SPRING 2018

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE

THE UNIVERSITY OF TEXAS AT DALLAS

ARIZONA STATE UNIVERSITY

ISSUES

IN SCIENCE AND TECHNOLOGY

Preparing for THE NEXT FLOOD

A Response to Reactionary Populism

Straight Talk on Workforce Skills

New Rules for Self-Driving Cars

Infrastructure in the Anthropocene

The Downton Abbey University System

Using Data to Make Better Decisions

USA $12.00 / CAN $13.00

8 1>

0 56698 57433 0

STEPHANIE WYKSTRA

Philosopher's Corner: What is "Fair"? Algorithms in Criminal Justice

On rare occasions, new technologies open up straightforward routes to a better world. But on many other occasions, the mirage of a simple path forward fades quickly. In the case of social policy algorithms, the promise was that systems from hiring to criminal justice can be improved through "objective" mathematical predictions. Today, however, communities of scholars and practitioners are calling for a closer look before we leap.

A "Governing Algorithms" conference at the University of California, Berkeley, a few years ago explored the "recent rise of algorithms as an object of interest in scholarship, policy, and practice." A group called Fairness, Accountability, and Transparency in Machine Learning (FAT-ML) formed soon after and now holds meetings each year. And New York University recently established the AI Now Institute, with scholars dedicated to investigating the "social implications of artificial intelligence." These conferences and communities aim to bring together researchers and practitioners to think through the use of algorithms in society. A common theme is the concern that predictive algorithms, far from leaving bias in the past, carry both the risk of bias and increased threats of unaccountability and opacity.

Take the use of algorithms in the criminal justice system. Although algorithms have been used in some form in criminal justice decision-making since the 1920s, they are gaining wider use in areas such as pretrial decision-making. The algorithmic tools take in a variety of inputs, ranging from just a few variables to over a hundred, and assign defendants a risk score based on probability of rearrest, failure to appear in court, or both. The score is then often shown to judges, who may choose to release defendants with low risk scores on their own recognizance or under some form of limited supervision. The others are held in jail or assigned bail (which often means remaining in jail because they lack the money to pay bail). For some criminal justice reformers, the hope is that the use of the tools will help to reduce jail populations. And in some places, including New Jersey, the jail population has decreased after adopting pretrial risk assessment algorithms such as the Arnold Foundation's Public Safety Assessment (PSA) and abandoning use of bail. But there has also been criticism of the PSA and other algorithms.

The ethics shaping the algorithm

In 2016, journalists at ProPublica criticized as unfair the risk assessment tool called COMPAS, developed by the company Northpointe (later renamed Equivant). After analyzing data they obtained from a jurisdiction in Florida that uses the algorithm, the reporters concluded that the algorithm is racially biased. They found that among defendants who are not rearrested within two years, 45% of those who are black—compared with 24% of the whites—had been assigned high risk scores. Yet when Northpointe responded to the critique, they pointed to a different statistic, supporting a different sort of fairness: within each risk category, black and white defendants had the same rearrest rate.

Could the algorithm have satisfied both conditions of fairness? A number of academic researchers have argued otherwise, purely on mathematical grounds. The reason is that there are different base rates of rearrest among the racial subgroups. Because the groups are different in that regard, achieving one kind of fairness automatically means that the other will not be achieved. To return to Northpointe's kind of fairness: when a judge sees a particular risk score, he or she can infer that this indicates the same chance of rearrest, regardless of the race of the person given the score. That is, arguably, a kind of fairness.

But achieving this kind of fairness creates another kind of unfairness. Because black defendants are arrested at higher rates, and because criminal history is a significant input into risk-assessment tools, a greater percentage are assigned high risk scores than white defendants. This affects other metrics, including the percentage of those who are not rearrested and yet are deemed high risk, which was the bias that ProPublica pointed out. Since being deemed higher risk makes it more likely that someone will end up in jail, this means there is a disparate impact for black defendants.

However, remedying the kind of unfairness that ProPublica points to has its own problems. It involves a trade-off with the first kind of fairness (i.e., risk scores that mean the same risk of rearrest, regardless of race). After exploring the trade-off between these two kinds of unfairness, a team of Stanford University researchers wrote in the *Washington Post*, "Imagine that…we systematically assigned whites higher risk scores than equally risky black defendants with the goal of mitigating ProPublica's criticism. We would consider that a violation of the fundamental tenet of equal treatment." In other words: it is a difficult and open question whether it would be ethical or constitutional to design an algorithm that treats people differently on the basis of race, in order to achieve the kind of fairness to which ProPublica points.

There's a key question of how much to prioritize safety risks over harming people and their families through incarceration.

Richard Berk, a statistician at the University of Pennsylvania who has developed criminal justice algorithms, emphasizes the trade-offs involved in their design. In a paper on fairness in assessment algorithms, Berk and coauthors write, "These are matters of values and law… They are not matters of science." They discuss kinds of fairness and note that increasing one kind of fairness can often involve decreasing other kinds of fairness, as well as reducing the predictive accuracy. They argue that it is mathematically impossible to achieve what they call "total fairness" in one algorithm. The authors conclude that "it will fall to stakeholders—not criminologists, not statisticians, and not computer scientists—to determine the tradeoffs."

