Responsible AI, Law, Ethics & Society

Fairness & AI Solutions To Bias In The Criminal Legal System:

Roni Berman

LL.B Law and Economics, Tel Aviv University

Adi Levi

B.S Information System Engineering, Technion

Abstract: Prediction algorithms are increasingly being used in criminal justice. In particular COMPAS software is used in courts across the US. Even though the use of AI is undoubtedly effective and beneficial, it raises many technical, legal, and ethical questions. A reverse model published by ProPublica was made to show that the tool is biased against African Americans in the US.

In this Report, we are going to question how a biased outcome affects human and constitutional rights, especially the defendant's rights. Then, we will discuss how affirmative action might solve the bias through 3 possible methods.

Our main question is, what are the possible solutions to assure fairness (applied in the case study of the ADM made by COMPAS)?

1. Artificial Intelligence in the Judicial System:

1.1. Digitization of court systems:

The process of digitizing the courts began at the end of the third industrial revolution.
Now, it provides an excellent opportunity to transform the judicial system into a quick, efficient, and high-quality array of services.

The digitization of court systems creates a foundational basis upon which available data can be used to identify the spaces in the judicial system where AI application can have significant impact.

1.2. Benefits:

Introducing AI to the justice system promises to improve procedural and administrative efficiency, aid in decision making processes for judges, lawyers, and litigants. Furthermore, it is likely to improve the transparency of the functioning of justice by improving the predictability of the application of the law and the consistency of case law.
Proponents believe AI could offer the key to reducing human error and bias in the courts. If algorithms could accurately predict recidivism, the increased fairness would permit courts to be more selective about who is imprisoned and for how long.[1]
There have been some positive results reported from the use of risk assessment tools. A risk assessment algorithm adopted as part of the 2017 New Jersey Criminal Justice Reform Act resulted in a 20 percent reduction in the number of people incarcerated while awaiting trial.[2]

1.3. Judicial Use of AI for Risk Assessment:

In recent years, courts around the country have been using AI-driven assessment tools to gauge the risk of recidivism of defendants in criminal cases. In this report we will talk about One popular application, called COMPAS.

1.4. COMPAS:

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, developed and owned by Northpointe was one of the most popular scores used nationwide and were increasingly used in pretrial and sentencing.
When most defendants are booked in jail, they respond to a COMPAS questionnaire. Their answers are fed into the COMPAS software to generate several scores including predictions of "Risk of Recidivism" and "Risk of Violent Recidivism."
Depending on the scores generated by this software, the judge can decide upon whether to detain the defendant prior to trial and/or when sentencing.

1.5. Data description:

ProPublica obtained pretrial defendant's COMPAS scores from the Broward County Sheriff's Office in Florida in 2013 – 2014.
Broward County is a large jurisdiction using the COMPAS tool in pretrial release decisions and Florida has strong open-records laws.
Each pretrial defendant received at least three COMPAS scores, each ranged from 1 to 10, with ten being the highest risk:

1. Risk of Recidivism- 'decile_score'
2. Risk of Violence- 'v_decile_score'
3. Risk of Failure to Appear

There was given a description of the scores: "Low" (1 – 4), "Medium" (5 – 7) and "High" (8 – 10)
It's worth noticing that the Defendants who are classified medium or high risk (scores of 5–10), are more likely to be held in prison while awaiting trial than those classified as low risk (scores of 1–4).

Additional attributes examined:

4. score_text- Risk of recidivism category
5. v_score_text- Risk of violence category
6. days_b_screening_arrest- number of days before COMPAS assessment being conducted
7. c_charge_degree- the degree of the charge

[1] Angwin et al., supra note 6.
[2] Heaven, supra note 3.
[3] Brennan, Dieterich et Ehret, Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System

8. priors_count- number of prior offences
9. is_recid- yes/no prediction of the model of whether the defendant will reoffend
10. two_year_recid- actual result over a two-year period
11. is_violent_recid- yes/no prediction of the model of whether the defendant will have a violent offence
12. juv_misd_count- number of juvenile misdemeanor crimes
13. juv_fel_count- number of juvenile felony crimes
14. juv_other_count- number of juvenile crimes with degree different than misdemeanor or felony

1.6. How We Defined Recidivism:

Northpointe defined recidivism as *"a finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code."*[3] meaning a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored.
Northpointe's practitioners guide says that its recidivism score is meant to predict *"a new misdemeanor or felony offense within two years of the COMPAS administration date."*

We looked at criminal defendants in Broward County, Florida, and **compared their predicted recidivism rates with the rate that actually occurred over a two-year period.**
We set out to assess the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

1.7. Data Preprocessing:

We filtered the underlying data from Broward County to include only those rows representing people who had either recidivated in two years or had at least two years outside of a correctional facility.
To match COMPAS scores with accompanying cases, we considered cases with arrest dates or charge dates within 30 days of a COMPAS assessment being conducted

We did not count traffic tickets and some municipal ordinance violations as recidivism because there is no jail time.
To determine if a person had been charged with a new crime subsequent to a crime for which they were COMPAS screened, we did not count people who were arrested for failing to appear at their court hearings, or people who were later charged with a crime that occurred prior to their COMPAS screening.

