Fairness in AI

Alex Eaton, Tushar Menon, Ethan Ariowitsch

Machine Bias

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Public and private sectors are increasingly turning to AI and machine learning algorithms to automate both simple and complex decisions

- The availability of data has allowed for the training of AI for use in a wide variety of sectors
  - Entertainment
  - Finance
  - Criminal justice
  - …
- While well suited to some tasks, it is difficult to define what makes an algorithm a fair and objective decision maker

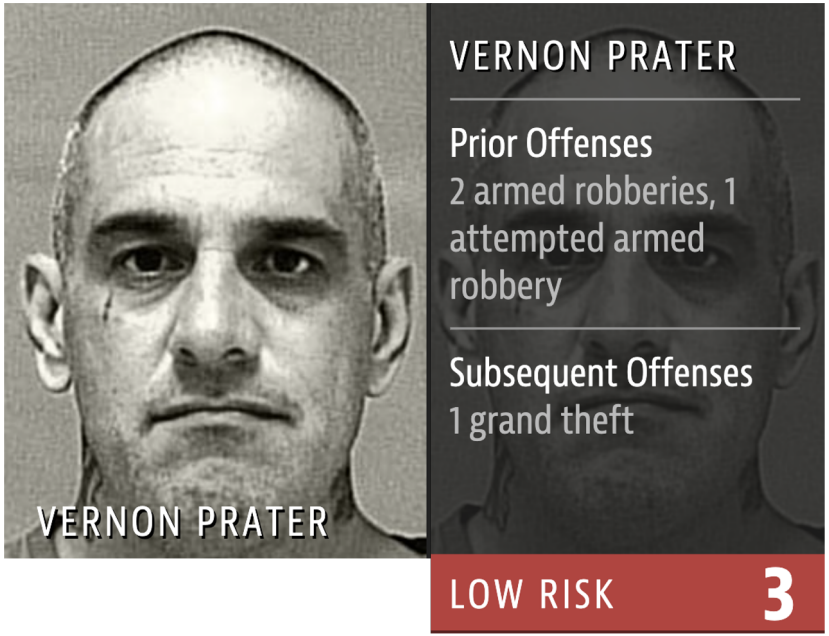
Background

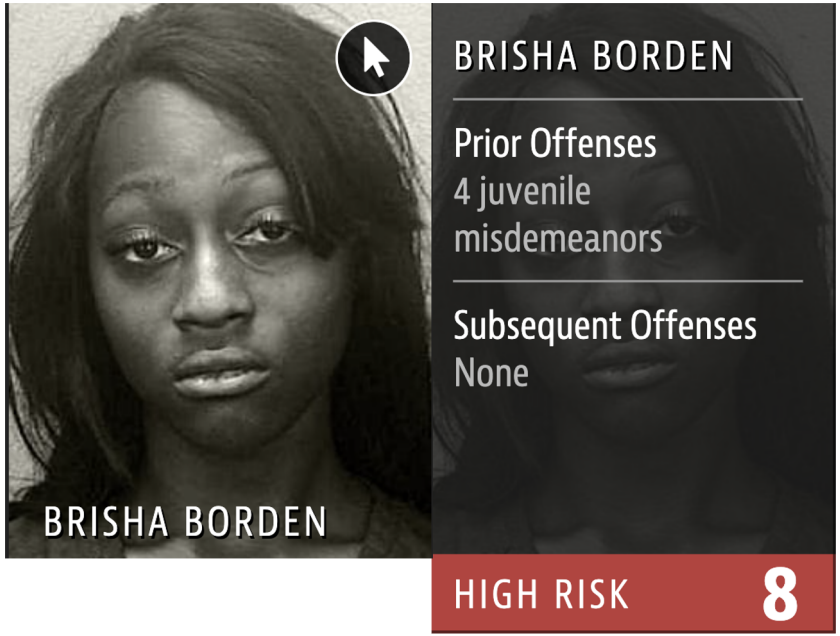
- Courts across the nation have incorporated “risk-assessment” algorithms that are designed to predict a defendant's likelihood of reoffending for a crime
  - Judges can see risk scores during sentencing scores in:
    - Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin
- The Justice Department’s National Institute of Corrections encourages the risk assessments combined with an evaluation of a defendant’s rehabilitation needs at every stage of the criminal justice process.
- A sentencing reform bill pending in Congress would mandate the use of such assessments in federal prisons.

