Fallyn Buckner

Professor Brad McHose

Data Science Ethics & Justice

16 May 2025

Word Count: **2068**

Option: **Option #1**

The COMPAS Algorithm: Perpetuating or Correcting Historical Bias

Introduction:

The Correctional Offender Management Profiling for Alternative Sanction (COMPAS) predicts the likelihood of criminal reoffending using a questionnaire of 137 questions.[1] The rating predicts reoffending based on factors that include age, sex, and criminal history.[2] The algorithm was created to give the justice system a standardized and logical mechanism for decisions. Despite this, it has faced harsh backlash for unfairly misclassifying Black defendants as high risk reoffenders compared to white defendants. This paper evaluates whether algorithmic decision making compounds historical injustices to a level that outweighs its potential to reduce human bias in judicial settings. There must be processes in place to audit the technological tools being integrated into everyday governmental proceedings.

Part I:

The COMPAS algorithm should not be used as a definitive factor in criminal sentencing for three reasons. First, the COMPAS algorithm uses proxies that correlate with racial factors. Secondly, historical and systemic racism have created unfair disadvantages for Black people. Thirdly, the likelihood statistical measure of reoffending that

[1] "Scores like this—known as risk assessments—are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts—as is the case in Fort Lauderdale—to even more fundamental decisions about defendants' freedom."Angwin, Julia, et al. "Machine Bias." ProPublica, 23 May 2016, p. 254.

[2] "The COMPASS 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." Sam Corbett-Davies et al., "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear," The Washington Post, 17 Oct. 2016, p. 2.

COMPAS algorithm attempts to compute is biased.[3] Integrating COMPAS into the justice system permanently would violate the principles of ethical justice. It is ethically invalid to determine an individual's punishment based on statistical predictions that may reflect and amplify systemic discrimination.

The biggest moral reason weighing against the permissibility of using the Northpointe algorithm to determine whether a defendant goes to jail or is let out on bail is the algorithm's ability to perpetuate historical racial injustices by having datasets filled with societal biases factored into automated decisions. I will map the specific criteria of structural discrimination to the process that the algorithm goes by in order to show how this occurs. Structural discrimination is defined as follows: when the rules of a society's major institutions reliably produce disproportionately disadvantageous outcomes for the members of certain salient social groups and the production of such outcomes is unjust, then there is structural discrimination against the members of the groups in question, apart from any direct discrimination in which the collective or individual agents of the society might engage.[4] The COMPAS algorithm is a facet of the justice system. Any standards or requirements produced from such a system are rules. Its implementation supports injustice that potentially undermines the dignity of marginalized communities. Reducing human beings to mere statistics provides a skewed perspective on personal narratives. The algorithm itself also reliably produces disproportionately disadvantageous outcomes for members of salient social groups. The salient social group in this case is Black people. Black people form a salient social group due to race. The collective identity that they share has been the basis of discrimination systematically. The production of disadvantageous outcomes is done in the form of how the measure of rearrest is computed. Proxies are the key component in the production of a score. Some examples of proxies are employment status, prior criminal history, education level, and residence information. A completely perfect algorithm computationally is still invalid if its outcome is a perpertuation of historical discrimination on a marginalized group of people.

---

[3] "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."Corbett-Davies, Sam, et al. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear." The Washington Post, 17 Oct. 2016, p. 5.

[4] Altman, Andrew. "Discrimination" Stanford Encyclopedia of Philosophy, Stanford University, 20 Apr. 2020, p. 1-33.