We can see that the data consist of many different races.

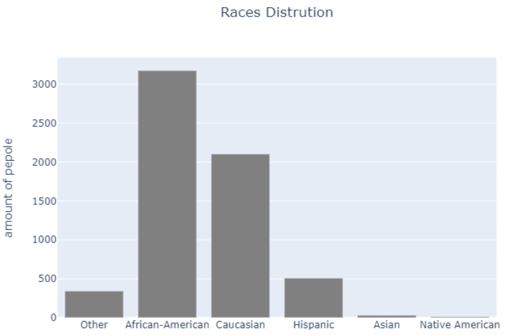


Figure 1: Race distribution in the dataset

We can see that Caucasian and African American represent 85.5% of the data, thus we will focus on those 2 groups because we want more accurate results

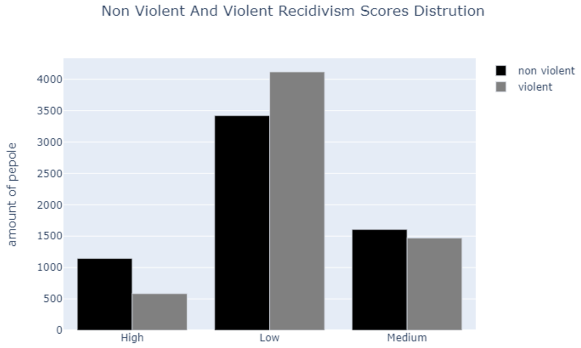


Figure 2:Non Violent And Violent Recidivism Scores Distrution

We can see that most of the results are from the 'LOW' category, furthermore according to Northpointe's practitioners guide: "scores in the medium and high range garner more interest from supervision agencies than low scores, as a low score would suggest there is little risk of general recidivism,"
So, we will consider scores that are higher than 'LOW' to indicate a risk of recidivism.

2. Bias In The Data:

From the analysis we made we found:

2.1. Risk of Recidivism:

We can see that 26.6% of African American received a “HIGH” score whereas only 11% of Caucasian individuals received a similar score, meaning that the rate of receiving a **“HIGH” score for African Americans is more 2.5 times that of Caucasians.**

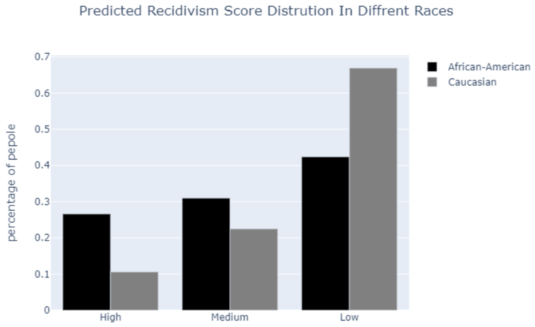


Figure 3:Predicted Recidivism Score Distrution In Diffrent Races

In addition, we can see that 67% of Caucasians received a ״LOW״ score whereas only 42.3% of African American individuals received a similar score, meaning that the rate of receiving a **״LOW״ score for Caucasians is more 1.6 times that of African Americans.**

In order to test the differences in the score distribution for the different races, we created a logistic regression model that considered race, criminal history, future recidivism, charge degree, gender and age.

We have found that, African Americans defendants were **45.31% more** likely to get a higher score than Caucasians defendants.

2.2. Risk of Violent Recidivism:

We can see that 12.02% of African American received a “HIGH” violence score whereas only 3.73% of Caucasian individuals received a similar score, meaning that the rate of receiving **a “HIGH” violence score for African Americans is about 3 times that of Caucasians.**

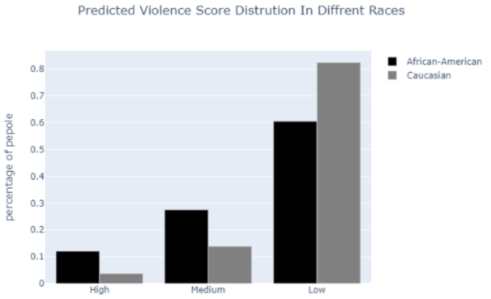


Figure 4:Predicted Violence Score Distrution In Diffrent Races

In addition, we can see that 60.46% of Caucasians received a ״LOW״ violence score whereas only 82.4% of African American individuals received a similar score, meaning that the rate of receiving a **״LOW״ violence score for Caucasians is more 1.36 times that of African Americans.**

To test the differences in the violence score distribution for the different races, we created a logistic regression model that considered race, criminal history, future recidivism, charge degree, gender, and age.

