In Search of True North

An Examination of COMPAS and Bias in Criminal Risk-Assessment

April 13, 2019

Anna Berman, Lidia Azucena Morales Vasquez, Sicong Zhao, Viggy Kumaresan, Yifei Wang

ABSTRACT

Abstract Text

Introduction

Consider two individuals; both charged with petty crimes. The first is Brisha Borden, an 18-year-old girl with no criminal record who attempted to steal a 6-year old's bike and razor scooter on her way to pick up her god-sister from school. The second is Vernon Prater, a 41-year-old man previously convicted of multiple attempted and completed arm robberies who has spent five years in jail for his crimes. He attempted to shoplift $86.35 worth of tools from Home Depot. Which of these individuals would you say is more likely to commit a future crime? According to COMPAS, an automatic risk assessment tool used in judicial systems across the country, Borden, who is black, is a high risk, while Prater, who is white, is a low risk (Angwin et al., 2016).

Advocates of machine learning argue that algorithms, unlike humans, have the ability to base decisions entirely on logic and therefore have the power to eliminate human biases from decision making entirely. However, studies have shown that this is far from the case (Barocas & Selbst, 2016; Mittelstadt et al., 2016; Caliskan & Narayanan, 2017). Without careful precautions, algorithmic decision making has been shown to result in inconsistent effects across classes of people - in other words, the algorithm shows unfairness between classes of people. "If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination" (Caliskan & Narayanan, 2017). Either with or without intention, a steadfast reliance on data for decision making has the potential to perpetuate human bias

"If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination." (Caliskan & Narayanan, 2017)

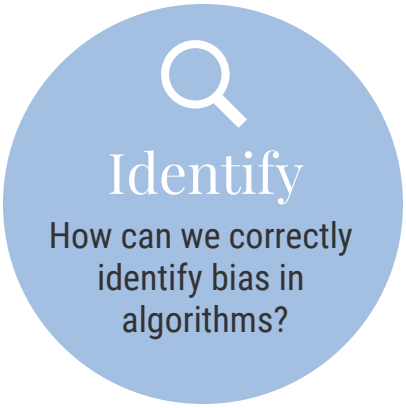
In the United States, regulations are in place that prohibit discrimination based on race, skin color, religion, sex, or national origin. However, excluding membership to a protected class as an input to any machine learning algorithm is more difficult than it may originally seem. Combinations of seemingly innocuous attributes can act as proxies for protected peoples that are much harder to control. For example, an attribute as harmless as postal code can be a reliable proxy for race (Consumer Financial Protection Bureau, 2014; Mittelstadt et al., 2016). Because the variable combinations that lead to this outcome are not necessarily straightforward to identify deconstruct, the discrimination can effectively be masked. The effect is only intensified as methods become increasingly complex and so opaque that data controllers are unable to comprehend how a given output is created (Mittelstadt et al., 2016).

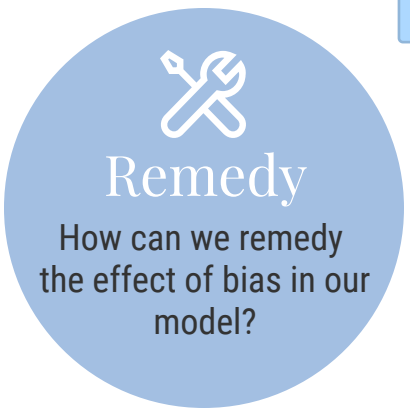
Algorithmic bias becomes even more problematic when applied in life-altering circumstances, such as in the criminal justice system. With current spending in the US reaching $13.7 billion on jailing and approximately $443,000 on pre-trial expenses each day, more and more justice systems are turning to algorithms to aid in decision-making, especially as the pre-trial burden in the United States continues to increase (Neufeld, 2017). Specifically, many courts have integrated risk assessment algorithms to help determine sentencing and bail for individuals (Angwin et al., 2016; Gershgorn, 2018; Yong, 2018). Unfortunately, there is little regulation or standard or rigor when it comes to assessing the accuracy of these methods. Thus, recommendations from algorithms already instituted in justice systems are often heavily skewed to target African-Americans (Angwin et al., 2016; Gershgorn, 2018).