Prediction algorithms can reflect and sometimes amplify pre-existing biases in training data

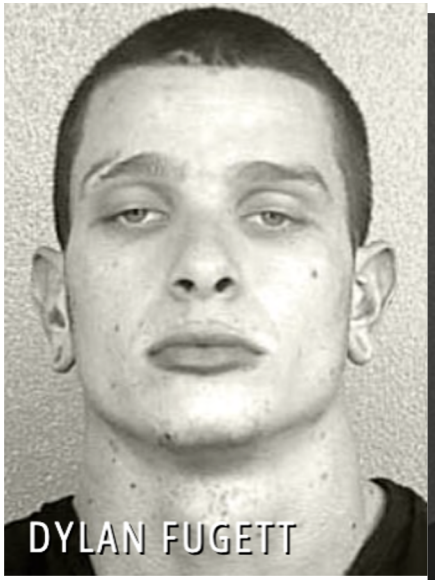
- It can be difficult to define whether an algorithm is fair
  - Just because an algorithm reflects the training data does not mean that it gives fair results
- If the training data is a product of systemic bias, algorithms trained on the model will perpetuate these biases unless corrected
- Two defendants in Coral Springs, Florida were charged with petty theft and received risk scores that did not make sense relative to one another considering the severity of the crime and their respective criminal histories

Two defendants both arrested for petty theft





Two defendants both arrested for petty theft

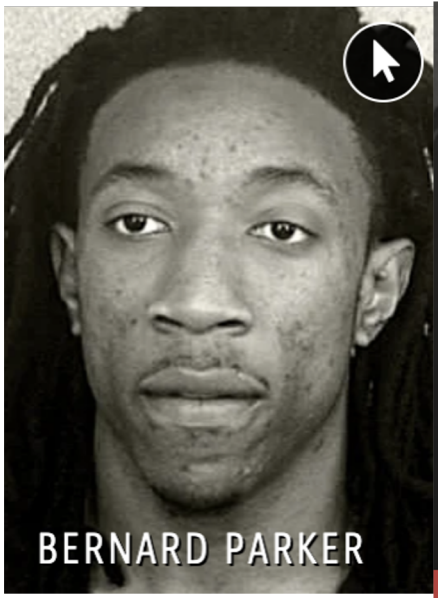


DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3



BERNARD PARKER

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

HIGH RISK 10

Despite clear issues with risk scores in some instances there have been few independent studies that have investigated racial bias in these algorithms

- The company that creates the algorithm responsible for the risk scores is Northpointe a private, for-profit company
- Northpointe maintains that the algorithm's scores are based on a series of questions that do not include the defendant's race
- The survey asks 137 questions like
  - "Was one of your parents ever sent to jail or prison?"
  - How many of your friends/acquaintances are taking drugs illegally?
  - How often did you get in fights while at school?
  - A hungry person has a right to steal (agree or disagree)

Northpointe and COMPAS

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was developed by Northpointe founders Tim Brennan and Dave Wells
  - It is intended assesses risk as well as two dozen categories representing “criminogenic needs” related to the major theories of criminality,
    - criminal personality
    - social isolation
    - substance abuse
    - residence/stability
- Defendants are ranked low, medium or high risk in each category.