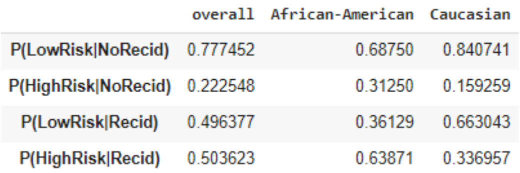
We have found that, African Americans defendants were **96.39% more** likely than Caucasians defendants to receive a higher score, correcting for criminal history and future violent recidivism.

We can clearly see from those results that the data collected was biased towards African American, but what about the COMPAS algorithm?

3. Bias In The Algorithm:

We were interested in how the COMPAS scores do at predicting recidivism and how their ability to predict depends on race.

We have trained a logistic Regression classifier on only the decile_score as a feature and then examine how the confusion matrices differ by race.

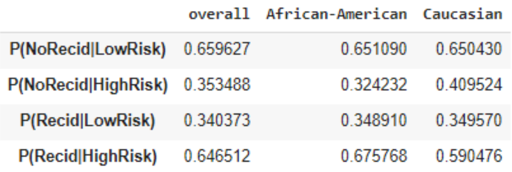


	overall	African-American	Caucasian
P(LowRisk|NoRecid)	0.777452	0.68750	0.840741
P(HighRisk|NoRecid)	0.222548	0.31250	0.159259
P(LowRisk|Recid)	0.496377	0.36129	0.663043
P(HighRisk|Recid)	0.503623	0.63871	0.336957

FNR - $\boldsymbol{P(LowRisk|Recid)}$ are higher for Caucasian than African American (0.663043 vs 0.36129), meaning we are more likely to misclassify Caucasian defendants who recidivated as a ״LOW״ risk than their African American counterparts.

FPR - $\boldsymbol{P(HighRisk|NoRecid)}$ are higher for African American than Caucasian (0.31250 vs 0.159259), meaning we are more likely to misclassify African American defendants who did not recidivated as a ״HIGH״ risk than their Caucasian counterparts.

It's worth noting that Northpointe, the company that produces COMPAS, argued that COMPAS is not biased because the probabilities of outcomes conditional on predictions (like $P(NoRecid|LowRisk)$) are approximately equal across races.



	overall	African-American	Caucasian
P(NoRecid|LowRisk)	0.659627	0.651090	0.650430
P(NoRecid|HighRisk)	0.353488	0.324232	0.409524
P(Recid|LowRisk)	0.340373	0.348910	0.349570
P(Recid|HighRisk)	0.646512	0.675768	0.590476

As we can see, the distribution of outcomes conditional on predictions does not vary too much with race. Moreover, if anything, it discriminates in favor of African Americans.

4. The Legal Perspective Of COMPAS Algorithm, Criminal Law & Affirmative Action:

4.1. Facts & Current Law:

Federal law in the US,[4] and the GDPR in the European Union[5] do not allow the use of biased algorithms to make automated decisions. For instance, the French government has prohibited the use of ML (machine learning) to evaluate judges and court clerks, with violators facing up to five years in prison. On the other hand, the Israeli Criminal Code does not address the use of algorithms regarding bias or sentencing.

The main questions we will ask in the light of democracy are:

- Is it adequate to use AI in criminal justice? How can guilt proof be assured when decisions are based on a machine?
- Is it appropriate to use AI for all offenses or only for fines and minor offenses?
- What is the right definition of fairness? And what is the appropriate tradeoff to make when balancing the fairness and accuracy of a ML tool?
- Should we use an algorithm to alleviate an overcrowded legal system despite the potential harm of due process?
- What should be the role of government regulations?
- Do people have the right to audit the data used to predict decisions about their lives that can affect their liberties if they think it is biased?
- The scope of the judge's role will need to be examined if we use COMPAS regularly in criminal systems - is a human in the loop needed?

The algorithm was found to be biased against African Americans in criminal proceedings. The baseline data used reflects discriminatory practices that contributed to injustices within historically disadvantaged minority groups.

It was found that African Americans defendants were **45.31%** more likely than Caucasian defendants to receive a higher score.
Is the bias justified? **According to the FNR and FPR, the answer is "NO."**

The FNR is higher for Caucasians than for African Americans, meaning we are more likely to misclassify Caucasian defendants who recidivated as a "LOW" risk than their African American counterparts.

The FPR is higher for African Americans than for Caucasians, meaning we are more likely to misclassify African American defendants who did not recidivate as a "HIGH" risk than their Caucasian counterparts.

4.2. Fairness:

Allows the equal treatment of individuals or groups based on certain characteristics (age, gender, ethnicity, sexual orientation, and religious belief). However, it is important to treat differences within the database (i.e., to allow affirmative action). It is to be noted that fairness is contextual.

