# COMPAS, algorithms and the fairness debate

Sae Koyama

Winter 2019-2020

**Abstract**

This essay discusses algorithmic fairness and the COMPAS algorithm for predicting recidivism among offenders. I consider the study conducted by Propublica claiming COMPAS is bias against Blacks and the subsequent rejoinder. I discuss why mathematical analysis can be contradictory and conclude that mathematical methods have limitations when it comes judging fairness, as fairness depends on context and can be sensibly described in contradicting ways.

# Contents

# 1 Introduction

Algorithms are increasingly being integrated into our everyday lives, having a direct impact on the decisions made around us. This raises important questions on the consequences, both intentional and unintentional, of using these algorithms. Here I focus on the issue of algorithmic fairness. Our society today is far from fair. While algorithms could be the future of reducing human bias, it is also vulnerable to picking up existing bias from data sets and cementing them in the form of an algorithm which is seemingly irrefutable to the majority of the population [12, 7].

One example of an algorithm used to make decisions is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The software was developed and is owned by Northpointe (now Equivant), for assessing the needs and risks of offenders in the criminal justice system. The Practitioner's Guide for COMPAS states that it can be used at "various decision points within the local criminal justice system and with various populations", including pre-trial screening and probation [10, p2]. COMPAS has three risk scales for which offenders are given a score from 1 to 10: general recidivism risk, violent recidivism risk[1] and risk of Failure-to-Appear (FTA). The scores 1 to 4 are classified 'low risk', 5 to 7 classified 'medium risk' and 8 to 10 classified 'high risk'. To make the classification, a form is completed, with 137 questions on the offender's criminal history, their background and their attitude to crime [2]. This is then compared with the responsive from a normative group[2]. The software then uses an algorithm to determine the offenders score on the three risk scales.

COMPAS came into the spotlight when Propublica wrote an article claiming that the COMPAS algorithm was bias against Blacks [2]. However this was disputed by both Northpointe and other researchers in algorithmic bias who claimed that Propublica's analysis was flawed and that COMPAS was not bias [7, 10]. Further articles tried to reconcile these two viewpoints by explaining how they stemmed from different definitions of fairness [2, 8]. In this essay, I summarise some of these arguments. Viewing fairness from both a mathematical perspective and a philosophical one, I conclude that while algorithms can improve equality in society, both of these disciplines must be understood by those using and developing the algorithm.

# 2 Intention and Use

Northpointe was a consulting and research firm, developing software and providing other services such as training to justice systems and policy makers in the US and Canada. In 2017, they merged with two other companies to form Equivant, which continues to provide tools for the justice industry, including COMPAS. Their website claims that "our tools are used nationwide for evidence-based decision making, helping to remove biases by equipping justice professionals

[1] COMPAS predicts risk of subsequent offences within two years of assessment.

[2] For the 2012 version, this was a sample of over 30,000 taken between January 2004 and November 2005 [10, p2].

with the research and rationale they need" [6]. Since it's development in 1998, COMPAS has been used to assess more than a million offenders [8].

Different elements of the COMPAS system is intended to be used at different stages. For example, the Failure to Appear (FTA) risk scale is intended for pretrial defendants to inform pretrial release decisions [10, 7], while the Field Guide suggests the use of the Violent Recidivism and Recidivism Scales at the probation stage. In their dispute of Propublica's analysis, Fores et al. claim that algorithms predicting recidivism are mostly developed for probationers and parolees [7]. However Propublica says that COMPAS is being used in pretrial and sentencing [1] and their original article focuses on how the recidivism risk scale was influencing prison time [2]. In fact, the COMPAS Field Guide does suggest assessing defendants with high FTA scores on all scales to get a 'holistic view'.

How exactly COMPAS is used is important because what is considered 'fair' is heavily context dependent and in their article *The Authority of "Fair" in Machine Learning* [13], Skirpan et al. note that questions such as "is it fair to make a machine learning (ML) system?", "is this ML system fair?" and "are the results fair?" are different questions. The usage of an algorithm can shift the focus to each question. Using risk scales to design effective support systems for offenders that are high risk because of say, criminal friends, could be considered a fair use of risk assessment and the focus is on whether the algorithm and its results are fair. Meanwhile, the very *use* of algorithms to give high risk offenders more prison time could be considered unfair.

