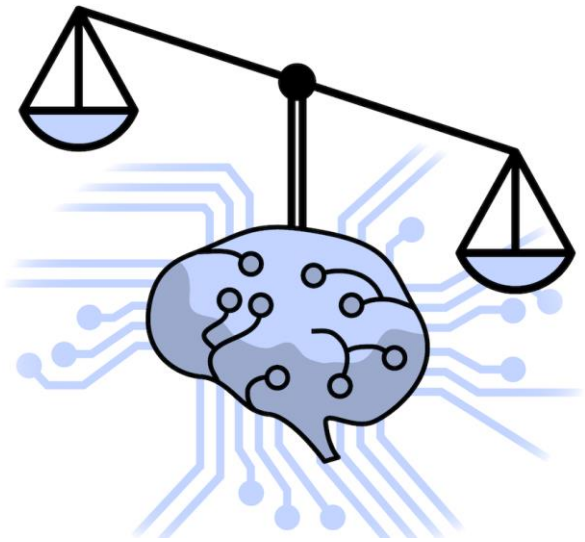

Fairness/Bias in Machine Learning

by Isabella Grabski
figures by Nicholas Lue

It's no secret that bias is present everywhere in our society, from our educational institutions to the criminal justice system. The manifestation of this bias can be as seemingly trivial as the timing of a judge's lunch break or, more often, as fraught as race or economic class. We tend to attribute such discrimination to our own internalized prejudices and our inability to make decisions in truly objective ways. Because of this, machine learning algorithms seem like a compelling solution: we can write software to look at the data, crunch the numbers, and tell us what decision we should make.

In reality, these algorithms can and do fall prey to the same biases as humans. One particularly chilling example is COMPAS, an algorithm used in several U.S. states to determine how likely a given defendant is to commit another crime in the future. This risk assessment is used to help determine high-impact consequences like probation and parole, but an analysis from ProPublica demonstrated that the algorithm's decisions can replicate racial discrimination. Researchers found that COMPAS is almost twice as likely to incorrectly predict black defendants as high risk than white defendants. Using this algorithm, then, can reinforce the same biases we are afraid of in our human decision-making.

COMPAS is not an isolated example. There are far too many to comprehensively list, leading to disparities in who is eligible for same-day Amazon deliveries, who will be shown science career opportunities, and who sees Facebook advertisements for certain types of housing (Figure 1). In short, machine learning algorithms and the biases they pick up can affect a huge component of our day-to-day lives. This issue has not gone unnoticed in the machine learning community and is referred to as the fairness problem.

Fairness is difficult to pin down, and its exact definition is the subject of much contention among researchers. One simplistic way to think about it is that a fair algorithm will make similar decisions for similar individuals, or similar decisions regardless of what demographic an individual belongs to. This definition is vague, of course. Part of the challenge is that we can't even define what a just and unbiased society should look like, let alone the decision-making processes that will bring us there.

Nevertheless, even if we can't state exactly what *fair* should look like, we often have a good idea of what *unfair* is. But where does the unfairness in machine learning algorithms come from, and how can we address it?

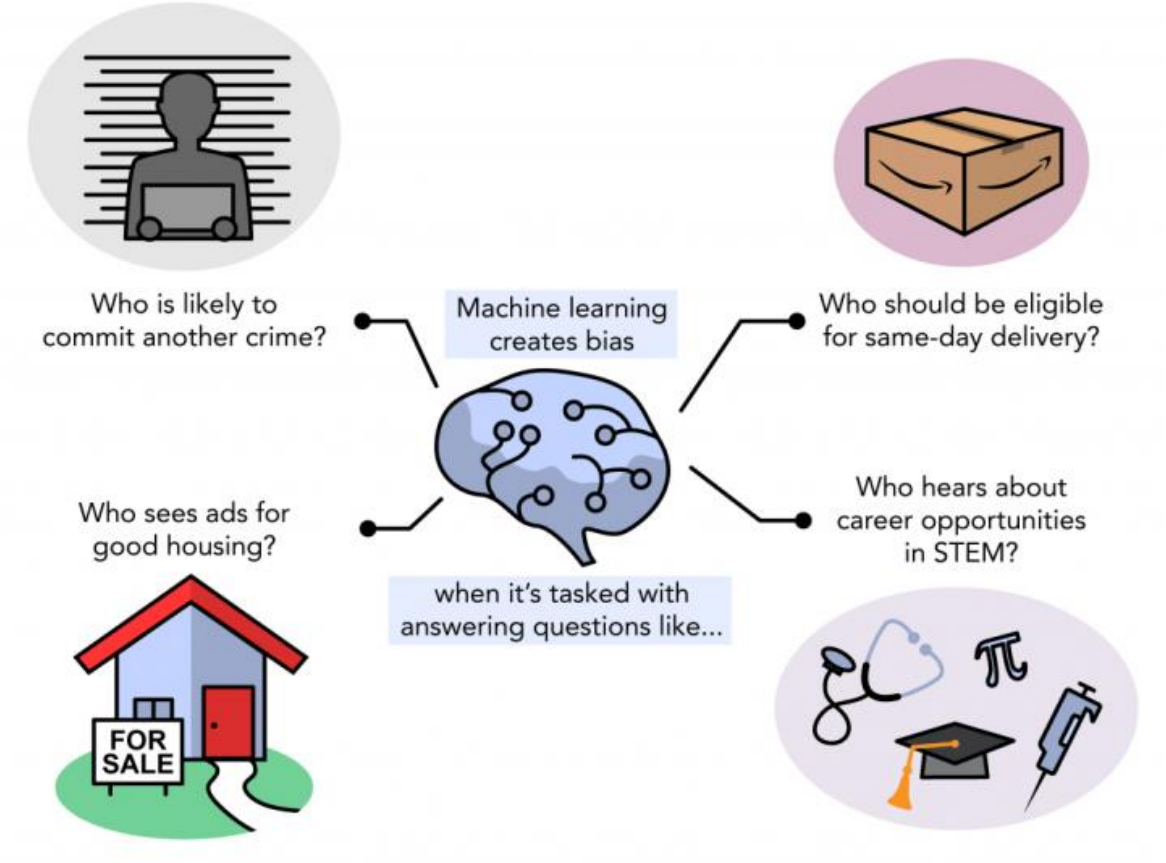

Figure 1. **Examples of how bias in machine learning can affect our daily lives.**

What causes unfairness?

