target and the actual target results in what we call **label choice bias**. In follow-up work since the example above, we've found bias in a wide range of other algorithms, nearly all driven by biased proxies, leading to biased outcomes. This bias, disturbing as it is, is just a symptom of a deeper problem: algorithms are often not doing what they are supposed to be doing. To catch problems like these, we need to understand algorithms in context. What do we want them to do? What are they actually doing? Is there a discrepancy, and if so, is it different for different groups?

To summarize, algorithms provide a key piece of information to a decisionmaker. So regulation must ensure that they are providing that information accurately, both overall and for specific groups. This approach, which focuses on the *output* of the algorithm—the accuracy of its predictions on an ideal target—has several key advantages. It does not require 'opening the black box': algorithms can be audited simply based on the scores they produce. This avoids compromising trade secrets. It also means regulators do not need to regulate the *inputs* of algorithms or understand the many reasons why they might be biased—non-representative training data, use of an explicit race correction, etc. Instead, regulators can focus on one simple question: is the algorithm predicting its ideal target accurately and equitably? This test will detect many forms of bias, like failure of an

algorithm to generalize from one population to another. But unlike many other measures of bias proposed, it will also detect label choice bias, which we've found is a major driver of algorithmic bias in many settings.

Translating the goalposts into action

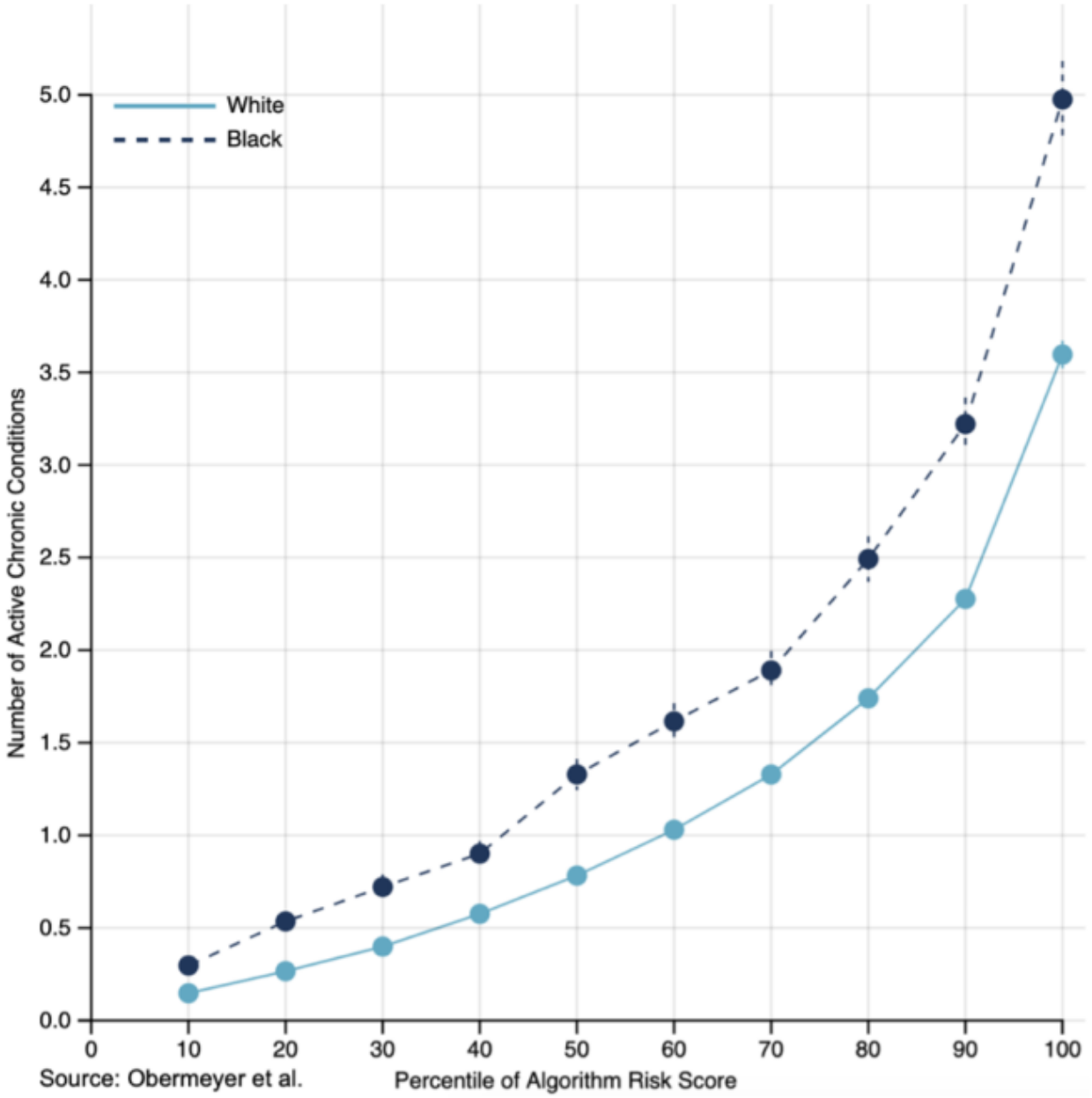
Defining the goalposts can be transformative. In our experience, most people who make algorithms do not want to scale up racial biases. They want to design products that work and to stay on the right side of regulators, but because this is a new and fast-moving field, they often don't know how. By providing guidance to industry on what bias looks like and how to avoid it, regulators can have a transformative effect on the future of algorithms in many fields.