Studies

- An assessment of how various risk scores are studied found that most risk score methodologies and that “in most cases, validity had only been examined in one or two studies”
  - Furthermore often the authors of the papers helped develop the model
  - The paper also found that through 2012 the risk assessment tools were only moderately effective
- New York adopted the Northpointe algorithm for a pilot program for assessing people on probation in 2010
  - Conducted a study in 2012 that found the algorithm had 71% accuracy and did not address racial discrepancies
- One 2016 study looked at 35,000 scores for Northpointe’s algorithm
  - concluded that although Black defendants had higher scores on average that the difference was not attributable to bias

Propublica investigated the results of the algorithm using data from Broward County, FL

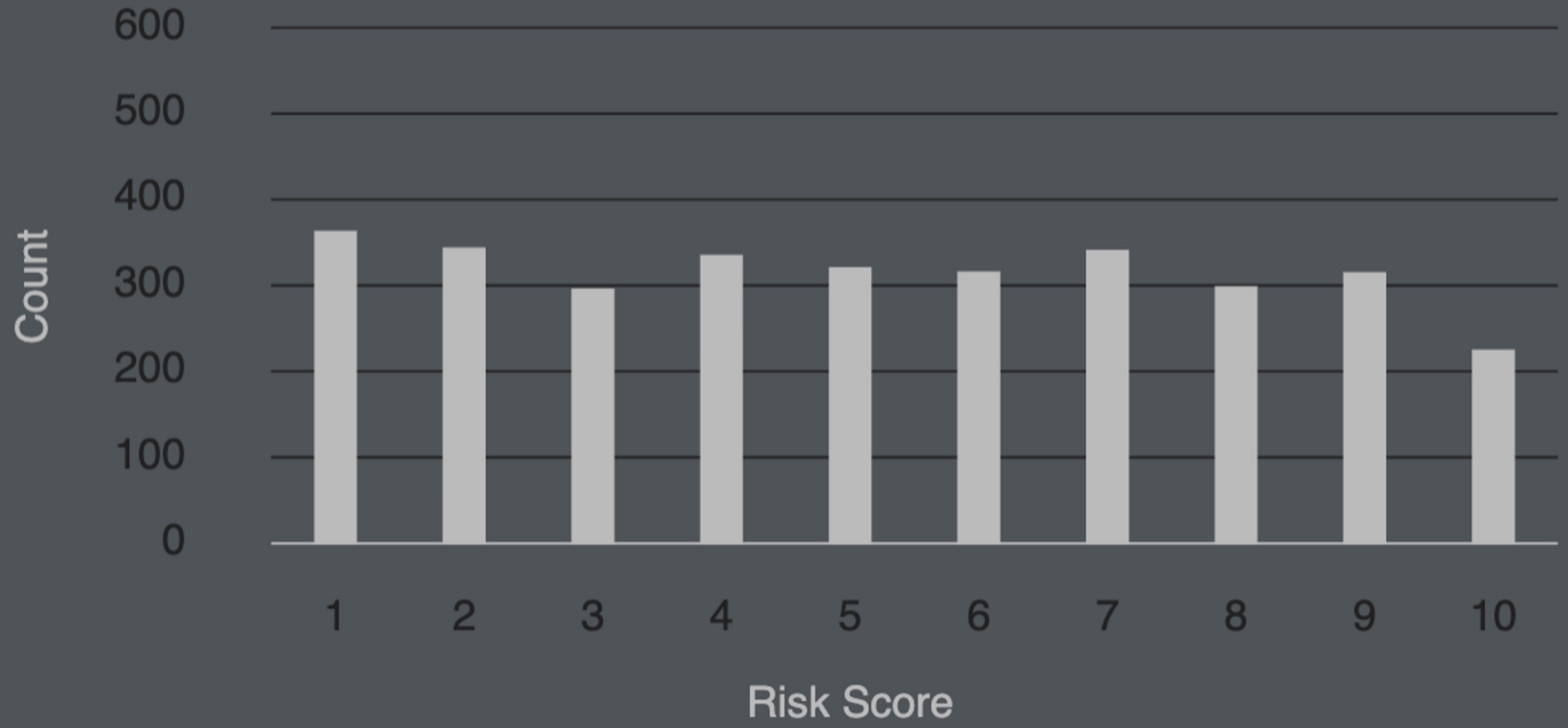
- 7,000 risk scores from Broward County
- From 2013-2014
- Propublica checked to see how many defendants reoffended in the following two years
    - This is the benchmark defined by the creators of the algorithm