On one hand, algorithms such as COMPAS could be a road to a more efficient and sustainable sentencing system. There is unconscious bias in human decision making; for example, defendants are more likely to get lighter sentences just after a lunch break than before [14]. While not removing them entirely, algorithms could be a way of at least reducing such biases. On the other hand, the use of algorithms in sentencing risks making the court process opaque and irrefutable. While human judgement can be examined further by other humans, the workings of an algorithm can be hidden as a trade secret [2]. Even with a the best intentions, decisions made by machine learning algorithms can be unintelligible to a human because of the complexity of the system or its use of concepts beyond our comprehension, such as the geometry of higher dimensional space [3].

The use of such algorithms raises a moral question beyond the problem that bias present in such algorithms may become impossible to remove or to challenge. A single judge influences decisions in one courthouse and will eventually retire, algorithms like COMPAS can be copied numerous times and used indefinitely.

While algorithms may be cheaper to implement than training specialists, they may not be more accurate. Dressel et al. [8] and Biswas et al. [5] did a study looking at how humans compare to COMPAS and found that the accuracy of human decision systems (majority voting) was comparable to that of COMPAS. The cost benefits of using algorithms like COMPAS has to be weighed with these consideration.

Another important question is how much influence COMPAS has. The Field Guide does note

that the COMPAS risk scales are unsuitable for some types of offences, such as sex offences or domestic abuse, and advises practitioners to look at mitigating circumstances. However many decisions can be made unconsciously so it would be interesting to study the impact of a high or low score on decision making, with different disclaimers. However this is beyond the scope of this essay.

# 3 Different analysis, different results

With this context in mind, I now discuss the analysis of Propublica and the subsequent rejoinder. Propublica is a non-profit organisation based in New York which aims raise public awareness in investigative journalism. Their investigations cover issues such as healthcare and hate crime [11].

Propublica studied the COMPAS results and subsequent recidivism rates of 10,000 pretrial criminal defendants from Broward County, Florida. Regarding a score of 1 to 4 as predicting negative and 5 to 10 as predicting positive, they found that accuracy in predictions was roughly the same for both races, but the false positives[3] and false negatives[4] rates differed. In particular, Black defendants were more likely to be mislabelled as high risk, while white defendants were more likely to be mislabelled as low risk [2]. In their article, *How we Analysed the COMPAS data* [1], Propublica also published further analysis on the data they had collected. They made a logistic regression model considering race, *future recidivism* and other factors such as age and gender etc., with *the probability of being given a medium or low score*. They found being Black increased the probability of a higher COMPAS score, adjusting for other factors. They also used a cox proportional hazard model with interaction term to see how race and score affected recidivism rates. They found that given an offender with a higher score, they were less likely to re-offend if they were Black.

In response to the article by ProPublica Flores et al. published a rejoinder [7] disputing the conclusions of ProPublica's study. Anthony Flores is a professor at the California State University, Christopher Lowenkamp is in the Administrative Office of the United States Courts, Probation and Pretrial Services Office and Kirstin Bechtel was at the Crime and Justice Institute at the Community Resources for Justice, a non-profit working in services and policy. They had five main points of contention. Firstly, they claim that the use of pre-trial defendants was inappropriate because actuarial risk assessment instruments (ARAI) are 'typically developed and administered to probationers and parolees'. Secondly, they dispute the decision to make both 'medium' and 'high' risk as positive predictions, as putting the 'low' and 'medium' categories together gives different results. However Propublica justified both decisions by the use

[3]If we let the percentage that were negative and classified as negative be TN (true negative), and the percentage that were negative and classified as positive be FP (false positive), the true positive rate is $FPR = FP/(FP + TN)$. The probability of mislabelled as high risk, given they do not re-offend.

[4]If we let the percentage that were positive and classified as positive be TP (true positive), and the percentage that were positive and classified as negative be FN (false negative), the true positive rate is $FPR = FN/(FN + TP)$. The probability of mislabelled as low risk, given they do re-offend.

of COMPAS in pre-trial screening and the Field Guide's assertion that 'medium' and 'high' risk scores garner attention. Thirdly, they state that differences in mean score between races is not bias. Fourthly, they claim 'well established and accepted standards exist to test for bias in risk assessment' and that the Propublica analysis did not use them. Finally, the p-values that the Propublica analysis found were not significant, given the sample size.