COMPAS, the Correctional Offender Management Profiling for Alternative Sanctions, is one such algorithmic risk assessment. Developed by a for-profit company called Equivant (formerly Northpointe), COMPAS was designed to predict a defendant's risk of committing future crimes. The algorithm takes in answers to a 137-item questionnaire and produces a risk-score scaled from one to ten, one being the least risky and ten being the most risky. Importantly, nowhere in the questionnaire, is there mention of race. Nevertheless, despite explicit exclusion of race as an input into prediction, outcomes differ by race. Specifically, prediction error, assigning a high risk-score to a defendant that did not ultimately reoffend and assigning low a risk-score to a defendant that ultimately did reoffend are unfavorably higher for black defendants (Angwin et al., 2016; Gershgorn, 2018; Yong, 2018).

Problems

Within the criminal justice context specifically, this paper will focus on three ethical issues that we have identified three concerns:





Identify: How can we identify bias in algorithms?

There are several difficulties associated with identifying bias in criminal justice algorithms. For the most part, these algorithms are created by for-profit companies that use proprietary knowledge, and the companies are usually unwilling to share their knowledge in order to preserve their bottom line. This ultimately means that these algorithms are extremely difficult to judge or assess, and we are only able to identify any issues after the harm is already done.

In the case of the COMPAS algorithm, all we know is that it "scores a person's risk of recidivism and assesses their "needs" based on 130-plus items including criminal history, age, gender and other information, such as whether their mother was ever arrested or whether they have trouble paying bills."1 ProPublica, a non-profit think tank, found that these factors were causing the algorithm to heavily discriminate against African-Americans, even to the point where African-Americans with no criminal record were being given higher risk scores than whites who did have a criminal record. Their methodology highlights a good example of identifying bias in a criminal justice algorithm, and we will show the potential of their methods for other solutions later.

Another difficulty with identifying bias is the nature of biased data collection. Rashida Richardson, policy director for AI Now, a nonprofit think tank dedicated to studying the societal impact of AI, says "A lot of these criminal justice algorithmic-based systems are relying on data collected through the criminal justice system. Police officers, probation officers, judges, none of the actors in the criminal justice system are data scientists, so you have data collection that's flawed with a lot of the same biases as the criminal justice system" (Gershgorn, 2018). Since there are no current checks or balances in place to monitor this data collection, there is often inherent bias in the collection that will affect the ultimate prediction made by the algorithm. Identifying the bias in these data collection mechanisms is another difficult question, but we will discuss proposed solutions to this problem later in the paper.

Quantify: How do we quantify or measure level of bias in our information?

As we have mentioned before, transparency is a barrier for bias, specially for quantification purposes. This not only includes the algorithm itself but the data in which it was trained. Unlike the identification of bias, quantification requires the full knowledge of the algorithm structures. It is necessary to know how the algorithm treat and manipulate each features and what is their impact on the outcomes before making any quantification. However, these algorithms are rarely fully transparent because of their profitable nature. ~~Nevertheless,~~ at the same time this lack of transparency prevents bad guys from cracking this algorithm, by intentionally changing to meet some highly weighted good features and commit crimes.

Moreover, quantification of algorithms requires the full knowledge of data, in order to infer where and to what extent the bias came from. Algorithms do not create new things but only generalize existed things, so it is possible for an algorithm to generate bias from biased et even if the algorithm itself is well-examined. However, crime-related data are always se and private because it contains personal identities and sensitive information. This nature make it hard to be fully examined by others.

We consider that algorithms that could have a high impact consequence in an individual, as those used in criminal justice, either should be transparent or need to be audit. In that sense, it is important to quantify in some way in what extent these algorithms are being biased. Once a model has being identified as biased, the importance of quantifying it is to have a better understanding of the unintended consequences or the impact of using it, therefore there should be a more objective metric. Therefore, explicit mathematical definitions of fairness should be proposed, before attempting a quantification of bias.

Remedy: How can we remedy the effect of bias in our model?

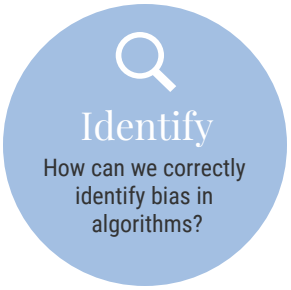
Once we have identified and quantified bias in an algorithm, then we must figure out how to use this information in order to improve/rectify the recommendations that are made. This is a difficult problem to address, since this often times involves addressing past mistakes that were made. In the case of COMPAS and other criminal justice algorithms, these mistakes can cause irreparable damage to people's lives. In this case, we are only able to deal with the ramifications of the model's mistakes: in 2016, an inmate in upstate New York was mistakenly denied parole, and in 2017, a man was released from a San Francisco jail based on a miscalculation, days before he allegedly committed a murder. These are just two cases of the effects of model bias, indicating that we need to identify bias in these algorithms before they affect lives.