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

ProPublica findings

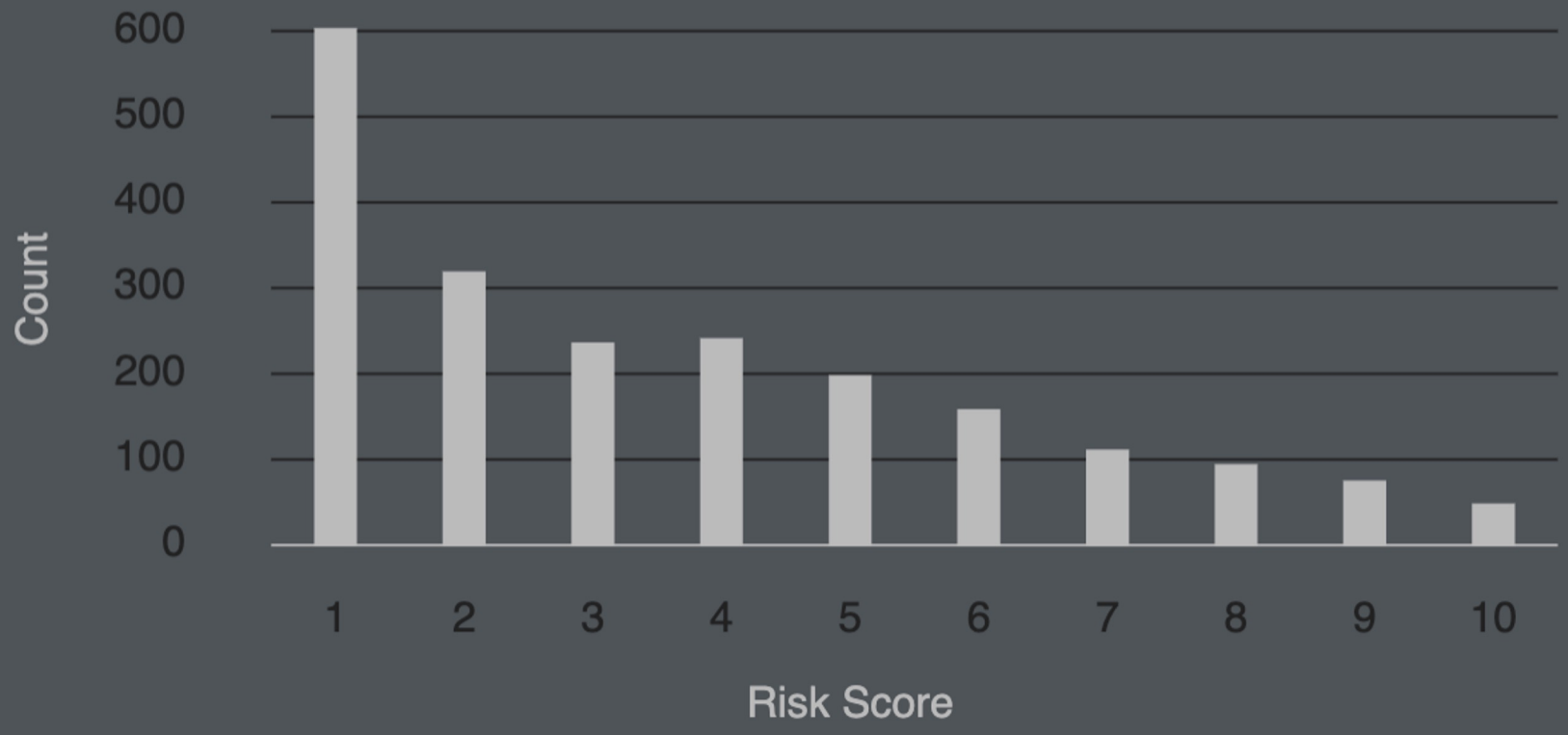
- ProPublica found that only 20% of people who were predicted to commit new violent crimes in the next two years actually did so
- For all crimes (including misdemeanors) the rate for successfully predicting re-offenders in two years was 61%
  - Only a little better than a coin flip
- There were also racial disparities in the algorithms predictions
  - The algorithm made mistakes at the same rate regardless of race
  - However the type of mistakes varied dramatically between black and white defendants on average

Black Defendants' Risk Scores



Source: ProPublica analysis of data from Broward County, Fla.