Flores et al. do their own analysis on the COMPAS data set. They analyse the distribution of recidivism for each decile score by looking at Area Under Curve - Receiver Operating Characteristics (AUC-ROC) scores for different races and conclude that they are similar. AUC-ROC scores depend on true positive rates (TPR) [5] and false negative rates (FNR), given a category is cut off at different scores. This suggests COMPAS is similarly good at separating high risk and low risk defendants regardless of race. They also did their own series of logistic regression models with *the probability of future recidivism* considering race and *COMPAS score* and find that *given the same risk score*, both Black and White offenders are equally likely to re-offend. They conclude that they could not find evidence of bias in the COMPAS data set.

To summarise Propublica's first two analysis suggests that given an offender with a set of characteristics and whether they re-offend, the distribution of COMPAS scores is different for different races. Meanwhile, Flores et al. find that given an offender with a set of characteristics and their COMPAS score, the distribution of re-offending is similar for different races[6].

Matthias Spielkamps article *Inspecting Algorithms for Bias* points out that the two sides of the argument are simply using different definitions of fairness [14]. Simplifying the argument, the positive predictive rate (PPV)[7] for the different races are similar as pointed out by North-Pointe, while the false positives and false negative rates differ as pointed out by ProPublica. The discrepancy in results arises primarily because of unequal recidivism rates across races. Recidivism rate is higher for Black offenders and they are incarcerated in numbers disproportionate to population [7, 15]. This is also found in the COMPAS dataset. The different recidivism rates mean that fairness is not as simple as statistical parity and furthermore, we cannot have both the PPV and false positive/false negative be equal[8].

Both PPV and false positive/false negative rates seem to be mathematical measures of bias but we have already seen that they fail to fully take in to account the different rates of recidivism across different races and seem 'unfair' in different ways. Other fairness metrics have been proposed; for example, Dwork et al. give an example of a mathematically model for fairness in their article *Fairness Mathematically* [9]. The key idea of the model they present is to 'treat similar individuals similarly'; that is, individuals with similar properties relevant to a particular task should have similar distributions of outcomes. However they note the difficulties that arise when two groups have different distributions of properties. For example, if high scores are

[5] If we let the percentage that were positive and classified as positive be TP (true positive), and the percentage that were positive and classified as negative (false negative), the true positive rate is $TPR = TP/(TP + FN)$ The probability of a high score given recidivism.

[6] Propublica concludes differently from their cox model, although the result is *almost* (and thus not quite) statistically significant.

[7] $PPV = TP/(TP + FP)$. The probability of re-offending given a high score.

[8] Elaborated in Appendix A.

given good outcomes and group A has much higher scores on average then group B, then it is possible to almost entirely exclude group B from good outcomes while still being 'fair' by this definition. There is also the matter of choosing 'relevant properties' in the first place, such as testing engineering skills when group A puts emphasis on engineering while group B puts emphasis on mathematics in their education, when both skills are 'relevant'. While this metric may be a useful framework, it should be used with the context in mind, such as selecting what properties are relevant and whether affirmative action is beneficial.

Furthermore, the conclusion that Black offenders have higher risk of recidivism so the greater likelihood of a high risk score is acceptable could also hide hidden reasons for the lack of statistical parity (re-offending rates are not the same as rearresting rates). It is important to bear in mind that algorithms can create feedback loops and testing is impossible since the algorithm influences the outcome. This also makes the algorithm impossible to falsify, leading to a wide variety of problems.

# 4 Fairness, Philosophically

As different assumptions about fairness give different results about the fairness of COMPAS and this suggests the contention is not mathematical but philosophical. In *Fairness in Machine Learning: Lessons from Political Philosophy*, Reuben Binns explains how mathematicians are trying to characters 'fairness' mathematically when we are far from a consensus on what even the most basic definitions of fairness and discrimination are [4].