Machine learning algorithms may seem like they should be objective, since decision-making is based entirely on the data. In a typical workflow, an algorithm is shown a large amount of representative data to learn from, and its decision-making process is refined by what it sees. However, any data we give the algorithm is describing, directly or not, the choices that have already been made in society. If black defendants are already falsely determined to be higher risk than white defendants, then an algorithm will learn that from the data as if it were factual. This bias in available training data creates a feedback loop, where the algorithm will make unfair decisions based on what it's learned, perpetuate further discrimination in society, and thus further taint data used down the road.

Unfairness can also arise from too much homogeneity in the data. A classic example is what happened with Nikon's facial recognition algorithm, which automatically detected the presence of blinking in photos. However, this algorithm mistakenly flagged Asian people as blinking at a substantially higher rate than other demographics. Although the exact reason was not explicitly revealed by Nikon, this situation is a textbook example of what might happen when an algorithm is primarily shown data from only one segment of the population. If the algorithm did not see many examples of Asian people, then it would not have been able to correctly learn what an Asian person blinking looks like.

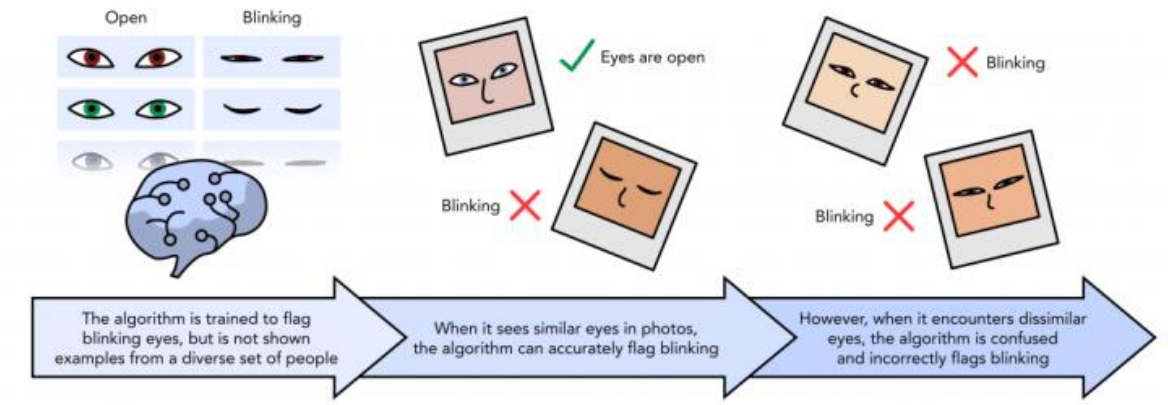

Figure 2: **Nikon's blink detection algorithm may only have been trained on certain types of eyes, leading it to misclassify new types of eyes it hadn't seen before.**

How can we correct unfairness?

Some forms of unfairness may be easier to correct than others. In the case of Nikon's facial recognition algorithm, balancing the data initially shown to the algorithm may have prevented the issue from arising at all. But in many other cases, when the data will always reflect pre-existing discrimination in society, it is much harder to prevent an algorithm from learning those same biases.

One approach is sometimes referred to as fairness through blindness. Here, the attributes at risk of discrimination are left out of the data entirely. In our criminal justice example, we might remove race entirely from the data we show an algorithm like COMPAS. The hope is that if COMPAS never sees what race a defendant is, it can only make decisions based on other characteristics.

The problem with this approach is that something like race does not exist in a vacuum. Many other attributes of a defendant are likely to be associated with their race, such as zip code or profession. These other attributes can then be used inadvertently as proxies for race and lead to essentially the same unfair results.

Another approach takes a completely different tack, and is sometimes called fairness through awareness. Instead of removing attributes that could lead to discrimination from the data, this approach focuses on these protected attributes and forces the algorithm to make comparable decisions among these different segments of the population. This idea can be implemented in several different ways, but the simplest approach would be to ensure that a positive outcome is given at equal rates across demographics. For example, an algorithm like COMPAS could be constrained to predict low risk among black defendants at the same rate as among white defendants.

This idea may seem promising, but it can lead to problems as well by oversimplifying what fairness looks like. There could still be imbalanced decisions within the attributes we correct for. For instance, even if there is an equal rate of low risk predictions between black and white defendants overall, lingering gender or class bias could still result in penalizing black men or low-income black defendants more frequently.

This significant shortcoming highlights a key challenge in trying to solve the fairness problem: our inability to identify what, exactly, fairness should look like. Fairness through awareness attempts to enforce a very simplistic ideal, but in doing so, it only creates more problems.

Some researchers are trying to approach fairness from a different angle. Instead of striving for literal equality at every step of the process, one way to think about fairness is determining how to make decisions that will improve the lives of disadvantaged demographics over time. If researchers can figure out what fairness should look like, then maybe the right machine learning algorithms can guide us there.

Isabella Grabski is a second-year Ph.D. student in the Biostatistics program at Harvard University

Nicholas Lue is a third-year Ph.D. student in the Chemical Biology program at Harvard.