White Defendants' Risk Scores



Source: ProPublica analysis of data from Broward County, Fla.

ProPublica findings cont.

- The algorithms was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants
  - Black defendants were flagged at almost twice the rate
- The algorithms was also much more likely to mislabel white defendants as low risk than black defendants
- ProPublica then isolated the effect of race from criminal history and recidivism, age and gender and reran the tests
  - Black defendants were still 77% more likely to be flagged for violent re-offenders
  - Also 45 percent more likely to be predicted to commit a future crime of any kind.

Risk Scores in Courts

- In theory judges are not supposed to use risk scores when determining sentencing
- Risk scores are intended to be used for determining probabation and treatment programs
- However judges have in many states can still view scores and some have even cited scores in sentencing decisions
  - In August 2013, a Wisconsin judge declared that a defendant “identified, through the COMPAS assessment, as an individual who is at high risk to the community.”
    - The judge then imposed a sentence of 8.5 years in prison
  - Wisconsin Assistant Attorney General Christine Remington stated “The risk score alone should not determine the sentence of an offender” in a State Senate meeting in 2016

NorthPointe vs Propublica: What Does It Mean For An Algorithm to be Fair?

Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

Background and Motivation

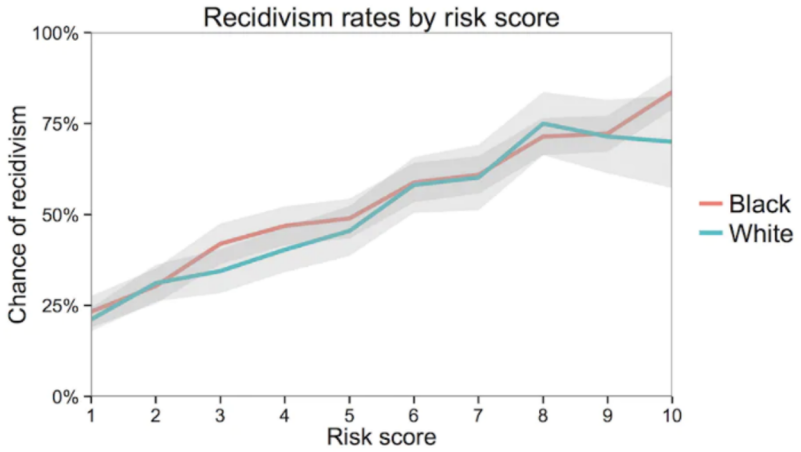
- COMPAS is an algorithm used to help make sentencing and bail decisions
- ProPublica claims COMPAS is biased against black people
- Northpointe (creator) denies those claims
- Is COMPAS fair? What is fairness?

COMPAS

- Assigns defendants scores from 0 to 10 that indicate how likely they are to reoffend
- Score is based on over 100 features: age, sex, criminal history etc. (but not race)
- Race however is correlated with many features that the score does rely on

Fair?

- Northpointe argues: Since the scores mean the same thing regardless of race, they are indeed fair. ex. Of all defendants who got a 7, 60% of the white defendants reoffended and 61% of the black defendants reoffended