Arvind Narayanan, a computer scientist at Princeton University, added in a recent talk that input from philosophers is needed. There is something incoherent, he said, in laying out technical definitions of fairness without bringing to bear the long history of work on justice and fairness in ethics. "It would be really helpful," he suggested, "to have scholars from philosophy talk about these trade-offs [in algorithms] and give us guidelines about how to go about resolving them."

Another significant normative challenge occurs when decisions are made about which variables (or "features") are included as inputs in the model. Although no algorithm includes race itself an input, some include features that are highly correlated with race and socioeconomic status. Including these variables will run the risk of increasing the disparities in how different communities are treated. By being more likely to deem people as high risk, for instance, they may increase their rate of being held in jail. And there is strong evidence that people who are held in jail as they await court hearings plead guilty at considerably higher rates than do people who are released. The resulting conviction would then serve as an additional data point held against them the next time they are arrested, leading to a vicious circle. To add to that, a criminal record can severely limit housing and job opportunities. Human Rights Watch has expressed this concern and has also pointed out that the use of algorithms does not guarantee a reduction in jail populations.

A third consideration in algorithm design involves weighing the trade-off between the harm done when someone is released and commits a crime and the harm done by holding in jail people who would not have committed any crime if released. In other words, there's a key question of how much to prioritize safety risks over harming people and their families through incarceration.

Finally, a fourth consideration, discussed by legal scholars such as Megan Stevenson and Sandra Mayson, is the question of which outcomes to try to predict. Many tools predict rearrest for any offense, including for low-level crimes and drug offenses. Focusing on the outcome of rearrest for a violent charge might make more sense, scholars suggest, if the primary concern is violent crime.

Connected to all of the normative questions is an (applied) epistemological one: How can we know what decisions were made in designing the algorithms?

Accountability and transparency

Critics of algorithms have also pointed to lack of transparency as a major problem. In 2016, a defendant in Wisconsin who pleaded guilty to eluding the police and operating a vehicle without its owner's consent sued the state, claiming that reliance on an algorithm (Northpointe's COMPAS) violated his rights to due process, in part because he was not able to see the algorithm's code. The company refused to share the algorithm, claiming that the tool's workings are proprietary. The Wisconsin Supreme Court eventually ruled against the defendant, and an appeal to the US Supreme Court was rejected. It also should be noted that the inputs and weights of several other algorithms, including the PSA, are publicly available.

Still, several big questions remain when it comes to transparency. Should individual defendants be notified about what their scores are, and should they have a right to see how the scores were calculated? Beyond that, should other people—for instance, independent researchers—have the access required to check a given tool for computational reproducibility, or whether they arrive at the same results, using the original data and code? In many scientific domains, there's been increasing concern about the reliability of research. This has been a major topic of discussion in fields such as psychology and biomedicine. Recently it's come up in criminal justice as well, with researcher David Roodman at the Open Philanthropy Project finding major methodological problems in seven of the eight criminal justice studies that he reanalyzed. Who, if anyone, is able to check the reproducibility of algorithms, in terms of the original calculations of the designer and whether they appropriately apply to the population to which they are being applied?

On the accountability of algorithms, there are major outstanding questions as well. How can stakeholders weigh in on values they'd like to see the algorithm represent? As we saw above, there are different concepts of fairness, and these concepts interact in complex ways with each other and with accuracy. As more and more jurisdictions adopt assessment algorithms, will not only policy-makers but also communities affected by the tools be able to provide input? After all, these are not objective decisions, but nuanced normative decisions with the potential to affect many thousands of lives.

What's next?

The challenges are daunting. Yet at the same time, as many commentators have pointed out, the baseline is the current US criminal justice system, which on any given day holds in jail nearly half a million people who haven't been convicted of a crime. Some defenders of algorithms also point out that judges conduct their own internal algorithmic decision-making, the workings of which are not transparent and can be affected by factors as irrelevant as whether the judges have eaten lunch. As long as risk-assessment algorithms stand a good chance of improving the status quo, proponents argue, it's worth the effort to work through the challenges.

This argument seems to be a good one, but at the same time, pretrial risk assessment should be used only under certain conditions. First, a tool should be made transparent. At the very least, it should be make public what inputs, outputs, and weights are used. Second, a tool should be regularly audited by independent researchers to check its design and its real-world impact on jail populations and racial inequity (the AI Now Institute has called this kind of check "algorithmic impact assessment"). In order to check effects, data need to be regularly collected and shared with auditors. Third, the wider community of stakeholders beyond algorithm designers should be included in deciding what an algorithm should look like: What kind of fairness is important to prioritize? What threshold of risk should be used for release, and what kind of risk makes sense to measure in this context? (As discussed above, *risk* sounds as if it's referring to public safety, but the algorithm measures the chance of any kind of rearrest, including many activities that pose no danger to the community). And finally, the tools should be used only to reduce jail populations, not to lock up more people unjustly as they await court hearings.

Many jurisdictions have adopted one pretrial tool or another, and as the legal scholar Sonya Starr writes, "It is an understatement to refer to risk assessment as a criminal justice trend. Rather we are already in the risk assessment era." Yet at the same time, it's important to recognize that risk assessment is only as good as the decisions that go into designing the tool. We should hold the tools and their designers accountable, and should use the tools only if they demonstrably serve the goal of dramatically reducing pretrial incarceration.

Stephanie Wykstra is a freelance writer and research consultant, with a focus on research transparency and criminal justice reform. She was previously an assistant professor of philosophy, specializing in epistemology.