Optimizing the Criminal Justice System with Algorithms: A Case Study of Technological Politics

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Alma Rivera

February 21, 2019

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____

Approved: _____ Date: _____
Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

Predictive algorithms are becoming increasingly prevalent in the criminal justice system, and are used for various purposes including crime prediction and defendant risk assessment. In 1998, one such algorithm, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was developed by Northpointe, Inc. (*Injustice Ex Machina*, 2019). COMPAS assigns a risk score to a given defendant based on factors such as prior offenses and a defendant's attributes and qualities. This score is then handed to a judge to determine the defendant's bond amount and even criminal sentencing (Julia Angwin, 2016). In 2016 an investigation led by journalists at ProPublica analyzed the data obtained from a jurisdiction in Florida and found that the predictions made by COMPAS were unreliable and biased against black defendants. Black defendants were incorrectly predicted to reoffend twice as often as white defendants and white defendants were much more likely than black defendants to be labeled low risk, yet go on to reoffend (Park, 2019).

The investigation carried out by ProPublica sparked a debate in the fields of computer science and data science and prompted a response by the creators of COMPAS in defense of the algorithm's fairness (Washington, 2019). The debate centered around the question of defining fairness, and how it could be quantified by algorithms. Although it is a valid question to address, it is equally as important to understand how technologies like COMPAS work to maintain and support power relations between groups, privileging certain groups while marginalizing others. If we continue to center the conversation around the attempts of COMPAS to quantify fairness with algorithmic methods, we will be unable to recognize the social and political dimensions of such technologies and miss out on the opportunity to critically analyze the work they do.

Using the technological politics framework, I argue that the COMPAS algorithm injects bias into the criminal justice system and thus, expresses and shapes power relations by privileging white defendants while further marginalizing black defendants**.** Specifically, I will examine the methods used by COMPAS to calculate risk scores and the way that COMPAS is integrated in the file of a defendant in the courtroom.

Background

COMPAS is a proprietary algorithm owned by Northpointe that was introduced in 1998 and has been used to assess over 1 million defendants since. It is an algorithm meant to predict a defendant's risk of committing another crime within 2 years of assessment and is derived from a questionnaire of 137 questions filled in based on the defendant's answers and the defendant's past criminal history. The possible score range is 1-10, with 10 being the highest risk and 1 being the lowest risk. Although the calculations used to arrive at the score are not publicly disclosed, Northpointe has repeatedly emphasized that race is not one of the questions that is asked and thus, their algorithm cannot be racially biased. However, in 2016 ProPublica published a study conducted on data obtained from Broward County, Florida on more than 7,000 people assessed with COMPAS. They statistically analyzed the acquired data and found that black defendants were disproportionately more likely to be scored at "high risk" for committing a future violent crime than their white counterparts (77 percent more likely). They were also 45 percent more likely to be predicted to commit a future crime of any kind than white defendants (Park, 2019). This controversial article sparked a debate among critics and defendants of COMPAS that would address the algorithm's "fairness."

Literature Review

Following the publication by ProPublica, within 18 months, there were over 200 academic papers that cited the article and produced findings of their own using alternate methods. The conversation was mainly centered around the definition of fairness and the mathematical achievement/validity of it. Various definitions of fairness were proposed and given mathematical relevance in order to apply statistical models and come to a conclusion about the effectiveness of risk assessment algorithms. Despite the breadth and extent of such published works, they generally fail to address the social and political work done by COMPAS, and other similar algorithms.

In *Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment* Laurel Eckhouse breaks down concerns about algorithmic bias into three layers: fair algorithms, data quality, and fundamental conceptual problems with the fairness of data-driven decisions. In the first layer, Eckhouse explains that there are three basic measures of fairness that are mathematically impossible to attain simultaneously. The first measure is based on calibration across groups, meaning that if 30% of black defendants are predicted to commit a crime, and 30% of white defendants are predicted to commit a crime, they should both offend at the same rate. Secondly, false positive rates should be the same across groups. Conversely, the third measure is based on the algorithm's ability to correctly predict false negative rates across groups. Since these measures cannot be balanced equally, Eckhouse suggests that the tradeoffs in fairness should be made explicit by an algorithm. In the data quality layer, Eckhouse focuses on the quality of the data, specifically on the use of sample sizes and the tradeoffs between small and large sample sizes. A large enough sample size may show statistically significant differences

