The Washington Post

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies ,
Emma Pierson ,
Avi Feller and
Sharad Goel
October 17, 2016

This past summer, a heated debate broke out about a tool used in courts across the country to help make bail and sentencing decisions. It's a controversy that touches on some of the big criminal justice questions facing our society. And it all turns on an algorithm.

The algorithm, called COMPAS, is used nationwide to decide whether defendants awaiting trial are too dangerous to be released on bail. In May, the investigative news organization ProPublica claimed that COMPAS is biased against black defendants. Northpointe, the Michigan-based company that created the tool, released its own report questioning ProPublica's analysis. ProPublica rebutted the rebuttal, academic researchers entered the fray, this newspaper's Wonkblog weighed in, and even the Wisconsin Supreme Court cited the controversy in its recent ruling that upheld the use of COMPAS in sentencing.

It's easy to get lost in the often technical back-and-forth between ProPublica and Northpointe, but at the heart of their disagreement is a subtle ethical question: What does it mean for an algorithm to be fair? Surprisingly, there is a mathematical limit to how fair any algorithm — or human decision-maker — can ever be.

How do you define 'fair'?

The COMPAS tool assigns defendants scores from 1 to 10 that indicate how likely they are to reoffend based on more than 100 factors, including age, sex and criminal history. Notably, race is not used. These scores profoundly affect defendants' lives: defendants who are defined as medium or high risk, with scores of 5-10, are more likely to be detained while awaiting trial than are low-risk defendants, with scores of 1-4.

We reanalyzed data collected by ProPublica on about 5,000 defendants assigned COMPAS scores in Broward County, Fla. (See the end of the post, after our names, for more technical details on our analysis.) For these cases, we find that scores are highly predictive of reoffending. Defendants assigned the highest risk score reoffended at almost four times the rate as those assigned the lowest score (81 percent vs. 22 percent).

But are the scores fair?

Northpointe contends they are indeed fair because scores mean essentially the same thing regardless of the defendant's race. For example, among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended.

Consequently, Northpointe argues, when judges see a defendant's risk score, they need not consider the defendant's race when interpreting it. The plot below shows this approximate equality between white and black defendants holds for every one of Northpointe's 10 risk levels.

But ProPublica points out that among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent). Even though these defendants did not go on to commit a crime, they are nonetheless subjected to harsher treatment by the courts. ProPublica argues that a fair algorithm cannot make these serious errors more frequently for one race group than for another.

You can't be fair in both ways at the same time

Here's the problem: it's actually impossible for a risk score to satisfy both fairness criteria at the same time.

The figure below shows the number of black and white defendants in each of two aggregate risk categories — "low" and "medium or high" — along with the number of defendants within each category who went on to commit another crime.

The plot illustrates four points:

Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race; this is Northpointe's definition of fairness.
The overall recidivism rate for black defendants is higher than for white defendants (52 percent vs. 39 percent).
Black defendants are more likely to be classified as medium or high risk (58 percent vs. 33 percent). While Northpointe's algorithm does not use race directly, many attributes that predict reoffending nonetheless vary by race. For example, black defendants are more likely to have prior arrests, and since prior arrests predict reoffending, the algorithm flags more black defendants as high risk even though it does not use race in the classification.
Black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend; this is ProPublica's criticism of the algorithm.

The key — but often overlooked — point is that the last two disparities in the list above are mathematically guaranteed given the first two observations.

If the recidivism rate for white and black defendants is the same within each risk category, and if black defendants have a higher overall recidivism rate, then a greater share of black defendants will be classified as high risk. And if a greater share of black defendants are classified as high risk, then, as the

plot illustrates, a greater share of black defendants who do not reoffend will also be classified as high risk.

If Northpointe's definition of fairness holds, and if the recidivism rate for black defendants is higher than for whites, the imbalance ProPublica highlighted will always occur. (Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan explore this idea further in their recent paper.)

What should we do?

It's hard to call a rule equitable if it does not meet Northpointe's notion of fairness. A risk score of seven for black defendants should mean the same thing as a score of seven for white defendants. Imagine if that were not so, and 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.

But we should not disregard ProPublica's findings as an unfortunate but inevitable outcome. To the contrary, since classification errors here disproportionately affect black defendants, we have an obligation to explore alternative policies. For example, rather than using risk scores to determine which defendants must pay money bail, jurisdictions might consider ending bail requirements altogether — shifting to, say, electronic monitoring so that no one is unnecessarily jailed.

COMPAS may still be biased, but we can't tell.

Northpointe has refused to disclose the details of its proprietary algorithm, making it impossible to fully assess the extent to which it may be unfair, however inadvertently. That's understandable: Northpointe needs to protect its bottom line. But it raises questions about relying on for-profit companies to develop risk assessment tools.

Moreover, rearrest, which the COMPAS algorithm is designed to predict, may be a biased measure of public safety. Because of heavier policing in predominantly black neighborhoods, or bias in the decision to make an arrest, blacks may be arrested more often than whites who commit the same offense.

Algorithms have the potential to dramatically improve the efficiency and equity of consequential decisions, but their use also prompts complex ethical and scientific questions. The solution is not to eliminate statistical risk assessments. The problems we discuss apply equally to human decision-makers, and humans are additionally biased in ways that machines are not. We must continue to investigate and debate these issues as algorithms play an increasingly prominent role in the criminal justice system.

Sam Corbett-Davies and Emma Pierson are PhD students in the computer science department at Stanford University.

Avi Feller is an assistant professor in the Goldman School of Public Policy at the University of California at Berkeley.

Sharad Goel is an assistant professor in the department of management science and engineering at Stanford University.

Note on methods: ProPublica obtained records for nearly 12,000 defendants in Broward County, Fla., who were assigned a COMPAS score in 2013-2014. ProPublica then determined which defendants were charged with new crimes in the subsequent two years, and made this data set publicly available. We focused on the 5,278 cases involving defendants who are either white or black, and for which a full two years of recidivism information is available. We excluded Hispanic defendants from our analysis because there are not many in this data set. The COMPAS tool also rates defendants on about two dozen other dimensions of risk, including likelihood to commit a violent crime, but here we consider only the overall recidivism score.