4.3. Affirmative Action:

Affirmative Action, also called positive discrimination, is defined as measures taken with the purpose of correcting systematic biases against minorities. In principle the practice consists in

[4] Aiming for truth, fairness, and equity in your company's use of A, The Federal Trade Commission (April 19, 2021).

[5] How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers. Human Rights Watch.

granting preferential treatment to members from discriminated and statistically under-represented groups. With regards to COMPAS, affirmative action should be used to mitigate negative effects that are the result of racial bias in the dataset.

4.4. The Algorithm Applications:

In the criminal proceeding, there are three stages: pre-trial proceedings, trial proceedings, and sentencing. When it comes to COMPAS, we think it is only useful for the sentencing phase, since that is the only time when the court's decisions are affected by how often a person breaks the law again.

4.5. Trail Proceedings Use Of The Algorithm:

The Israeli law and case law forbid judges from using a defendant’s criminal record during the trial proceedings (clause 163 of the Israel Criminal Procedure Law). As such, there is no use for a recidivism prediction algorithm during said proceedings.

According to the European Commission’s Proposal for a Regulation on AI (2021), AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned—recidivism prediction algorithms shall probably be considered as such.

4.6. Sentencing Use Of The Algorithm:

In accordance with the Israeli Criminal Code, clause 37, "probation officer survey”, includes the defendant's criminal record that will be submitted to a judge. Then according to clause 40A-40O, the court’s sentencing is divided into two parts:

Firstly, deciding on an appropriate sentencing guideline, based on similar case law; and secondly, deciding what is the appropriate sentence for the individual. In the EU, the process is not unified but very similar.

The aforementioned algorithm, in its current form, has relevance in the second part of sentencing. The penal code allows the use of recidivism as a factor in sentencing, specifically as a factor for sentencing that exceeds the sentencing guidelines. The methods do not consider mitigating circumstances when sentencing, such as marital or mental status, and therefore, when sentencing, the judge or committee will have to make a decision that gives sufficient weight even to these external parameters in order to give concrete justice to each.

Algorithms might violate human rights. Moreover, what is **the impact of AI on human emotional experiences**, including the ways in which AIs address (or fail to do so) cultural sensitivities and emotional harm? Immanuel Kant said that “*rational human beings should be treated as an end in themselves and not as a means to something else*”.

Therefore, human rights shall not be violated using private software that maximizes profits while ignoring the value of liberty (and, negative liberty).

Furthermore, what will be the impact on the next generations if bias is transmitted forward instead of managed?

5. AI Solutions To Bias:

We saw that there is a clear bias in the datasets - African American defendants were more likely to be misclassified as higher risk compared to their Caucasian counterparts. To achieve fairness, we will try to achieve an important aggregated fairness measures called Equalized Odds parity.

5.1. Equalized Odds Parity:

Equalized Odds parity ensures parity between the subgroups of each race with label 1 in the training set, and parity between the subgroups of each race with label 0 in the training set.

This means that the subgroups of each race who reoffended are equally likely to be predicted to reoffend. Similarly, there is parity between subgroups of each race without recidivism.

In mathematical terms:

$$TPR_{African\ American} = TPR_{Caucasian}$$

$$P(predicted\ recidivism\ |African\ American, recidivism)$$
$$= P(predicted\ recidivism\ |\ Caucasian, recidivism)$$

&

$$FPR_{African\ American} = FPR_{Caucasian}$$

$$P(predicted\ recidivism\ |African\ American, no\ recidivism)$$
$$= P(predicted\ recidivism\ |Caucasian, no\ recidivism)$$

We can note that $TPR = 1 - FNR$,thus minimizing the diffrence between the TPR in the two group will also minimize the diffrence between the FNR in the two group

The observed probabilities in the data before applying any solution were:

$$TPR_{African\ American} = 0.7001$$

$$TPR_{Caucasian} = 0.3934$$

The diffrence: 0.3075

$$FPR_{African\ American} = 0.3446$$

$$FPR_{Caucasian} = 0.1662$$

The diffrence: 0.1784

5.2. Logistic Regression:

Logistic regression models the probabilities for classification problems with two possible outcomes. For example: given the parameters, will the defendant reoffend or not?

It's an extension of the linear regression model for classification problems. The logistic regression model uses the logistic function - sigmoid:

$$\sigma(w^T x_i) = \frac{e^{w^T x_i}}{1 + e^{w^T x_i}}$$

If we feed an output value to the sigmoid function, it will return the probability of the outcome between 0 and 1. This probability is the model's confidence score to the label he predicted

If $\sigma(w^T x_i) \geq 0.5$ then the label will be 1 (will reoffend) otherwise it will be 0 (won't reoffend).