It may be unclear how an algorithm discriminates on the basis of race if race itself is not a factor. Slavery and historical racism cultivated a system where Black people were put at a socio-economic disadvantage at the hands of the United States government. Think of those who have been targeted for activism in the Black Panther movement, jailed for violating Jim Crow laws in the South, or wrongfully convicted simply for being caught in a sundown town. Black people endured several wrongful convictions that were tied to racism. The COMPAS algorithm will ask a blatant question such as "Was one of your parents ever sent to jail or prison?"[5] It will then proceed to rank the defendant negatively because of what happened in their family in the past, without any contextualization. Blacks, in this case, suffer presently because of their racial membership in a group historically discriminated against. This is shown in how the algorithm itself is more likely to incorrectly predict that Black defendants will commit future crimes. It is also more likely to predict wrongly that white defendants won't commit future crimes.[6] The production of such outcomes is clearly unjust in this case. Being white gives people benefits in the algorithm. Being Black causes people to face harsher treatment. Group membership affects how individuals are judged.[7] The COMPAS algorithm prevents each person from being judged on their own merit. The harm goes beyond instances of misclassification. Its implementation supports injustice that systematically undermines the dignity and individual agency of marginalized communities. Reducing human beings to mere statistics provides a skewed perspective on personal narratives. Harm itself is problematic, even when it comes from elements which appear accurate on the surface level.

This is not to say all judgements bearing negative results on a group automatically qualify as systemic discrimination. Here's an example of something that seems like it could qualify as systemic discrimination, but in actuality, fails to meet the criteria. My internship search for the summer has been filled with numerous job

---

[5] "The survey asks defendants such things as: "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school" Angwin, Julia, et al. "Machine Bias." ProPublica, 23 May 2016, p. 3.

[6]"In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways." "White defendants were mislabeled as low risk more often than black defendants." Angwin, Julia, et al. "Machine Bias." ProPublica, 23 May 2016, p. 3.

[7]"Standard accounts hold that discrimination consists of actions, practices, or policies that are—in some appropriate sense—based on the (perceived) social group to which those discriminated against belong and that the relevant groups must be socially salient in that they structure interaction in important social contexts." Altman, Andrew. "Discrimination." Stanford Encyclopedia of Philosophy, Stanford University, 20 Apr. 2020, p. 2.

descriptions; let's assume the company used AI to rank resume quality due to the increased number of applicants. The AI would determine who gets through to the hiring manager or denied via email. The job description states that candidates must be pursuing an undergraduate degree from an accredited university, graduating by December 2025 or May 2026, with an excellent academic record. The organization is trying to measure the quality of candidates for the Information Security intern program. Being a part of an accredited institution is a proxy for how well one will fulfill this criteria. Having an excellent academic record is the second proxy.

It may seem like there is some discrimination here at first glance. Some brilliant hackers lack college degrees and some information security students struggle academically due to being first generation college students. These skilled individuals might be overlooked despite their potential. The AI would automatically deny any resume that did not meet the degree requirements. These systemic barriers disproportionately impact marginalized communities, including minority groups, low income backgrounds, and those with nontraditional educational paths.

But, this case is distinctly different from circumstances like the one mentioned above. Although internship requirements place constraints on certain groups, using AI to judge and auto-reject do not fully meet the criteria of structural discrimination. First, proxies do, in fact, measure what skills are necessary to succeed in the role, even though education is more accessible to certain groups. The proxies, such as criminal history, reflect much more systemic burden than academic standards do. The education requirement is an accessory to corporate business interest. The COMPAS algorithm's computations are based on data that may not effectively carry out its objective of increasing public safety.

The second reason is the weight of one receiving an internship or not is not the same as having one's life put on pause for multiple days.[8] The latter has the potential to have negative effects on one's social responsibilities, employment, and security of family. Employment requirements offer flexibility. The third issue is that the COMPAS algorithm offers no room for alternatives to proxy measurement in the algorithm. Certifications, work experience, and projects can be enough to replace formal education in the corporate example. Job candidates receive holistic evaluation. Defendants face algorithmic judgment with no chance to present their full circumstances. COMPAS

[8] “The girls spent two nights in jail before being released on bond. ‘We literally sat there and cried’ the whole time they were in jail, Jones recalled... Jones, who had never been arrested before, was rated a medium risk...‘I went to McDonald's and a dollar store, and they all said no because of my background,’ she said. ‘It's all kind of difficult and unnecessary.’” “Angwin, Julia, et al. "Machine Bias.” ProPublica, 23 May 2016, p. 264.

algorithm amplifies systemic discrimination in a way that is distinctly different from typical proxy measurements used in other qualification assessment computations.