Not only could individuals spend significant time in jail, but they also have to struggle with finding jobs afterwards, since they will carry this criminal record with them for the rest of their lives. Some non-profit organizations, like the Innocence Project and the Exoneration Initiative, are working to remedy the effects of these models in a broader context, but we will propose solutions that specifically address the remediation of biased algorithmic decisions in the criminal justice system.

This remediation also involves another aspect, which is concerned with updating the algorithm to remove any discriminatory effects that are found. This is a difficult task not only because the for-profit companies that are in charge of these algorithms are unwilling to alter their proprietary formulas, but also because they have no guidance on how to do so, even if they were willing. Additionally, these bias adjustments often come at the cost of accuracy, which is the primary metric for which the algorithms are being judged on.

Solutions

These three problem areas (and their respective proposed solutions in the next section) lay out a pathway for the development of future criminal justice algorithms. By running these ethical checks at different points throughout the creation, implementation, and assessment of these algorithms, we hope that we can move towards a system that is fair and just. This is the ultimate ideal that we should be holding ourselves to in the criminal justice system, and while algorithms are seemingly more complex (in certain ways) than other components of the criminal justice system, we owe it to ourselves to ensure that this system is equitable.



We identified proprietary knowledge and lack of transparency as one of the problems associated with identifying bias in these criminal justice algorithms. Despite pushback from justice advocates and the larger community, these companies still refuse to share the information that goes into their predictions, for they believe they would be at-risk of harming their bottom line. For this reason, we propose a proactive approach to developing these algorithms that involves external third-party auditing and public transparency throughout the process. We believe that the enforcement of these two factors will go a long way towards addressing the issue of hidden knowledge that we raised.

External audits from ethical and legal organizations would ensure that outside eyes are able to look at all of the working mechanisms in the algorithm, evaluating them for any potential bias that the original creator unintentionally included. By having different stakeholders involved, the probability of identifying bias increases dramatically, and this is to the benefit of everyone. In the case of COMPAS, legislators have already scheduled a review for 2023 that checks the new system for bias. This is a good step, but four years of holding people in jail before their trial is still a long time, which is why we are advocating for more proactive measures.

This concept is also linked to transparency, since certain individuals will be allowed to view the workings of the algorithm, but we also propose taking it a step further. If these algorithms are going to be evaluated properly, then a hidden prediction is not fair to the general public. As Cynthia Rudin, associate professor at Duke University, says "Transparent models are strictly better for the justice system in every possible way" (Rudin, 2018). Rudin, along with professors at Harvard University and University of California-Berkeley, developed their own machine learning model called CORELS that was based purely off of publicly available data. They created this model to show that a public alternative to a system like COMPAS could still make accurate predictions while being transparent (their model performed just as well as COMPAS or other state-of-the-art models on both blacks and whites). With this evidence, we strongly suggest that the move to transparency in criminal justice algorithms is a clear solution to the proposed problem.

For future algorithms, there is various legislation being created that is focused on this topic of transparency. In California, state senator Robert Hertzberg has already committed to introducing new legislation that will ensure transparency of risk assessment tools, which would affect future models like COMPAS. And in 2018, New York City, passed the country's first legislation to subject such algorithms to greater public scrutiny. Known as the Algorithmic Accountability Bill, it established a task force to examine how algorithms are used by city agencies (Gershgorn, 2018). This type of proactive legislation is exactly what we are proposing and advocating for, since it will ensure that bias is identified early in the process for algorithms like COMPAS, before individuals are harmed.

Bias could arise from the whole process of data collection, so we proposed a series of in-place checks at each stage of data collection to prevent biased data. At the very beginning, the data collection team should understand where the data comes from and how it is being collected. After ensuring that the data is as neutral as possible, the team should clarify what kinds of assumptions they are making for using this data. Once the data is finally fed into the algorithm, there should still be continued evaluation of the integrity of the data. One possible method for doing this is by running the algorithm on simulated samples with variant backgrounds, and then recording and analyzing the results carefully to find the algorithm's weak points. This feedback can then go to the data collection team, reducing bias in future data.

In the literature, this problem of assessing an algorithm's fairness has been addressed before, and there are different methods that are available to us. We will define the same fairness criteria mentioned by Chouldechova (Chouldechova, A, 2017), within the context of criminal justice, specifically for the COMPAS example. It is important to be aware of these categories and how they differ in evaluating fairness, for this will allow us to better understand the impact of failure. Before defining these categories, it is also important to understand another concept: classifiers, such as the COMPAS model, give a score to each individual to determine their future risk. In this case, the score was on a scale from one to ten, where closer to ten means that the model predicts the person of having a greater risk for recidivism.