Let's observe the distribution of the predictions of African American & Caucasian defendants with Logistic Regression model:

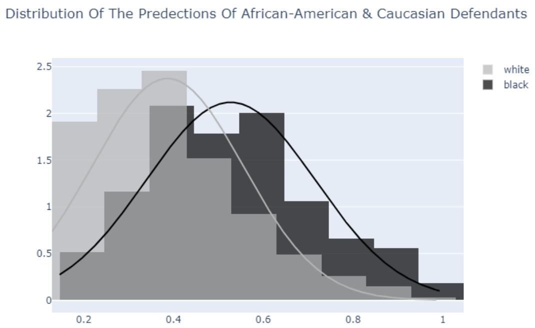


Figure 5:Distribution Of The Predections Of African-American & Caucasian Defendants with Logistic Regression model

We can clearly see that there is a major difference between the two races.

The probabilities of African American defendants are much higher of Caucasian defendants since the distribution of African American is to the right of the Caucasian defendants' distribution.

Also, the distribution of the Caucasian defendants is not as symmetric as of African American defendants, with more samples in the left side of the distribution, showing a clear high FNR for Caucasian defendants.

5.3. Method 1 - Threshold-Moving:

Many machine learning algorithms are capable of predicting a probability or scoring of class membership.

Those probabilities can be mapped to class label by using decision threshold where all values equal or greater than the threshold are mapped to one class and all other values are mapped to another class

The default value for the threshold is 0.5.

For those classification problems that have a severe class imbalance, the default threshold can result in poor performance.

As such, a simple solution for reducing the disparities is using group-specific decision threshold that will minimize the differences of FNR and FPR between the two races, thus enabling to achieve Equalized Odds parity.

After applying the method, we can see that in the train dataset we achieved:

Best difference achieved: 0.0224

African American best threshold achieved: 0.67

Caucasian best threshold achieved: 0.53

$$TPR_{African\ American} = 0.3168$$

$$TPR_{Caucasian} = 0.3103$$

The diffrence: 0. 0064

$$FPR_{African\ American} = 0.0909$$

$$FPR_{Caucasian} = 0.1068$$

The diffrence: 0. 0159

We can see that, the difference between the probabilities in the train dataset are really small.

To validate the results, in the train dataset we achieved:

$$TPR_{African\ American} = 0.3226$$

$$TPR_{Caucasian} = 0.2826$$

The diffrence: 0. 04

$$FPR_{African\ American} = 0.1184$$

$$FPR_{Caucasian} = 0.1037$$

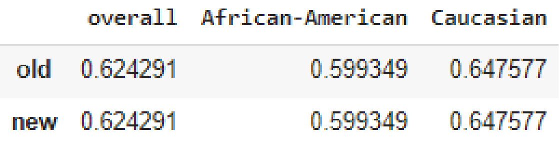
The diffrence: 0. 0147

We can see that, we indeed got small differences between the probabilities also in the test dataset (0.03 ,0.014) comparing to the original model (0.3, 0.18)

Since the difference in both the train dataset and the test dataset are really small, we can say that we have reached Equalized Odd parity.

Advantages:

1. Since we do not change the distribution of the risk of the defendants to re-offend in the model, we will not lower the accuracy of the model and/or the accuracy of the subgroups.



	overall	African-American	Caucasian
old	0.624291	0.599349	0.647577
new	0.624291	0.599349	0.647577

2. The main advantage of the method is that it reaches the local maxima, meaning we can be sure that we cannot get a better result using this method from what we have right now, something that cannot always be guaranteed using the “black box” method.

3. The method is interpretable, we can easily understand how the prediction was made and what considerations were taken into account. It can be changed to adjust other considerations by simply changing the threshold.

Disadvantages:

1. The main problem with the method is that we treat all African American defendants as one group. We can have a African American defendant who did recidivate in the past two years and the method will prioritize him over a Caucasian defendant even though he does not deserve it, just based on his race (African American).

2. Another problem of the method is that it could over-prioritize one group over the other since it is based on a small sample and not the actual distribution in the population. The algorithm uses a threshold of 67% for African Americans, and a threshold of 53% for Caucasians.

3. Although the algorithm is interpretable, a typical judge would not be able to understand why the model predicted the way he did unless he had an expert that would describe to him what the method takes under consideration.

5.4. Method 2 - Normalize The Prediction:

If we want to make the distribution of the predictions more similar, we can normalize the prediction of African American defendants by the prior of the distribution of Caucasian defendants and leave the prediction of Caucasian defendants as they are.