between groups, however a small sample size may not exacerbate these differences. In the final layer, Eckhouse sheds light on the problem of using prior data about other people to evaluate a particular defendant's risk and questions the constitutionality of the method (Eckhouse et al., 2019). Although Eckhouse successfully unpacks the complexity of assessing an algorithm, she fails to address what the social and political consequences of a failing algorithm like COMPAS might be.

In *Inherent Trade-Offs in the Fair Determination of Risk Scores* Jon Kleinberg defines three similar fairness properties for risk assignments: calibration within groups, balance for the negative class, and balance for the positive class. Calibration within groups requires that the scores accurately reflect outcomes, regardless of group. Balance for negative and positive classes stipulates that two individuals from different groups should be comparably treated if their behavior in the future is also comparable. The first point they state is that from a mathematical standpoint, all three of the fairness properties can be achieved only in the two cases of perfect prediction or equal base rates, meaning that any two groups have equal amounts of members in negative and positive classes. They further explain that in cases that are any more complex than those two base cases, some fairness condition will inherently suffer. To address the ability of COMPAS to estimate risk, they show that a perfect prediction is mathematically impossible, and they partly explain this shortcoming of the algorithm as a simple difference in base rates between groups. Kleinberg parallels COMPAS to determining the risk that a person is a carrier of a disease, given that women carriers are higher in proportion to men carriers. He points out that any test developed will have a trade-off among the three fairness properties, and that this is a result of differing base rates (Kleinberg, 2016). In this article, Kleinberg explores the definition

of fairness and how it can be achieved when incorporating algorithms like COMPAS; however, he over-simplifies the problem and unjustly attempts to reduce it to a matter of differing base rates. By doing such over-simplification, he is able to avoid addressing the underlying bias that affects the outcomes of such predictive algorithms, and the larger consequences that biased algorithms carry into their scope of reach.

It is undoubtedly important to investigate the mathematical abilities and limitations of COMPAS and similar algorithms; however, it is equally critical to weigh the power and privilege wielded by those same algorithms. Currently, the debate continues to be centered around finding an adequate definition of fairness, and translating that in a mathematically efficient manner to predictive algorithms like COMPAS. This paper will seek to fill in some of the gaps in understanding and provide not only an analysis of the way that COMPAS functions, but also an evaluation through the lens of technological politics to elucidate the relationships of power and privilege expressed by COMPAS.

Conceptual Framework

My analysis of COMPAS draws on Langdon Winner's Theory of Technological Politics to understand how COMPAS privileges white defendants over black defendants. Technological politics describes technical devices, systems, or artifacts which appear to require or to be strongly compatible with particular kinds of power dynamics that advantage certain groups over others (Winner, 1980).

In his theory, Winner argues that technological artifacts have political meaning. It is important to note that Winner specifically defines politics as "arrangements of power and authority in human associations as well as the activities that take place within those

arrangements" (Winner, 1980). One way that Winner argues that technologies contain political properties is that they function as a way of settling an issue in a particular community. Therefore, technologies can support existing forms of social order or create new forms of social order. He further expounds that in these cases, "the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results counted as wonderful breakthroughs by some social interests and crushing setbacks by others" (Winner, 1980). Thus, for these technologies, intentionality behind the design cannot be easily ascribed, but rather these technologies should be weighed in parallel to existing forms of social order. Only then can these technologies be understood in terms of the social and political work they do.

Under this theory, the reader will be able to understand that COMPAS is a technology that is arguably not intentionally designed to be racially biased; however, it is inevitably a relic of long-standing forms of social order that privilege whites over blacks. In what follows, I will analyze the ways that COMPAS expresses social and political relationships of power by investigating the COMPAS risk score calculation methods and explaining the role that COMPAS has in the courtroom.