At the same time, regulators also need to be prepared to enforce the standards and make sure the industry is staying within those goalposts that have been set. There are a couple key tools regulators can use to investigate suspected cases of bias.

First, start with some simple statistical tests to see if the algorithm is performing well on the variable it is actually predicting. These tests are as simple as making a graph or running a regression in any standard statistical package: simply compare the variable being predicted to the algorithms score, by group. This should detect problems caused by non-diverse training data, failures to generalize, or poor performance in underserved groups (this is termed "calibration" in the literature). But keep in mind: good performance in one protected group does not mean the algorithm is free of bias.

Second, hold the algorithm accountable for predicting the *ideal* target for underserved populations. Once you define your ideal target variable, you can check whether the algorithm does a good job of predicting that variable across populations. Again, this is as simple as a graph or a regression comparing the ideal target to the algorithm score, by group. Notice that we don't need to understand or explain the inner workings of the "black box" in order to find bias. We just need the outputs of that box (the algorithm's predictions) and some data representing the outcome we care about (the ideal target) for our population.

Often, we find that algorithms are poorly calibrated when you consider the outcome that is truly important. In the health care example above, where an algorithm informs whether people are offered an extra help program, we found that if you take two patients, one Black and one white, who are equally sick (as defined by active number of chronic conditions), the algorithm is more likely to recommend extra care for the white patient than the Black one (see graph below). This is a case where something has certainly gone wrong.



Source: Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Solving these problems is simply a matter of matching the actual target (what the algorithm is predicting) more closely to the ideal target (what we want the algorithm to predict). Again considering the health care example, instead of predicting cost, we should predict the variable we truly care about—health care needs. In our research, adjusting the target led to almost double the number of Black patients being selected for the program. Fixing algorithmic bias means solving the problems we are really trying to solve.

Accountability structures to prevent algorithmic bias

Policymakers also need to provide guidance on how to set up preventative structures and practices within one's organization. In our Algorithmic Bias Playbook we recommend a few steps, and we think that most of these are broadly applicable to most organizations.

In any organization, it is crucial that a single person be responsible for algorithmic bias. This person, whom we've termed a **steward,** should have broad oversight over strategic decisions so that they are equipped to provide executive support to the team but also so they can be truly accountable for mitigating bias. While a single person should be ultimately responsible, they shouldn't work alone. They must be advised by a diverse group of stakeholders for this effort to be truly successful. Individuals from different organizational capacities with different backgrounds, both internal and external voices, should all be included in an advisory group to the steward so that decisions about algorithmic bias are made collectively.

Aside from a clear accountability structure, organizations also need clear **documentation**. Algorithms live throughout an organization and provide input on decisions that affect thousands or even millions of people. We recommend that the steward and her team create a master inventory of algorithms that can provide a record of items such as the following: the source, the owner, the intended purpose/ideal target, the actual target, the population served, and any other information that the organization deems useful. Keeping records like these will help uncover where bias can arise and help prioritize efforts to ensure equitable outcomes.

Finally, be clear that ignorance is no protection—as our own ongoing work with state law enforcement agencies has taught us. It is no longer reasonable for an organization to be unaware of algorithmic bias as a potential problem. Negligence is in not doing the work to find, mitigate, and prevent it.

Conclusions

Algorithmic bias is everywhere. Our work has shown that label choice bias is a particularly dangerous form of bias: it is widespread because we often don't have the specific variable we're interested in. We've discussed an example from health care, where it's difficult to measure the true outcome we care about—health—and instead organizations often fall back on convenient proxies like cost. Any industry where the variables we care about are not truly available will have this issue. In finance, we may want to measure "creditworthiness," but instead we measure income, employment, demographics, etc. In crime, we may want to know "propensity to commit crime," but instead we look at arrest record, education, employment, etc. These are all proxies for the true variable of interest, which means there is room in every one of these cases for label choice bias to exist. They must be investigated and prevented.

The good news is that we can reduce bias in algorithms by asking organizations to do a better job defining that "ideal target" —the thing we really care about. Organizations across the health care industry have come to us for support as they've worked to identify and mitigate bias in their own contexts. Some problems we've helped address include population health management, operational efficiency, strategic patient engagement, and others. In addition to workshopping specific algorithms, we help organizations apply our framework in practice and establish processes for proactively preventing bias.

Regulators are central to solving the problem of algorithmic bias. While some examples like the Federal Trade Commission guidelines and Food and drug administration regulatory framework are good starts, much more needs to be done. Regulators must identify targets for prompt investigations grounded in real world use cases and provide clear guidance and accountability structures for organizations to follow. Working together, we can move the needle towards more equitable outcomes across sectors.

Ziad Obermeyer has received speaking or consulting fees from AcademyHealth, Anthem, Independence Blue Cross, Premier Inc, and The Academy. The authors did not receive financial support from any firm or person for this article or, other than the aforementioned, from any firm or person with a financial or political interest in this article. They are currently not an officer, director, or board member of any organization with an interest in this article.

Footnotes

1. 1 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.