Calibration

A score is said to be well-calibrated if there is the same probability of recidivism given that two individuals have the same score based on the same covariates and belong to different groups. This is also referred as being free from predictive bias.

Predictive Parity

After predicting a score for each individual, a threshold value should be selected. Above that value, the model will classify a person as a high-risk. A score have predictive parity if the probability of recidivism is the same given that both individuals scores are above that threshold and they belong to two different groups. In the COMPAS example, the company argued that they have the same positive predictive value, which means across African-Americans and White-Americans groups have the same proportion of correctly classified high-risk individuals.

Statistical Parity

A score satisfies statistical parity if the probability of being above a threshold is the same for different groups. This is also known as equal acceptance rates. This category is not useful within our context, but is useful in others as a hiring process.

Error Rate Balance

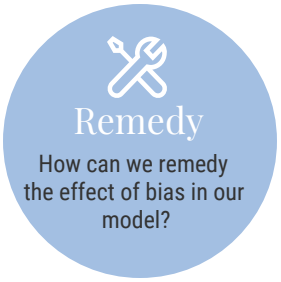
A score satisfies is error rate balance if the probability of them having a score higher than a threshold, this means they are categorized as high-risk, are the same given that they are not and they belong to different groups. This is known as a false positive. It also satisfies that it has the same probability if they are categorized as low-risk given that they are not and they belong to different groups, which is known as false negative rates. This is also known as equal opportunity.

In the example of COMPAS, the heated argument between ProPublica and Equivant results from using different definitions of fairness. The COMPAS model suffered from failing to comply with error rate balance, since it had a high imbalance between false positive and false negative rate between different races. This was the problem that ProPublica argued against; even if the model was able to predict high-risk recidivism between the two groups with the same accuracy, the proportion of those that were misclassified either as high-risk or low-risk was different among groups.

This problem was also examined by Chouldechova (Chouldechova, A, 2017), stating that predictive parity, equal false-positive error and equal false-negative error rates are statistically impossible to balance at the same time if there are differences across two groups. In this case, the difference is that members of the group are rearrested in different rates; for black defendants in the data, their recidivism rate was 51%, while for white defendants the rate was 39%.

It is important to be aware of these different categories, because an algorithm can be adjusted to comply with one category but might fail to satisfy another. Since choosing one category over another has serious consequences, a decision should be made about what the state considers fair in a specific context. This decision also needs to consider the risks for failing in each category. In this case, COMPAS gave priority to predictive parity, which means that their model is fair at predicting high-risk, but the proportion for misclassifications varies between groups. According to ProPublica, black individuals were more likely to be misclassified as high or medium risk than white people. This will affect their opportunities to be released from jail.

According to Rayid Ghani, director of the Center for Data Science and Public Policy, governments need to define the metrics that algorithms should be measuring and make sure that they are fair. One disadvantage with this solution is that the decision-makers in government might not be that familiar with these metrics. Thus, we believe it is important for those in charge of designing these models to help the users understand there is a tradeoff between these categories, and also communicate the potential model risks to the users.



Now that we've outlined methods that can be used to identify and quantify bias in our algorithms, we now have to face the problems in the 'Remedy' step: correcting the past effects of the algorithm, and improving the future implementation of the algorithm.

For the past effects, the biased results could affect a defendant in two ways: either deny their legitimate rights (ex. wrongful denial of parole) or give them rights which they might not deserve (ex. early release from prison). These situations needed to be handled differently and on a case-by-case basis, but we propose a pathway for correction.

For a defendant whose legitimate rights have been denied, we propose offering reparations to the defendant as soon as possible, and these reparations should attempt to fully offset the effect of the biased algorithm and its effect on the defendant's life. In the case we proposed above, an inmate in upstate New York was mistakenly denied parole in 2016. If this inmate suffered from financial loss due to the denial, the government should actively evaluate the financial loss fairly and pay back a certain amount with interest. There are also other considerations, such as the effect on the defendant's health and future employment status, as well as the impact on the defendant's family. In these cases, there are statutes and precedents set for compensating the wrongfully convicted 1, so states can use these as a baseline for their reparations.

IThe other case, where a defendant is wrongfully given rights which they might not deserve, is a more complicated situation. We do not claim to know all of the legal intricacies that are involved here, so we will not suggest any specific solutions, but it will fall upon the government to figure out what the best course of action will be. We hope that by following the 'Identify', 'Quantify', 'Remedy' process, the number of cases that fall in this category will lessen over time, but in the meantime there will need to be a specified procedure put into place.