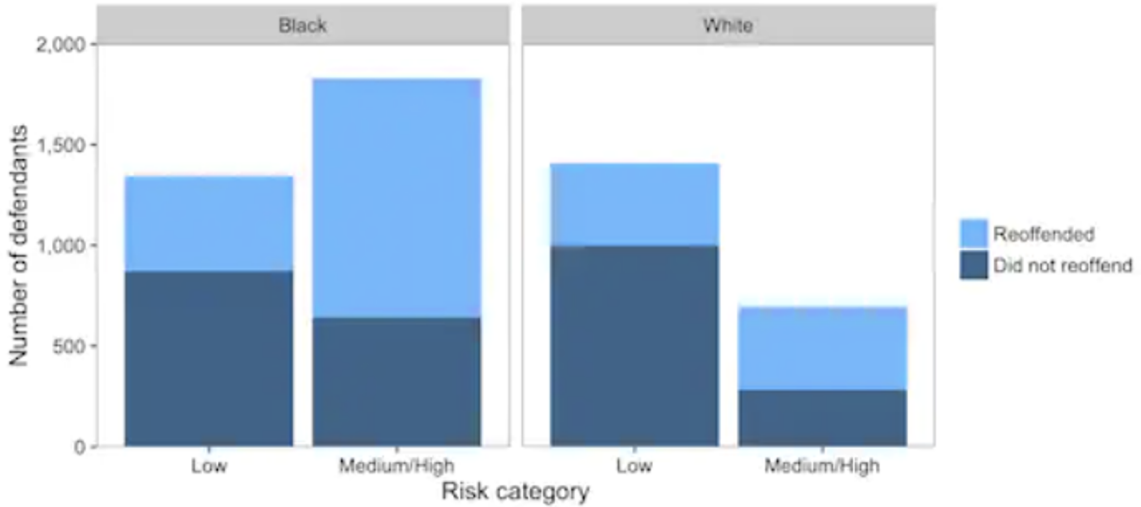


Not fair?

- Propublica argues: among the defendants who did not reoffend, black people were more than twice as likely to be classified as high risk
- Even though they didn’t reoffend, their sentencing and bail decisions were more severe

Who is correct?

- The score cannot satisfy both Northpointe and ProPublica



Who is correct?

- Northpointe: Within each risk category, the ratio of defendants who reoffend is the same regardless of race
- The overall recidivism rate for black people is higher than for white people (52% vs 39%)
- The algorithm is more likely to classify black people as high risk vs white people (58% vs 33%)
- Propublica: Black people who don’t reoffend are more likely to be classified as high risk than white people who don’t reoffend

Who is correct?

- Is Northpointe’s interpretation wrong? It’s hard to feel that way
- What about ProPublica?

Imbalanced classification

- ‘Performance’ of a classifier can mean several things

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Imbalanced Classification

- Precision: TP / (TP + FP)
- Inverse Recall: FP / (FP + TN)

What to do?

- There are multiple measures of performance of a classifier, which each reflect different values
- When we use a classifier whose predictions have such gravity, we have to think about what are the things we value
- Would we rather put dangerous people in jail, running the risk of also putting harmless people in jail?
- Or would we rather ensure no harmless people get put in jail, running the risk of letting dangerous people roam free?

What to do?

- We must explore alternative policies - not just algorithmic, but we have to examine the sociological conditions and reimagine the system that raises the question in the first place
- Ex. Why is the recidivism rate higher for black people than white people?
- Ex. Instead of using an algorithm to decide which defendants must pay bail, why not end bail requirements altogether?

Discussion

1. Why do you think the recidivism rates for black people are higher than for white people?
2. Do you think the values reflected by Northpointe's measure of classifier fairness (same precision for both races) are reflective of the US judicial system?

Extra

- Runaway Feedback Loops in Predictive Policing (Ensign et al. 2018)
- US has highest incarceration rate of any country in the world (WPR)
- Capitalizing on Mass Incarceration: U.S. Growth in Private Prisons (Gotsch, Basti 2018)

Fair prediction with disparate impact

Alexandra Chouldechova

Background

- This paper is essentially a smarter, more robust way of analyzing the “paradox” we discussed in the previous paper
- Connects test fairness and classification error
- Shows that using a classifier that has different FNR and FPR across groups leads to disparate impact when the positive group has stricter penalties

Test fairness (psychometrics)

- A score S is fair if the probability of positive prediction is the same despite the group membership of the person (race)

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w).$$

Test fairness (psychometric)

	$S_c = \text{Low-Risk}$	$S_c = \text{High-Risk}$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

$$\text{PPV}(S_c \mid R = r) \equiv \mathbb{P}(Y = 1 \mid S_c = \text{HR}, R = r)$$

Recidivism prevalence:

$$p_r \equiv \mathbb{P}(Y = 1 \mid R = r).$$

Test fairness (psychometric)