Binns focuses on the notion of 'discrimination'. Some philosophers argue that an action is discriminating if there is 'systematic animosity or preference' towards certain groups. By this definition, algorithms cannot be discriminating because they do not exhibit conciousness. Another view point is that discrimination occurs when individuals are not treated as individuals but as part of a general group, which would make algorithms discriminating by its very nature. A further viewpoint is that discrimination leads to some sort of inequality but what is meant to be 'equal' depends heavily on context. For example, it could be argued that a job applicant should have equal possibilities but not equal outcomes, while the right to vote should be uniformly distributed. We cannot judge an algorithm to be mathematically fair or not without addressing the underlying philosophical issues.

Does this mean we should drop all attempt to create a mathematical 'fairness metric'? If we have algorithms that affect lives, then it is essential that we monitor its results to prevent continued discrimination against minorities going undetected. With the vast amounts of data which is processed by these algorithms, this seems difficult to do without some computing help. On the other hand, we have seen how different definitions of fairness are contradictory and no single metric can cover all the 'obvious' points of fairness. What we gain out of these metrics are the different measures that contribute to fairness, not a measure of 'fairness' itself. Instead of relying on a single metric to judge the fairness of an algorithm as a binary result, it is perhaps more constructive to use several different metrics to see in what way an algorithm is fair, in

what way it is not, whether the two can be reconciled and what underlying elements mean they cannot.

# 5 Conclusion

In conclusion, whether COMPAS is fair or not depends on how it is used and your definition of fairness. On one hand, it has differing false positive, false negative values for Black and white offenders which suggests that a Black offender is more likely to be labelled higher risk. On the other hand, it has similar ROC-AUC values and PPR for different races which suggests that offenders labelled higher risk have similar probabilities of re-offending, regardless of race. Both definitions of fairness might be worth considering but it is impossible for COMPAS to be fair in both senses because of the different recidivism rates between races. While Equivant has a case that COMPAS is fair, or at least better than human judgement, it could be argued that they should be more active in informing users of the problems associated with using algorithms to make predictions about recidivism. While Propublica has a case that COMPAS is unfair, it could be argued that they should have included the affect of higher recidivism rates among Black people in their main article and that they overstated their conclusions. Both sides fail to convey in their analysis that their discussion is a *philosophical one* rather than a dispute on what mathematical method is correct.

The philosophical and moral debate about the use of algorithms is unfinished at every stage, from whether algorithms should be used in the first place if they do not treat people as individuals and are not understandable by humans, to whether statistical parity is a measure of fairness for different groups. Fairness in general has many different definitions and these cannot be reconciled within a single, well-defined metric. At most, these measures of 'fairness' can only be snapshots as to how an algorithm is fair *in what way*. It is up to humans to interpret and act on these results. In the case of recidivism, the mathematical discrepancies highlight the underlying problem of recidivism rates among Black people, with futher underlying causes such as poverty and lack of education which should be tackled [15]. While Equivant claims that algorithms are the future of a less bias society, these problems clearly show that algorithms are not the entire solution.

There are several important points of discussion which this essay does not cover. What responsibilities does Equivant have to insure the limitations of their software are well understood and that it is being used correctly? It would be better if their algorithm was more transparent and constructed in a way which allowed individual offenders to dispute the results (by presenting mitigation circumstances for factors which lead to high scores), however this is currently hidden as a trade secret. How could laws be changed to make this change possible? How can we tackle the problem of falsifiability and feedback loops? How could this be balanced with the accuracy of the system, when it is not that accurate anyway? How could the underlying problem of Black recidivism be tackled? Would positive affirmative action implemented within systems like COMPAS be harmful or helpful? While algorithms have the *potential* of doing good, they are not magic boxes that will fix all of our problems. This must be understood by the people

using them.

# References

[1] Angwin J, Larson J, Mattu S, Kirchner L. *How We Analysed the COMPAS Recidivism Algorithm* ProPublica. [2016 May 23]. Available from: `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`

[2] Angwin J, Larson J, Mattu S, Kirchner L. *Machine Bias. There's a software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. [2016 May 23] Available from `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

[3] Bathaee Y. The Algorithmic Intelligence Black Box and the Failure of Intent and Causation *Harvard Journal of Law & Technology* 2018;31(2):890-938.

