Machine Bias
Algorithmic injustice and the formulas that increasingly influence our lives.

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson
ProPublica, Dec. 30, 2016, 4:44 p.m.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

The racial bias that ProPublica found in a formula used by courts and parole boards to forecast future criminal behavior arises inevitably from the test's design, according to new research.

The findings were described in scholarly papers published or circulated over the past several months. Taken together, they represent the most far-reaching critique to date of the fairness of algorithms that seek to provide an objective measure of the likelihood a defendant will commit further crimes.

Increasingly, criminal justice officials are using similar risk prediction equations to inform their decisions about bail, sentencing and early release.

The researchers found that the formula, and others like it, have been written in a way that guarantees black defendants will be inaccurately identified as future criminals more often than their white counterparts.

The studies, by four groups of scholars working independently, suggests the possibility that the widely used algorithms could be revised to reduce the number of blacks who were unfairly categorized without sacrificing the ability to predict future crimes.

The author of one of the papers said that her ongoing research suggests that this result could be achieved through a modest change in the working of the formula ProPublica studied, which is known as COMPAS.

An article published earlier this year by ProPublica focused attention on possible racial biases in the COMPAS algorithm. We collected the COMPAS scores for more than 10,000 people arrested for crimes in Florida's Broward's County and checked to see how many were charged with further crimes within two years.

When we looked at the people who did not go on to be arrested for new crimes but were dubbed higher risk by the formula, we found a racial disparity. The data showed that black defendants were twice as likely to be incorrectly labeled as higher risk than white defendants. Conversely, white defendants labeled low risk were far more likely to end up being charged with new offenses than blacks with comparably low COMPAS risk scores.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks. Read the story.

Northpointe, the company that sells COMPAS, said in response that the test was racially neutral. To support that assertion, company officials pointed to another of our findings, which was that the rate of accuracy for COMPAS scores — about 60 percent — was the same for black and white defendants. The company said it had devised the algorithm to achieve this goal. A test that is correct in equal proportions for all groups cannot be biased, the company said.

This question of how an algorithm could simultaneously be fair and unfair intrigued some of the nation's top researchers at Stanford University, Cornell University, Harvard University, Carnegie Mellon University, University of Chicago and Google.

The scholars set out to address this question: Since blacks are re-arrested more often than whites, is it possible to create a formula that is equally predictive for all races without disparities in who suffers the harm of incorrect predictions?

Working separately and using different methodologies, four groups of scholars all reached the same conclusion. It's not.

Revealing their preliminary findings on a Washington Post blog, a group of Stanford researchers wrote: "It's actually impossible for a risk score to satisfy both fairness criteria at the same time."

The problem, several said in interviews, arises from the characteristic that criminologists have used as the cornerstone for creating fair algorithms, which is that formula must generate equally accurate forecasts for all racial groups.

The researchers found that an algorithm crafted to achieve that goal, known as "predictive parity," inevitably leads to disparities in what sorts of people are incorrectly classified as high risk when two groups have different arrest rates.

"'Predictive parity' actually corresponds to 'optimal discrimination,'" said Nathan Srebro, associate professor of computer science at the University of Chicago and the Toyota Technological Institute at Chicago. That's because predictive parity results in a higher proportion of black defendants being wrongly rated as high-risk.

Srebro's research paper, "Equality of Opportunity in Supervised Learning," was co-authored with Google research scientist Moritz Hardt and University of Texas at Austin computer science professor Eric Price in October. Their paper proposed a definition of "nondiscrimination" that requires the error rates between groups be equalized. Otherwise, Srebro said, one group ends up "paying the price for the uncertainty" of the algorithm.

The need to look at the harms that arise when a test is inaccurate arises frequently in statistics, particularly in fields like health care. When researchers weigh the merits of

exams like mammograms, they want to know both how often they correctly detect breast cancer and how often they falsely indicate that patients have the disease.

False findings are significant in medicine because they can cause patients to unnecessarily undergo painful procedures like breast biopsies. It's entirely possible that a test could correctly identify most breast cancers, showing what's known as "positive predictive value," and yet make so many mistakes that it is viewed as unusable.

When he first heard about the COMPAS debate, Jon Kleinberg, a computer science professor at Cornell University, hoped he could figure out a way to reduce false findings while keeping the positive predictive value intact. "We thought, can we fix it?" he said.

But after he, his graduate student Manish Raghavan and Harvard economics professor Sendhil Mullainathan downloaded and crunched ProPublica's data, they realized that the problem was not resolvable. A risk score, they found, could either be equally predictive or equally wrong for all races — but not both.

The reason was the difference in the frequency with which blacks and whites were charged with new crimes. ""If you have two populations that have unequal base rates," Kleinberg said, "then you can't satisfy both definitions of fairness at the same time."

Kleinberg and his colleagues went on to construct a mathematical proof that the two notions of fairness are incompatible. The paper, "Inherent Trade-Offs in the Fair Determination of Risk Scores" was posted online in September.

In the criminal justice context, false findings can have far-reaching effects on the lives of people charged with crimes. Judges, prosecutors and parole boards use the scores to help decide whether defendants can be sent to rehab programs instead of prison or be given shorter sentences.

Defendants inaccurately classed as "high risk" and deemed more likely to be arrested in the future may be treated more harshly than is just or necessary, said Alexandra Chouldechova, Assistant Professor of Statistics & Public Policy at Carnegie Mellon University, who also studied ProPublica's COMPAS findings.

Chouldechova said focusing on outcomes might be a better definition of fairness. To create equal outcomes, she said, "You would have to treat people differently." Chouldechova's paper, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," was posted online in October.

Chouldechova is continuing to research ways to improve the likelihood of equal outcomes.

Using the Broward County data we made public, Chouldechova rearranged how the COMPAS scores are interpreted so that they were wrong equally often about black and white defendants.

This shift meant that the algorithm's predictions of future criminal behavior were no longer the same for all races. Chouldechova said her revised formula was unchanged for white defendants (59 percent correct) while its predictive accuracy rose from 63 to 69 percent for black defendants.

Northpointe, the company that sells the COMPAS tool, said it had no comment on the critiques. And officials in Broward County said they have made no changes in how they use the COMPAS scores in response to both ProPublica's initial findings and the research papers that followed.

Like this story? Sign up for our daily newsletter to get more of our best work.

Steal Our Stories
Unless otherwise noted, you can republish our stories for free if you follow these rules.

Download Our Data

Send Us Tips or Documents Securely