- Given PPV, if I vary p across groups, they will have different FPR and FNR

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR}).$$

Different Impacts

$$\begin{aligned}\Delta &\equiv \mathbb{E}_{\text{MinMax}}(T_{b,y_1} - T_{w,y_2}) \\ &= (t_H - t_L)\big[\mathbb{P}(S_c = HR \mid R = b, Y = y_1) \\ &\qquad - \mathbb{P}(S_c = HR \mid R = w, Y = y_2)\big]\end{aligned}$$

- Given recidivism rate difference across groups and test fairness constraints, it will be the case the group with higher recidivism rate will have higher FPR and lower FNR. So greater penalties for defendants in that group

Inherent Tradeoffs in the Fair Determination of Risk Scores

Kleinberg et al.

Problem domains

- Many problem domains in which decision criteria should be uniform across groups
- Criminal justice system
  - Consistent scores across groups
- Advertising
  - Equal advertisement probabilities across groups
- Medical testing and diagnosis
  - Medical decisions uniform across groups

Structural commonalities

- Different problem domains will have similar structures
  1. Algorithmic estimates are used as input to larger decision framework
     - Risk scores, test scores, machine learning outputs
  2. Underlying task is classifying whether people possess a relevant property
     - Recidivism, medical condition, interest in product
  3. Algorithmic estimates are not binary yes/no, but rather probability estimates about if given person is positive or negative class

Three Fairness Criteria

- Well calibrated
  - Algorithm's assigned probability for a positive class should accurately reflect the percentage of positive instances in the population

- Positive class balance
  - Average score of people in positive classes should be the same across groups

- Negative class balance
  - Average score of people in negative classes should be the same across groups

Trade offs among guarantees

- All three of the fairness conditions seem to variants of the same goal
- Therefore, it should be easy to satisfy all three at the same time, right?
- **WRONG** - these fairness conditions are incompatible with each other, and can only be simultaneously satisfied under very specific cases

Formulating the problem

- For any problem domain, there is a group of people who constitutes a positive or negative instance of the classification problem
- *Positive Class* - people who constitute positive instances
  - E.g. defendants who will be arrested again
- *Negative Class* - people who constitute negative instances
  - E.g. defendants who will not be arrested again
- The decision algorithm attempts to estimate these classes

Formulating the problem

- Feature vectors
  - Each person has a feature vector σ representing data we know about them. p_σ is the fraction of people with feature vector σ who belong to the positive class.
- Groups
  - Each person belongs to group 1 or 2. Both groups have different distributions over features vectors, but people from each group have the same probability p_σ of belonging to the positive class assuming their feature vector is σ .
- Risk Assignments
  - Risk assessments are ways of dividing people into bins b with associated score v_b using a person's feature vector σ. These assignments require a mapping from feature vectors to bins.

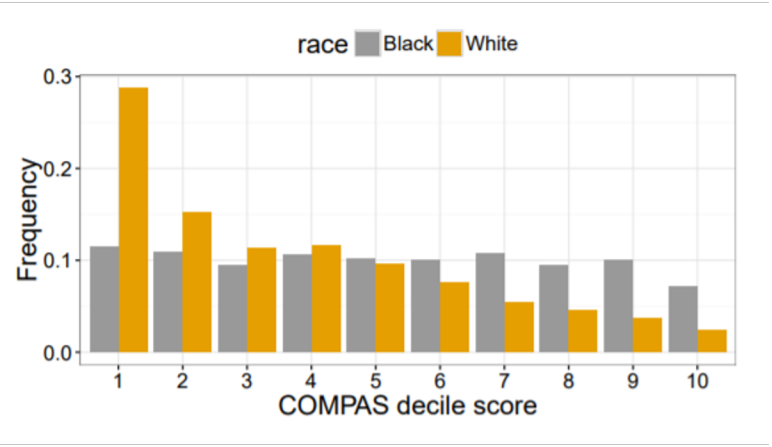
Fairness properties for risk assignments

- Calibration within groups
  - For each group *t* and each bin *b* with score v_b , the expected number of people from group *t* in *b* who belong to the positive class should be a v_b fraction of the expected number of people from group *t* assigned to *b.*
- Balance for positive class
  - Average score of people from group 1 and 2 assigned to positive class should be equal
- Balance for negative class
  - Average score of people from group 1 and 2 assigned to negative class should be equal