This means we are only considering the probability for recidivism given a race as an estimator for our bias, thus we can normalize the prediction "Post-Training" of the model

African American-Prior:

$$P(Recid|African\ American) = \frac{P(African\ American, Recid)}{P(African\ American)} = 0.53$$

Caucasian-Prior:

$$P(Recid|Caucasian) = \frac{P(Caucasian, Recid)}{P(Caucasian)} = 0.39$$

After applying the method, we can see that in the train dataset we achieved:

$$TPR_{African\ American} = 0.2963$$

$$TPR_{Caucasian} = 0.3715$$

The diffrence: 0. 0752

$$FPR_{African\ American} = 0.08$$

$$FPR_{Caucasian} = 0.1515$$

The diffrence: 0.0715

We can see that, the difference between the probabilities in the train dataset are really small.

Let's observe the distribution of the predictions of African American defendants and Caucasian defendants with the method on the train dataset:

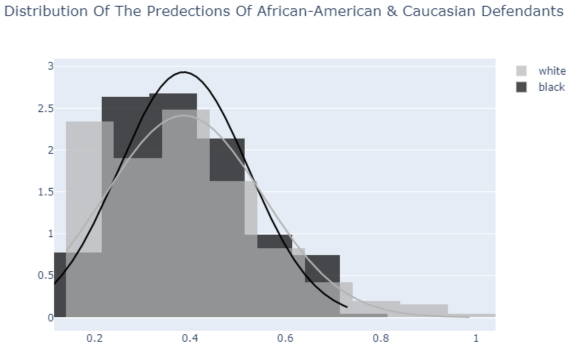


Figure 6: Distribution Of The Predections Of African-American & Caucasian Defendants with the method on the train dataset

We can see from the graph that the distribution of the African American and Caucasian defendants is very closely aligned together, which means that the difference between the probabilities in the train dataset are really small.

To validate the results, in the train dataset we achieved:

$$TPR_{African\ American} = 0.2683$$

$$TPR_{Caucasian} = 0.3214$$

The diffrence: 0. 0531

$$FPR_{African\ American} = 0.0794$$

$$FPR_{Caucasian} = 0.151$$

The diffrence: 0. 0717

Let's observe the distribution of the predictions of African American defendants and Caucasian defendants with the method on the test dataset:

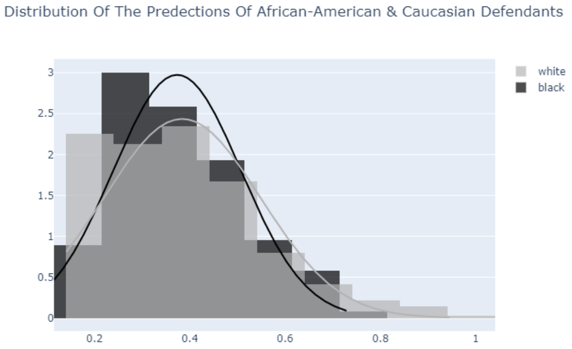


Figure 7:Distribution Of The Predections Of African-American & Caucasian Defendants with the method on the test dataset

We can see from the graph that the distribution of the African American and Caucasian defendants is very closely aligned together, which means that the difference between the probabilities in the test dataset are really small as well.

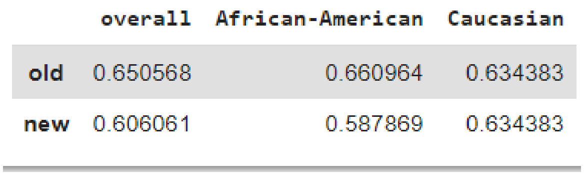
Advantages:

1. The method is interpretable; we can easily understand how the prediction was made and what considerations were taken into account. We can choose which samples will have greater importance in the prediction of the risk, and which would have less by simply changing the normalization factor.

2. It could be applied post-training, meaning if we got a dataset of a model which was already trained. We can achieve fairer results on the given prediction he made without applying different methodologies to recreate the model itself in order to train it on the data with different weights for each group.

Disadvantages:

1. The main problem with the method is that we treat all African American defendants as one group. We can have a African American defendant who did recidivate in the past two years and the method will prioritize him over a Caucasian defendant even though he does not deserve it, just based on his affiliation to the African American race.



	overall	African-American	Caucasian
old	0.650568	0.660964	0.634383
new	0.606061	0.587869	0.634383

2. Another problem of the method is that it could over-prioritize one group over the other since it is based on a small sample and not the actual distribution in the population.
3. When we are changing the distribution of the risk of African American defendants to re-offend in the model, we can lower the accuracy of the model and/or the accuracy of the subgroups.

5.5. Method 3 - SGD With Weighted Samples:

SGD Classifier: Stochastic Gradient Descent (SGD) is a simple, yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function.
In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression.

The advantages of Stochastic Gradient Descent are Efficiency and Ease of implementation

The main disadvantage of Stochastic Gradient Descent is his sensitivity to feature scaling - data normalization. So, it is highly recommended to scale your data.
Note that the same scaling must be applied to the test vector to obtain meaningful results.