Analysis

The COMPAS algorithm developed by Northpointe is a technology that is inherently political under the technological politics framework. It privileges and benefits white defendants while marginalizing and disenfranchising black defendants, and I will explore this by analyzing the risk score calculation methods and the role that COMPAS plays in the courtroom. Although COMPAS may not have been intentionally designed to shape relationships of power, it inevitably does so because of the system within which it is embedded, a system that has a history of

reflecting and reifying white privilege as a form of social order. The following paragraphs will explore each of these aspects individually and present key evidence and details that indicate how COMPAS is able to express relationships of social and political power.

Risk Score Calculation Methods

One mechanism by which COMPAS favors white defendants over black defendants is the racial bias present in the risk score calculation methods themselves. COMPAS is a proprietary algorithm, which means its methodology is not made public and it essentially functions as a black box model; thus, its process cannot be verified by outside sources. However, according to information put out by Northpointe, COMPAS is intended to be more comprehensive than other risk assessment tools by taking into account a defendant's social isolation, leisure time, and family criminality. In total, COMPAS is based on 15 factors, and it measures risk based on theories of criminality including "criminal personality," "social isolation," "substance abuse," and "residence/stability" (Julia Angwin, 2016). When a defendant is arrested, he/she is prompted to answer a 137 question survey. Some of the questions are answered by the defendant themselves, and some are filled in by the police officer or by the defendant's past criminal history. This questionnaire makes up one third of the defendant's COMPAS assessment, with the other two thirds composed of information collected from official records and an interview with the defendant. Based on the results of the COMPAS assessment, he/she is assigned three different scores ranging from 1-10: risk of recidivism, risk of violent recidivism, and risk of failure to appear (Bloomberg et al., 2010). Recidivism is the act of reoffending within two years of the initial assessment. Although the questions are meant to be unbiased and objective, some of

the questions may inherently inject bias into the questionnaire. Below are some example questions pulled from the COMPAS questionnaire:

a. How often do you barely have enough money to get by?
b. How hard is it for you to find a job ABOVE minimum wage compared to others?
c. How many of your friends have ever been arrested?
d. Is this person a suspected gang member? [answered by the police officer]

Although none of the questions explicitly ask about race or socioeconomic status, they may serve as direct proxies for asking about such characteristics. (Brackey, 2019). Question a and b, for example, implicitly ask about a defendant's financial status by asking about jobs, salary, and financial resources. Seemingly, these questions are unbiased, however, when we consider the historical racial wage gap in the United States, it is simple to see how black defendentants are at a disadvantage. In 2017, black men were paid only 69.7 cents on the white male dollar and black women were paid 60.8 cents on the white male dollar (Gould et al., 2018). There has been little progress on closing gender and racial wage gaps since 2000, with the exception of white women, whose gap has since narrowed. Thus, although questions about a defendant's financial resources may not directly ask about race, they are politically charged because of historical gaps in earnings based on race. Question c on the other hand, asks the defendant about personal ties to past offenders that have been arrested. This is a problematic question to ask because of the racial disparities in incarceration rates and the United State's problem with mass incarceration. Mass incarceration was the result of law enforcement and sentencing policy changes that began with the War on Drugs in the 1980s. This has translated to dramatic increases in the prison population and longer jail terms, with the U.S. jail population accounting for nearly 25% of the world's

prison population (*Criminal Justice Facts*, 2019). It is important to note that black and hispanic communities have been marginalized and targeted by these policy changes, and this is reflected in their disproportionate representation in the jail population. This begins with increased police contact due to higher policing rates in communities of color, which then inherently leads to higher arrest rates as evidenced by black defendants currently comprising 27% of all arrests in the U.S. Furthermore, statistically one out of three black boys can be expected to go to jail in their lifetime compared to one out of every seventeen white boys (*Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System*, 2018). Thus, question c is fundamentally unfair to black defendants since the likelihood of people in their community being arrested is systematically rigged. Finally, question d is a direct injection of the police officer's potential bias about the defendant because it asks the police officer to make a judgement about the defendant based on what they perceive about him/her.

