Ivey | Publishing

W42647

MACHINE LEARNING BIAS: ALGORITHMS IN THE COURTROOM

Professors Yasser Rahrovani and Lauren Cipriano wrote this case solely to provide material for class discussion. The authors do not intend to illustrate either effective or ineffective handling of managerial situation. The authors may have disguised certain names and other identifying information to protect confidentiality.

CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a risk-assessment tool designed to predict the likelihood of a defendant reoffending. It was developed by Northpointe Inc. (now Equivant) starting in 1998, with the recidivism risk scale being used since 2000.[1] The tool was created with the aim of standardizing decision-making within the criminal justice system, intending to reduce human error and bias in court rulings—therefore, "simplifying justice."[2]

The COMPAS tool is a complex, multi-layered artificial intelligence (AI) computing system that uses a combination of a defendant's criminal history, demographic information, and answers to a questionnaire to generate a risk score. The questionnaire includes around 137 questions that cover various topics such as behaviour tendencies, social ties, family history, and personal attitudes. These data points are scored against historical data and adjusted to Northpointe-appointed norm groups. The scores range from 1 to 10, categorized as low risk (1–4), medium risk (5–7), and high risk (8–10).[3]

COMPAS was adopted by criminal justice systems in numerous US states, including New York, Wisconsin, and California, as well as other municipalities. It was used in decisions related to pre-trial release, parole, and probation. By 2015, over 60,000 COMPAS assessments were evaluated each year in New York state.[4]

PROPUBLICA'S CRITICISM OF THE COMPAS TOOL

In 2016, ProPublica, Inc. published a groundbreaking investigative report titled *Machine Bias*, which scrutinized the COMPAS risk-assessment software and argued that it perpetuates racial bias against Black defendants in the US criminal justice system.[5] This report raised critical concerns about the fairness and reliability of algorithmic tools used in high-stakes decision-making, particularly in determining whether defendants should be released or detained before trial.

[1] Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (2018), DOI: 10.1126/sciadv.aao5580.
[2] "Home Page," Equivant, accessed January 18, 2025, https://www.equivant.com/.
[3] Dressel and Farid "The Accuracy, Fairness, and Limits."
[4] S. Thomas, "The Fairness Fallacy: Northpointe and the COMPAS Recidivism Prediction Algorithm" (unpublished undergraduate thesis, Institute for the Study of Human Rights, Columbia University, 2023).
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

The first major issue identified in the report was the racial disparities in COMPAS's predictions. ProPublica found that while the software had similar overall accuracy rates for Black and white defendants, the types of errors it made were strikingly unequal. Black defendants who did not reoffend were incorrectly predicted to reoffend at a rate of 44.9 per cent, nearly double the 23.5 per cent rate for white defendants. Conversely, white defendants who did reoffend were mistakenly classified as low risk at a rate of 47.7 per cent, almost twice the 28 per cent rate for Black defendants. These findings underscored how COMPAS's predictive errors disproportionately disadvantaged Black individuals.

Another central argument in the report was the disparate impact of COMPAS's predictions. The imbalanced error rates effectively favoured white defendants while penalizing Black defendants, creating a system that had a disproportionate adverse effect on a marginalized group. This disparate impact was troubling because it suggested that the tool reinforced inequality even without explicit racial intent.

The report further highlighted the problematic use of historical data in training the COMPAS algorithm. The software was built on data from the US criminal justice system, which had a well-documented history of over-policing Black communities. This overrepresentation of Black individuals in the training data likely contributed to the biased outcomes, as it reflected and perpetuated systemic inequities.

ProPublica also pointed out that while the COMPAS algorithm did not explicitly use race as a factor, it incorporated variables served as proxies for race. For example, it considered whether a defendant's friends used drugs or a parent had been arrested. These variables, tied to socio-economic conditions that disproportionately affected Black communities, indirectly encoded racial bias into the predictions.

Last, the report argued that COMPAS not only mirrored existing biases in the criminal justice system but exacerbated them. By aligning with sentencing trends that historically disadvantaged Black defendants, the algorithm reinforced and magnified systemic inequities it was ostensibly designed to mitigate.

ProPublica's report sparked significant media attention, with major news outlets highlighting concerns about racial bias in criminal justice algorithms. For instance, the *New York Times* discussed the implications of algorithmic bias in sentencing, emphasizing the need for transparency and fairness in such tools.[6] Similarly, the *Atlantic* examined the reliability of predictive algorithms in the justice system, questioning their effectiveness and potential for discrimination.[7]

NORTHPOINTE'S RESPONSE

Northpointe responded to ProPublica's report with a detailed rebuttal titled *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*.[8] In its paper, Northpointe rejected the claim that the COMPAS algorithm was racially biased against Black defendants. It presented several counter-arguments, aiming to demonstrate that its tool operated equitably and accurately across racial groups.

A central focus of Northpointe's argument was the concept of predictive parity. Predictive parity meant that, among defendants classified as high-risk, the likelihood of reoffending was the same regardless of race. Northpointe asserted that COMPAS achieved this standard, arguing that both Black and white individuals labelled as high-risk by the software had an equal probability of committing new offences. This, according to the company, was evidence that the tool treated racial groups fairly.