The observed probabilities in the data before applying any solution were:

$$TPR_{African\ American} = 0.6789$$

$$TPR_{Caucasian} = 0.373$$

The diffrence: 0.3058

$$FPR_{African\ American} = 0.3327$$

$$FPR_{Caucasian} = 0.1554$$

The diffrence: 0.1773

Let's observe the distribution of the predictions of African American and Caucasian defendants:

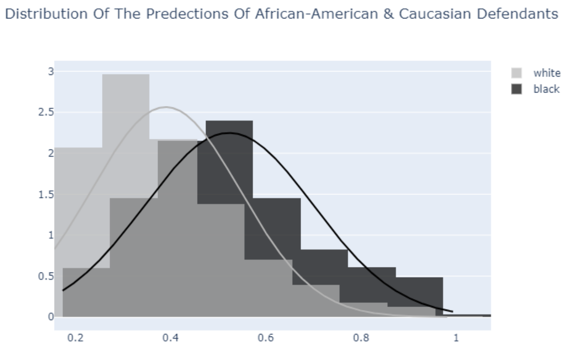


Figure 8: Distribution Of The Predections Of African-American & Caucasian Defendants with SGD model

We can clearly see that there is a major difference between the two races.
The probabilities of African American defendants are much higher of Caucasian defendants since the distribution of African American is to the right of the Caucasian defendants' distribution.

We see that with the traditional model we can't achieve Equalized Odds parity, so what can we do?

Importance sampling in SGD:
Importance sampling prioritizes training examples for SGD in a principled way. The technique suggests sampling example i with probability proportional to the norm of loss term i's gradient. This distribution both prioritizes challenging examples and minimizes the stochastic gradient's variance.
We can apply Importance sampling in our data to prioritize African American defendants who did not recidivate over Caucasian defendants who did not recidivate in order to achieve Equalized Odds parity

African American-Priors:

$$P(Recid|African\ American) = \frac{P(African\ American, Recid)}{P(African\ American)} = 0.53$$

$$P(no\ Recid|African\ American) = \frac{P(African\ American, no\ Recid)}{P(African\ American)} = 0.47$$

Caucasian-Priors:

$$P(Recid|Caucasian) = \frac{P(Caucasian, Recid)}{P(Caucasian)} = 0.39$$

$$P(no\ Recid|Caucasian) = \frac{P(Caucasian, no\ Recid)}{P(Caucasian)} = 0.61$$

For example: If we have a have an African American man which recidivated, the proportion in the population out of all African American defendants is 0.53, so we are going to give him a weight of 0.39 (The proportion of Caucasian defendants who recidivated in the population out of all Caucasian defendants)

If we have a have an African American man which did not recidivated, the proportion in the population out of all African American defendants is 0.47, so we are going to give him a weight of 0.61 (The proportion of Caucasian defendants who did not recidivated in the population out of all Caucasian defendants)
Similarly, we will do for Caucasian defendants.

After applying the method, we can see that in the train dataset we achieved:

$$TPR_{African\ American} = 0.3968$$

$$TPR_{Caucasian} = 0.4021$$

The diffrence: 0. 0052

$$FPR_{African\ American} = 0.1284$$

$$FPR_{Caucasian} = 0.1689$$

The diffrence: 0.0405

We can see that, the difference between the probabilities in the train dataset are really small

Let's observe the distribution of the predictions of African American defendants and Caucasian defendants with the method on the train dataset:

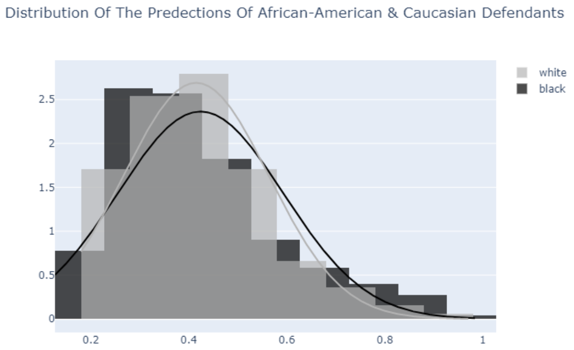


Figure 9: Distribution Of The Predections Of African-American & Caucasian Defendants with the method on the train dataset

We can see from the graph that the distribution of the African American and Caucasian defendants is very closely aligned together, which means that the difference between the probabilities in the train dataset are really small.