[4] Binns R. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81:1-11, 2018

[5] Biswas A, Kolczynska M, Rantanen S, Rozenshtein P. *Comparing human recidivism risk assessment with the COMPAS algorithm.* Available from: `https://www.dropbox.com/s/xc2ukfqbwog7718/EuroCSS.pdf`

[6] Equivant homepage. Available from: `https://www.equivant.com/`

[7] Flores A W, Lowenkamp C T, Bechtel K. *False Positives, False Negatives, and False Analyses: A rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks"* Available from: `http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf`

[8] Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism *Science Advances* 2018; 4(1):eaao5580.

[9] Dwork C, Hardty M, Pitassiz T, Reingoldx P. *Fairness Through Awareness* November 30, 2011

[10] *Practitioner's Guide to COMPAS* Northpointe. 2012 Aug. 17.

[11] Wikipedia. *Propublica* Available from: `https://en.wikipedia.org/wiki/ProPublica`

[12] Kirkpatrick K. It's not the Algorithm, its the Data. *Communications of the ACM.* 2017; 60(2):21-23.

[13] Skirpan M, Gorelick M. The Authority of "Fair" in Machine Learning. 2017.

[14] Spielkamp M. *Inspecting Algorithms for Bias* MIT Technology Review. [2017 Jun 12] Available from: `https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/`

[15] Wehrman M W. Race, concentrated disadvantage, and recidivism: A test of interaction effects. *Journal of Criminal Justice* 2010;38(4):538-544.

# Appendix A PPV vs FP/FN

Consider an algorithm which classifies offenders as high risk (H) or low risk (L) of recidivism. There are two groups of people, A and B, with the rate of recidivism $p_1$ and $p_2$ respectively. Say $x \in R$ if $x$ is a recidivist. For an element $x$, suppose

$$P(x \in R | x \in H) = q_1$$

$$P(x \notin R | x \in L) = q_2$$

Note, $q_1$ and $q_2$ are the positive predictive value (PPV) and negative predictive value (NPV) respectively. The fact these are independent of group could be a sign that the algorithm is 'fair' as the probability that $x$ is a recidivist given it's classification is independent of the group it is in.

Let $r_1$, $r_2$ be the probability that $x \in H$ given $x \in A, B$ respectively. $r_1$ can be determined from $p_1, q_1, q_2$ as follows.

Given $x \in A$,

$$\begin{aligned} p_1 = P(x \in R) &= P(x \in H \cap x \in R) + P(x \in L \cap x \in R) \\ &= P(x \in H)P(x \in R | x \in H) + P(x \in L)P(x \in R | x \in L) \\ &= r_1 q_1 + (1 - r_1) q_2 \end{aligned}$$

Rearranging,

$$r_1 = \frac{p_1 - q_2}{q_1 - q_2}$$

and similarly,

$$r_2 = \frac{p_2 - q_2}{q_1 - q_2}$$

Now we can make a truth tables (Table 1 and Table 2).

| A | High | Low |
|---|---|---|
| $x \in R$ | $\frac{p_1-q_2}{q_1-q_2} \times q_1$ | $(1 - \frac{p_1-q_2}{q_1-q_2})q_2$ |
| $x \notin R$ | $\frac{p_1-q_2}{q_1-q_2} \times (1-q_1)$ | $(1 - \frac{p_1-q_2}{q_1-q_2})(1-q_2)$ |

Table 1. Truth table for group A

| B | High | Low |
|---|---|---|
| $x \in R$ | $\frac{p_2-q_2}{q_1-q_2} \times q_1$ | $(1 - \frac{p_2-q_2}{q_1-q_2})q_2$ |
| $x \notin R$ | $\frac{p_2-q_2}{q_1-q_2} \times (1-q_1)$ | $(1 - \frac{p_2-q_2}{q_1-q_2})(1-q_2)$ |

Table 2. Truth table for group B

Note that for $p_1 > p_2$, the rate of false positives for group A is higher than that for group B, while the rate of false negatives is lower and so while this algorithm is 'fair' in one sense (PPV and NPV same for both groups), it is not 'fair' because it is more likely to misclassify people in group A as high risk.