---

[6] Kate Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*, June 25, 2016, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

[7] Megan Garber, "When Algorithms Take the Stand," The Atlantic, June 30, 2016, https://www.theatlantic.com/technology/archive/2016/06/when-algorithms-take-the-stand/489566/.

[8] William Dieterich, Christina Mendoza, and Tim Brennan, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, July 8, 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616/.

Northpointe also highlighted what it called "accuracy equity." It argued that COMPAS could discriminate between individuals who did and did not reoffend equally effectively for Black and white defendants. This claim was based on statistical measures such as the area under the receiver-operating characteristic curve (ROC). Northpointe stated that the ROC values for Black and white defendants were nearly identical, suggesting that the algorithm did not favour one group over the other.

Another significant point of contention was ProPublica's use of classification statistics. Northpointe criticized ProPublica for focusing on metrics that ignored differences in the recidivism base rates between Black and white defendants. According to Northpointe, ProPublica analyzed what it called "model errors," which evaluated errors independently for individuals who reoffended and for those who did not. Northpointe argued that this approach was less relevant than analyzing "target population errors," which considered the overall base rate of recidivism in the population. By accounting for these base rates, Northpointe claimed that its tool demonstrated fairness and that ProPublica's conclusions were flawed.

Base rates and score distribution were also key aspects of Northpointe's defence. The company explained that the observed differences in error rates—higher false positives for Black defendants and higher false negatives for white defendants—were a natural outcome of using unbiased scoring rules in populations with differing recidivism base rates. For instance, in Broward County, Florida, Black defendants had a recidivism rate of 51 per cent, compared to 39 per cent for white defendants. Northpointe argued that these differences in recidivism rates influenced the sensitivity and specificity trade-offs in the algorithm's predictions rather than indicate bias.

Ultimately, Northpointe maintained that COMPAS was both accurate and equitable. The company emphasized that differences in scores and error rates across racial groups did not stem from bias in the algorithm itself but, rather, reflected real disparities in recidivism rates between those groups. Northpointe concluded that when analyzed using the correct classification statistics, the data did not support ProPublica's claims of racial bias. Instead, the company argued that the COMPAS tool operated fairly by adhering to objective scoring rules that were not influenced by race.

In defending its software, Northpointe framed COMPAS as a tool that achieved both predictive parity and accuracy equity, aiming to reassure stakeholders that the algorithm was neither biased nor discriminatory.

DO YOU SIDE WITH PROPUBLICA OR NORTHPOINTE IN THIS DEBATE?

As we navigate an increasingly digital world, developing a structured understanding of AI bias and its sources is essential. Achieving this requires looking beyond the surface and examining the processes involved in the development of machine learning systems. Only by "lifting the hood" can we uncover the complexities and challenges that lead to bias in AI.

To prepare for class, scan the QR code to listen to the podcast[9] about bias in COMPAS. Identify and articulate the various types of biases mentioned in the podcast. You do not need to focus on labelling them precisely—what matters is distinguishing various types of biases and reflecting on their implications.

SCAN ME

[9] Yasser Rahrovani, "ML Bias: AI in the courtroom," January 20, 2025, CaseCast, podcast, MP3 audio, 42:56, https://www.buzzsprout.com/2293124/episodes/16470939.

EXHIBIT 1: SUMMARIZING THE STATISTICS IN NORTHPOINT-PROPUBLICA DEBATE

Label	Probability	Black	White
Positivity rate	P("HR")	0.587	0.347
True positive rate (TPR)	P("HR"|R)	0.720	0.523
True negative rate (TNR)	P("LR"|noR)	0.551	0.765
False negative rate (FNR)	1-P("HR"|R)	0.280	0.477
False positive rate (FPR)	1-P("LR"|noR)	0.449	0.235

ProPublica identified that the FNR and FPR are different by race.

Label	Probability	Black	White
Base rate	P(R)	0.510	0.390
Positive predictive value (PPV)	P(R|"HR")	0.630	0.590
Negative predictive value (NPV)	P(nR|"LR")	0.650	0.710
False discovery rate (FDR)	1-P(R|"HR")	0.370	0.410
False omission rate (FOR)	1-P(nR|"LR")	0.350	0.290

Northpointe counter argued that the PPV (and, thus, FDR) are very similar by race.

Source: created by the author.
Notes: “HR” and “LR” indicate the label of “High Risk” and “Low Risk”; R and nR indicate the truth, whether the person re-offended or not within 2 years; P(x) indicates the probability of an outcome x. P(x|y) indicates the probability of an outcome x conditional on the outcome y; P(“HR”) is the probability of being labelled “High Risk” and P(“HR”|R) is the probability of being labelled “High Risk” conditional on the fact that the person will ultimately re-offend. Similarly, P(R) is the probability of recidivism and P(R|”HR”) is the probability of recidivism conditional on being labelled “High Risk”.