To validate the results, in the train dataset we achieved:

$$TPR_{African\ American} = 0.3689$$

$$TPR_{Caucasian} = 0.369$$

The diffrence: 0. 0001

$$FPR_{African\ American} = 0.1238$$

$$FPR_{Caucasian} = 0.1755$$

The diffrence: 0.0517

Let's observe the distribution of the predictions of African American defendants and Caucasian defendants with the method on the test dataset

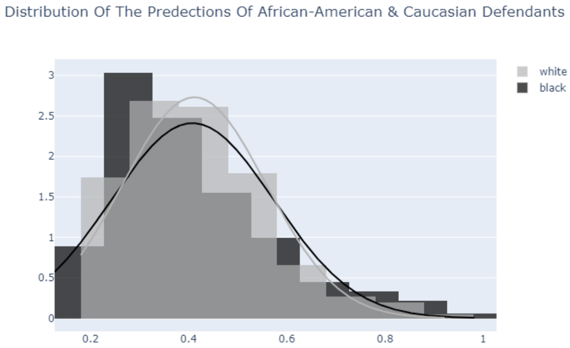


Figure 7:Distribution Of The Predections Of African-American & Caucasian Defendants with the method on the test dataset

We can see from the graph that the distribution of the African American and Caucasian defendants is very closely aligned together, which means that the difference between the probabilities in the test dataset are really small as well.

Since the difference in both the train dataset and the test dataset are really small, we can say that we have reached Equalized Odd parity.

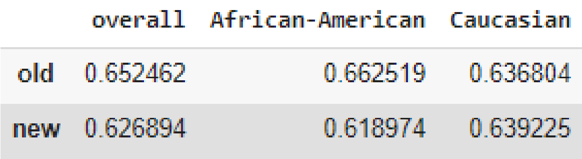
Advantages:

1. The main advantage of the method is that we do not treat each race as one separate group, we can still prioritize African American defendants over Caucasian defendants while considering whether they did recidivate or not. The method weights 4 kinds of groups: Caucasian defendants who recidivate\ not recidivate and African American defendants who recidivate\ not recidivate. Thus, we can assure that we do not prioritize defendants just based on their affiliation to the African American race.

2. The method is interpretable, we can easily understand how the prediction was made and what considerations were taken into account. We are able to choose which samples will have greater importance in the prediction of the risk and which would have less.

Disadvantages:

1. When we are changing the distribution of the risk of African American and Caucasian defendants to re-offend in the model, we can lower the accuracy of the model and/or the accuracy of the subgroups.



	overall	African-American	Caucasian
old	0.652462	0.662519	0.636804
new	0.626894	0.618974	0.639225

2. Another problem of the method is that it could over-prioritize one group over the other since it is based on a small sample and not the actual distribution in the population.
3.

6. General Legal Commentaries On The Methods:

The legal literature and existing laws do not allow the software to classify the algorithm according to the characteristics of race, sex or religion. However, in order to promote affirmative action using the algorithm itself, we need to take these variables out and control them so that we can promote substantive equality.

6.1. Transparency and accountability:

Transparency and accountability are core principles in justice. Judges explain their decisions in a way that can be reviewed and evaluated by others. The lack of transparency and accountability with its algorithm is a barrier to COMPAS's integration into the criminal justice system. Allowing closed source software such as COMPAS to assign a recidivism rate to defendants, without any oversight or review, has the potential to violate human rights. We want supervision from an "objective" factor that can explain how decisions are made, especially when it comes to human rights.

6.2. New Governance Model:

Therefore, a new **governance model** should be promoted in which the legal system is aware of the vulnerabilities of the system and treats different sensitivity-rates between different details. In other words, one must recognize the different characteristics that each group has, adopt them, and then give equal weight to different parameters between the groups.

6.3. Human In The Loop:

In addition, adding a **human in the loop**, by allowing judges to use the information at their discretion, might solve bias in specific cases, since the algorithm prediction will be taken into account in addition to other indicators that are external to those examined by the algorithm. However, it can also lead to relying on the data without truly testing the algorithm.

6.4. Other:

In justice, algorithms are increasingly being used to modernize practices, reduce bias, and deliver justice. Algorithmic-based sentencing is one application of algorithms in the law that is constitutionally, technically, and morally troubling. In sentencing decisions, states are increasingly allowing judges to consider actuarial risk assessment scores.

From a moral standpoint, the Wisconsin Supreme Court's decision on COMPAS, a sentencing algorithm, was troubling. **It could be said that the decision to let an algorithm play a role in taking away someone's freedom was against the Constitution.**

With a better framework, algorithms might be able to help the justice system without causing constitutional, technical, or moral problems.

Thus, we would like to conclude this part by highlighting **the need for supervision and regulation over private companies and the need to include ethics and responsible AI inside the methods** since they have an effect on our human rights.

Conclusion: Given the core civil rights that are at stake, like freedom of movement, liberty, human dignity, autonomy, and due process, it is crucial that the COMPAS algorithm accounts for affirmative action while making its decisions.

Acknowledgement: We want to thank Prof. Avigdor Gal, Prof. Niva Elkin-Koren and M.Sc Shir Lissak for guiding us and sharing there wisdom and experience in the field.