Part II:

The biggest moral reason to favor COMPAS is its ability to balance community safety concerns with defendants' rights in a system that aims to protect the public while providing equitable treatment in a system plagued with both inconsistency and implicit biases in human judgment. The algorithm's overarching objective is to ensure public safety.

Inherently, humans are susceptible to having biases and differing interpretations that lead to general inconsistency. Judges are not exempt. Personal experiences, prejudice, and divisive rhetoric all influence judges in different ways. Equal treatment for all is voided when the judicial system relies on traditional means. Consistency and objectivity[9] can be ensured through the use of the COMPAS algorithm. Providing a statistical measure for likelihood of reoffending as opposed to relying solely on human judgement provides a consistent amount of reliability. The instances of personal discrimination in cases will be minimized through using the algorithm. A fair standard for judgment is provided, even though the algorithm is not one hundred percent neutral in terms of calculation. [10]

The concept of harm is key in unpacking the COMPAS algorithm. Lippert-Rasmussen provides a framework to evaluate harm and the extent to which it impacts humans. He defines non-moralized intrapersonal harm as "X harms Y if X brings it about that Y is worse off than Y would have been had X been unable to exercise

---

[9]"More plausibly and more generally, every substitution of more rather than less individuation increases the costs of scrutiny, and also increases the possibility of error. Individuation requires individuators, and often the errors consequent on the use of imperfect proxies will, in some contexts, be less than" Schauer, Frederick. "Statistical (And Non-Statistical) Discrimination." The Routledge Handbook of the Ethics of Discrimination, edited by Kasper Lippert-Rasmussen, Routledge, 2017, p. 51.

[10]"That some people believe that a statistical (or probabilistic) relationship exists between some proxy and what it is a proxy for does not mean that they are correct in so believing. Indeed, much of the history of pernicious discrimination is a history of beliefs about the existence of some supposedly valid statistical instrumental relationship that turns out to have no sound empirical basis whatsoever." Schauer, Frederick. "Statistical (And Non-Statistical) Discrimination." The Routledge Handbook of the Ethics of Discrimination, edited by Kasper Lippert-Rasmussen, Routledge, 2017, p. 46.

his agency in the situation"[11] and moralized harm as "X harms Y if X brings it about that the gap between the level of benefits Y enjoys and the level Y ought to enjoy becomes greater than it would have been had X been unable to exercise his agency in the situation."[12] The COMPAS algorithm fits the criteria of causing non moralized harm. The harm would be on some defendants who receive higher risk scores, and as a result, more harsh treatment. Though, the algorithm offers the reduction of moralized harm spread across the entire judicial system. It offers the benefit of decreasing biases and inconsistencies. The application of standards and personal biases by judges creates a form of moralized harm. Defendants have a gap between actual treatment and how they "ought to be treated"based on the framework. COMPAS narrows the gap spread across the system as a whole. This is the case even though the weight faced on an individual basis seems unfair on the surface.[13] The existence of such raises an essential point: what factors matter the most in algorithmic justice? Some will surely argue that helping the most people (entire system) at the cost of a few people violates the courts requirement to treat each person as an individual.

Those in favor of the algorithm claim that it is a consistent standard across all groups, despite consistency itself not resolving every ethical concern. There is no way to ensure an unbiased and consistent algorithmic standard. This system does, though, take the only available data provided and attempt to standardize it to get consistency. Society is simply using technology, working with what they have, to attempt to make a strong improvement.[14] The

---

[11]"Non-moralized intrapersonal harm: X harms Y if, and only if, X brings it about that Y is worse off than Y would have been had X been unable to exercise his agency in the situation." Lippert-Rasmussen, Kasper. "Is There a Duty Not to Compound Injustice?" Law and Philosophy, vol. 42, 2023, p. 107.