Why do these conditions correspond to fairness

- Calibration
  - This condition essentially means that scores actually mean what they say they mean
  - Lack of calibration between groups incentivizes decision makers to treat people of different groups with the same score differently
- Positive/Negative class balance
  - This means that if two people in different groups exhibit similar future behavior (positive or negative), then they should be treated similarly by the procedure
  - Violating this could lead to some groups receiving consistently higher/lower scores

Fairness Example



- COMPAS algorithm
- Angwin et al. observe that COMPAS breaks fairness criteria 2 and 3
    - Average score of white defendants in positive and negative classes lower than average score of black defendants
- Counter-arguments established that COMPAS upholds the first calibration criteria
- Although the first criteria is very important, it could be equally crucial that the other criteria aren’t met depending on the problem domain

Determining what is achievable

The three fairness conditions can only be met in two simple cases

1. *Perfect Prediction*
   - For each feature vector, we have $p_\sigma = 0$ or $p_\sigma = 1$, giving us perfect prediction. Assign feature vectors with $p_\sigma = 0$ to a bin with score $v_b = 0$ and feature vectors with $p_\sigma = 1$ to a bin with score $v_b = 1$. This satisfies all three fairness conditions.
2. *Equal Base Rates*
   - The two groups have the same fraction of members in the positive class; that is, the average value of p_σ is the same for group 1 and 2 . In this case, make a single bin with score p_σ and assign everyone to this bin. This satisfies all three fairness conditions.

Theorem 1

Theorem 1.1 *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction (with p_σ equal to 0 or 1 for all σ) or have equal base rates.*

- This theorem states that for any instance that is more complex than the two simple cases above, at least one of the fairness criteria will be violated by the risk assignment
- This result holds regardless of how the risk assignment is computed
  - Under this framework, risk assignments are arbitrary functions from feature vectors to bins, therefore it applies independently of the method that's used to construct the risk assignment

Theorem 2

Theorem 1.2 *There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\varepsilon > 0$, and any instance of the problem with a risk assignment satisfying the ε-approximate versions of fairness conditions (A), (B), and (C), the instance must satisfy either the $f(\varepsilon)$-approximate version of perfect prediction or the $f(\varepsilon)$-approximate version of equal base rates.*

- The results of the first theorem can be relaxed in a continuous fashion where the fairness conditions are only approximate
  - For any ε > 0 , ε-approximate versions of each condition require that the equalities between groups only hold to an error of ε
- In other words, anything that approximately satisfies the fairness conditions must approximately look like the two above simplified cases
- Intention of these theorems is to establish unavoidable tradeoffs, not to make recommendations on how fairness conflicts should be handled

Conclusions

- There exist three basic criteria of fairness of equal importance for these predictive algorithms
- It is not possible to satisfy all at once and there are inherent tradeoffs in prioritizing each criteria
- These results hold regardless of problem domain and method to compute risk assignment
- These tradeoffs are not well understood

Conclusions

- There may be alternate settings to these problem domains in which algorithmic mistakes may be weighted differently
  - E.g. the cost of false negatives may be much greater than the cost of false positives
  - Could correspond to finding risk assignments that satisfy the calibration condition, and one of the balance conditions
- Up to the algorithm creators to determine which criteria are the most important for their problem

Discussion

- Now knowing that it's impossible for these algorithms to satisfy all fairness conditions, have your initial opinions on using algorithms in the criminal justice system changed at all?

- What are some rule/procedures you can think of that might promote fairness when training and implementing AI?

- What are ways in which we can reimagine the role of algorithms in society?
  - E.g. algorithm use in policing, advertisement, etc.

Takeaways

- Predictive scoring algorithms can be used across many different problem domains, and often hold a concerning amount of decision power
- In all practical cases, it’s mathematically impossible for these algorithms to be truly unbiased between groups
- The obligation to explore alternative policies if we recognise that our algorithms are unfair - solution doesn’t always have to be an algorithm