Ostensibly these questions appear neutral and unbiased and thus, "[the questions] are subject to much more deferential standards than would be clearly race-based questions" (Kehl & Kessler, 2017). Trying to prove that the COMPAS questionnaire is racially biased would be extremely difficult because an explicit demonstration of discriminatory intent would have to be available. This would be difficult to prove because the question structure masks it well.

As demonstrated by the evidence above, the COMPAS questionnaire is embedded within a particular economic structure and criminal justice system that situates black defendants at a disadvantage compared to their white defendant counterparts. Thus, the outcomes of COMPAS are not unbiased; but rather, they appear to be compatible with a racial power dynamic that privileges white defendants while marginalizing black defendants.

How COMPAS is Used in the Courtroom

Another mechanism by which COMPAS is racially biased in favor of white defendants is the way that COMPAS is inconsistently used in the courtroom. The COMPAS score is not made available to defendants, instead it is presented to the judge as a component of the defendant's profile during sentencing. Consequently, many defendants do not realize the importance of their score until they are facing a possible sentencing in court. Judges are theoretically intended to use the COMPAS scores of a defendant only to determine probation and treatment program eligibility. Nonetheless, judges have previously cited COMPAS scores in their sentencing decisions. One of the most notable cases of this occurrence is the State v. Loomis case of 2016 where Wisconsin charged Eric Loomis with five criminal counts relating to a drive-by shooting. Loomis had been assigned a COMPAS score that identified him as a high risk to the community. In the following court excerpt from Loomis's sentencing, the judge cites Loomis's COMPAS score before handing him a sentence of eight years and six months in prison, "You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend. " (*State v. Loomis*, 2016). By using COMPAS scores as a reason for sentencing, the judge is mishandling a tool that was not meant for sentencing purposes. Furthemore, COMPAS scores have been shown to be unreliable in predicting future crimes. For example, in a group of 7,000 people arrested, only 20% of those predicted to commit violent crimes actually went on to commit violent crimes in the future. COMPAS is also almost twice as likely to wrongly flag black defendants as future

criminals compared to white defendants (Julia Angwin, 2016). This means that a flawed algorithm that fails to provide equitable results across race could be used to stake the future of defendants beyond probation and treatment programs.

It was also documented that during the hearing of Loomis, the court had to stop to repeatedly ask the counsels about how the COMPAS scores worked. In the following excerpt from the court transcript, an expert witness, Dr. David Thompson testifies about the court's use of COMPAS, "The Court does not know how the COMPAS compares that individual's history with the population that it's comparing them with. The Court doesn't even know whether that population is a Wisconsin population, a New York population, a California population.... There's all kinds of information that the court doesn't have, and what we're doing is we're mis-informing the court when we put these graphs in front of them and let them use it for sentence" (*State v. Loomis*, 2016). Dr. Thompson is expressing a valid concern that judges are not fully informed about the way that COMPAS calculates a score about a defendant. This is a crucial failure within the justice system, because without context and proper information about the methods COMPAS uses, the standalone score becomes a dangerous tool that is available for use at the prerogative of the prosecutors and officials. Furthemore, it lends itself to automation bias in the hands of judges which could cause judges to over-rely on COMPAS scores because of the automation appeal that comes with technologies like COMPAS. Automation bias is "the tendency to ascribe greater power and authority to automated aids than to other sources of advice" (Park, 2019). This may be of concern with the use of COMPAS in the courtroom since judges are often faced with complex cases and significant time constraints. This could potentially lead to judges placing

undue weight on COMPAS scores and as a result, may rob defendants of their due process rights and perpetuate racial bias.

I have argued that COMPAS is an algorithm that injects bias into the criminal justice system that puts black defendants at a disadvantage compared to white defendants. However, some may argue that even without the use of COMPAS in the criminal justice system, there is still a "black box" making decisions, referring to judges making decisions about defendants in a non-transparent way (Rudin et al., 2019). Although this is true to some extent, the key difference between a judge and an algorithm like COMPAS is that defendants can contest a judge's ruling. In fact, 10.9% of filed cases are appealed and of those, about 9% result in reversal (Eisenberg, 2004). By contrast, defendants cannot question COMPAS due to its proprietary methods that completely obscure the decision-making process. In addition, COMPAS scores legitimize biased, potentially flawed scores in the courtroom in a way that a judge's opinion does not because of the perceived superiority of "artificial intelligence". Defendants are not made aware of their COMPAS score, or of the consequences of having a poor COMPAS score, and moreover, COMPAS methods cannot be verified by outside sources (Julia Angwin, 2016). Although a judge may still have unconscious bias that plays a role in decision making, there are several opportunities for this bias to be checked and questioned, a benefit that is not afforded by COMPAS.