[12] "Moralized interpersonal harm: X harms Y if, and only if, X brings it about that the gap between how Z's and Y's levels of advantage ought to be in comparison with one another, on the one hand, and how it in fact is, on the other, is greater than it would have been had X been unable to exercise his agency in the situation." Lippert-Rasmussen, Kasper. "Is There a Duty Not to Compound Injustice?" Law and Philosophy, vol. 42, 2023, p. 107.

[13] "Perhaps the children of female victims of domestic violence are indirectly wronged, and thus indirect victims of injustice and harmed when their mothers are subjected to domestic abuse… On the narrow victim view, neither in the case of the children nor in the case of the shop owners is there any duty not to compound the harm or (in the case of the children, who are indirect victims of injustice) not to compound the injustice done to their mothers." Lippert-Rasmussen, Kasper. "Is There a Duty Not to Compound Injustice?" Law and Philosophy, vol. 42, 2023, p. 98.

[14] "Drawing the analogy to social biases, our social environment is shaped by a variety of racist and discriminatory practices. So, if a machine learning program is aiming to make predictions in line with our current social landscape—i.e., built to navigate our current social environment—it necessarily adopts and utilizes assumptions that mimic patterns presently existing in the data on which it is trained. Thus, assumptions that encode problematic stereotypes will inevitably be adopted and perpetuated by machine learning programs. Johnson, Gabbrielle M. "Algorithmic Bias: On the Implicit Biases of Social Technology." Synthese, vol. 198, 2021, p. 9951.

ability to protect people from rulings where judges may be discriminatory and keep the same standard among trials can help to improve data in the future with continued use. Technology must evolve in the current state to benefit the people it was created for. The algorithm is beneficial in this current climate. Denying use of it is denying the opportunity to evolve with technology and create real improvement in the world.

Part III:

There are several alternative approaches that should be compared when trying to unpack the current implementation of the COMPAS algorithm. Firstly, void of algorithm assessment, society would be solely reliant upon human judgement. Historically, human judgement is not always consistent. It opens the room for bias while preserving individual assessment of every case. Secondly, consistency would be maximized through using COMPAS to automatically determine verdicts void of judicial input. Though, there would be an elimination of contextualization that may be relevant ethically. Finally, integrating both approaches by having a judge use the algorithmic rating amongst other factors provides the benefit of both options; it also includes the risk of algorithmic bias being an influence on verdicts.

The moral concerns about perpetuating discrimination outweigh the potential benefits of the COMPAS algorithm. While COMPAS was developed to complement judicial decision making by providing standardized risk assessments, its operation in practice creates more harm than good. Algorithmic tools built on datasets containing historical biases amplify historical wrongs rather than mitigate systemic injustices. The objective of ensuring public safety does not justify a system that risks reinforcing discriminatory patterns in the criminal justice system. Human judges can already evaluate criminal histories directly.

The most optimal approach of the three is the third. It is not fully perfect, but becomes preferable if the judges are aware of the algorithm's actual capabilities and flaws. Substituting contextualized human judgment with statistically computed risk scores ultimately undermines the justice system's commitment to equitable treatment under the law.

Works Cited

Altman, Andrew. “Discrimination” Stanford Encyclopedia of Philosophy, Stanford University, 20 Apr. 2020, p. 1-33.

Angwin, Julia, et al. “Machine Bias.” ProPublica, 23 May 2016.

Corbett-Davies, Sam, et al. “A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.” The Washington Post, 17 Oct. 2016.

Drösser, Christoph. “In Order Not to Discriminate, We Might Have to Discriminate.” Simons Institute for the Theory of Computing, 22 Dec. 2017.

Lippert-Rasmussen, Kasper. “Is There a Duty Not to Compound Injustice?” Law and Philosophy, vol. 42, 2023, p. 93-113.

Schauer, Frederick. “Statistical (And Non-Statistical) Discrimination.” The Routledge Handbook of the Ethics of Discrimination, edited by Kasper Lippert-Rasmussen, Routledge, 2017, p. 42-53.

Johnson, Gabbrielle M. “Algorithmic Bias: On the Implicit Biases of Social Technology.” Synthese, vol. 198, 2021, p. 9941-9961.