Apart from correcting past mistakes, we also identified the need to modify future use of the algorithm. For this case, we propose continued evaluation and feedback on the performance of algorithm. If in a specific time period (ex. last one year), the error rate is over a certain threshold (ex. 0.1%), then the algorithm should be pulled and examined further. There could also be certain information output by the algorithm that aids in decision-making; for example, the algorithm could include confidence scores in its report, allowing the human decision-maker (judge) to take this information into account when using the algorithm in the future. For the feedback loop, we propose continuous review of the mistakes that the algorithm makes and integrating new changes to keep the algorithm updated. In our 'Identify' section, we proposed external audits to continually evaluate the bias present in these algorithms. This information could then be used to quantify the bias and ,ultimately, remedy the bias by modifying the algorithm. We strongly believe that this loop will go a long way in improving these algorithms and preventing discrimination in the future.

Conclusion

To identify bias, we propose external audits, publicly accessible/transparent models, and data checks throughout the data collection process. We mentioned that the main problems we found when attempting to identify bias in these algorithms were related to proprietary knowledge and biased data collection, so we created our proposed solutions in direct accordance with these problems. By creating algorithms in the future that are transparent and continually audited by external organizations, then we increase our chances of identifying bias in our algorithms before they make incorrect or biased predictions. In regards to data collection, implementing checks in various stages of data collection allows an organization to identify bias before the information is input into the algorithm, leading to better recommendations that are less likely to be imbued with bias.

To quantify or measure level of bias in our information, we propose establishing a fairness criterion, monitoring how well the specified algorithm does for this criterion, and clearly communicate the results of this to the public. We mentioned that there are trade offs between different definitions of fairness, and it should be up to the model user to decide on which criterion is best for their purposes, as long as they also communicate the risks associated with their choice. We also mentioned different metrics that can be used to quantify bias once this criterion is established, including calibration, predictive parity, statistical parity, and error rate balance. All of these are useful examples of concrete measurements that can be applied to any criminal justice algorithm, which is useful for evidence-based evaluation of implemented models.

To remedy the effect of biased information in our models, we propose reparations for wronged individuals and iterative improvements for algorithms that we have identified and quantified as being biased. We mentioned how criminal justice algorithms could output risk scores and recommendations that lead to wrongful pretrial detainment or sentencing, and this raises concerns about how to make up for these mistakes to the individuals. We recognize that this is a difficult problem that requires a case-by-case approach, but the use of any criminal justice algorithm with bias necessitates a form of reparation. In terms of modifying algorithms in order to prevent future biased predictions, we proposed iterative feedback and improvement cycles in order to continually ensure that our model is not discriminating against members of protected classes. By continually identifying, quantifying, and removing bias in these algorithms, we have a process for increasing the probability that the recommendations are equitable.

By adding qualitative information and contextual thinking to the decision-making process, we can improve criminal justice algorithms while mitigating the risk of unknowingly discriminating against certain groups in our population. Instead of solely following a model's recommendation, we encourage an inquisitive mindset on the behalf of the decision-maker, enabling them to ultimately make a decision that takes multiple perspectives into account. We also encourage this same mindset for the development of future criminal justice algorithms; we've outlined a workflow loop in this paper that can be used to ensure that these algorithms continue to be scrutinized and improved, which will ultimately move us closer to an equitable solution. The road ahead for this technology is murky and unclear, but with the right principles and guidelines in place, we can create a criminal justice system that is fair for all citizens.

References

Angwin et al., (2016) Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. Propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

Consumer Financial Protection Bureau (2014). Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

Courtlan, R. (2018). Bias detectives: the researchers striving to make algorithms fair. Nature. https://www.nature.com/articles/d41586-018-05469-3

Gershgorn, D. (2018). California just replaced cash bail with algorithms. Quartz. https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

Neufeld, A. (2017). In Defense of Risk-Assessment Tools: Algorithms can help the criminal justice system, but only alongside thoughtful humans. The Marshall Project. https://www.themarshallproject.org/2017/10/22/in-defense-of-risk-assessment-tools

Rudin, C. (2018). Algorithms and Justice: Scrapping the 'Black Box'. The Crime Report. https://thecrimereport.org/2018/01/26/algorithms-and-justice-scrapping-the-black-box/

Yong, E. (2018). A Popular Algorithm Is No Better at Predicting Crimes Than Random People. The Atlantic. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/