Conclusion

The developers of COMPAS may not have intentionally designed it to be a political technology, however, COMPAS is embedded within the criminal justice system, which makes it inherently political. The criminal justice system is largely and historically based on racial

injustices, a statement that is time and time again reinforced by various events and statistics: Jim Crow laws that lasted from 1870-1965, the “War on Drugs” of the 1980s and 1990s that was publicly admitted to be aimed at minorities, juvenile drug arrests decreasing 28% for whites from 1980 to 1993 and increasing 231% for blacks during the same time period, etc. (Rosich, 2007). Therefore, COMPAS, which uses past data to predict future outcomes of crime, inherently becomes a tool that reinforces white privilege as a social order within a biased criminal justice system.

Engineers who produce algorithms like COMPAS should understand that their designs will not be politically neutral because they are created within existing forms of social order. Engineering work, although not typically considered a form of social and political work, has the ability to do such work because it forms part of a larger system that pushes certain political agendas like maintaining white privilege. Thus, engineers are not only responsible for designing solutions that provide solutions and optimize processes, but they are also responsible for making sure their work functions equitably. Although this may not be entirely possible from a mathematical standpoint, engineers must always strive towards improving their designs to minimize the negative consequences of their work.

Word Count: 3777

References

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks,. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bloomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *Validation of the COMPAS risk assessment classification instrument*. Florida State University. http://criminology.fsu.edu/wp-content/uploads/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf

Brackey, A. (2019). *Analysis of racial bias in Northpointe's COMPAS Algorithm—ProQuest*. https://search.proquest.com/openview/492b784291d8ff8e65dfdc3b7fe92206/1?pq-origsite=gscholar&cbl=18750&diss=y

Criminal Justice Facts. (2019). The Sentencing Project. https://www.sentencingproject.org/criminal-justice-facts/

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: a unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, *46*(2), 185–209. https://doi.org/10.1177/0093854818811379

Eisenberg, T. (2004). Appeal rates and outcomes in tried and nontried cases: further exploration of anti-plaintiff appellate outcomes. *Journal of Empirical Legal Studies*, *1*(3), 659–688. https://doi.org/10.1111/j.1740-1461.2004.00019.x

Kehl, D. L., & Kessler, S. A. (2017). *Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing*. https://dash.harvard.edu/handle/1/33746041

Kleinberg, J. (2016). Inherent trade-offs in the fair determination of risk scores. *ACM SIGMETRICS Performance Evaluation Review, 46(1).*

https://dl.acm.org/doi/10.1145/3292040.3219634

Park, A. (2019, February 19). *Injustice ex machina: predictive algorithms in criminal sentencing*. UCLA Law Review. https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/

Report to the united nations on racial disparities in the U.S. criminal justice system. (2018). The Sentencing Project. https://www.sentencingproject.org/publications/un-report-on-racial-disparities/

Rosich, K. (2007). *Race, ethnicity, and the criminal justice system*. Washington, DC: American Sociological Association. http://asanet.org

State v. Loomis. (2016). Supreme Court of Wisconsin. https://casetext.com/case/state-v-loomis-22/?PHONE_NUMBER_GROUP=P&NEW_CASE_PAGE=N

Washington, A. L. (2019). *How to argue with an algorithm: lessons from the COMPAS ProPublica Debate* (SSRN Scholarly Paper ID 3357874). Social Science Research Network. https://papers.ssrn.com/abstract=3357874

Winner, L. (1980). *Do artifacts have politics?* Daedalus 109. https://www.jstor.org/stable/20024652?origin=JSTOR-pdf&seq=1#metadata_info_tab_contents