What does it mean for an algorithm to be fair?

Please read this article:

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Please take a short survey!

Algorithms increasingly help make critical decisions

Criminal justice

Medical diagnosis

Important that algorithmic decisions be *fair*

But what does it mean for an algorithm to be fair?

Nuanced, controversial question!

Example: COMPAS criminal risk prediction algorithm

Algorithm helps judges make bail and sentencing decisions

Assigns defendants a score from 1 - 10 based on how likely they are to commit another crime. Scores of 1-4 = low risk.

Does not use a defendant’s race

Is this algorithm *fair*?

News organization ProPublica decides to investigate…

Who thinks, after reading the article, that COMPAS algorithm is fair?

Is COMPAS algorithm fair?

ProPublica: **No**

Observation	**Fairness principle**
Black defendants are more likely than white defendants to be classified as high risk	*Statistical parity* (equal rates across groups)

Is COMPAS algorithm fair?

ProPublica: **No**

Observation	**Fairness principle**
Black defendants are more likely than white defendants to be classified as high risk	*Statistical parity* (equal rates across groups)
Black defendants *who do not commit another crime* are more likely than white defendants *who do not commit another crime* to be classified as high risk.	*Predictive equality* (equal false positive rates across groups)

Is COMPAS algorithm fair?

Northpointe: **Yes**

Observation	**Fairness principle**
Black defendants are more likely than white defendants to be classified as high risk	*Statistical parity* (equal rates across groups)
Black defendants *who do not commit another crime* are more likely than white defendants *who do not commit another crime* to be classified as high risk.	*Predictive equality* (equal false positive rates across groups)
Algorithm does not use race	*Fairness through blindness* (do not use sensitive feature)

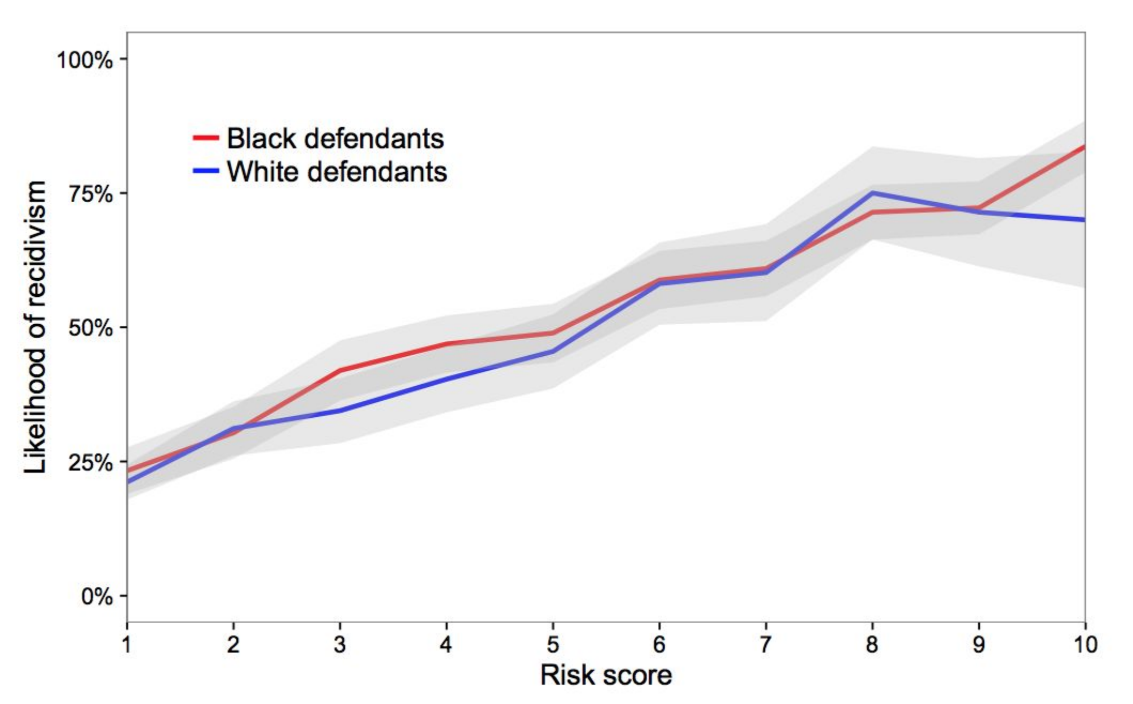
Is COMPAS algorithm fair?

Northpointe: **Yes**

Observation	**Fairness principle**
Black defendants are more likely than white defendants to be classified as high risk	*Statistical parity* (equal rates across groups)
Black defendants *who do not commit another crime* are more likely than white defendants *who do not commit another crime* to be classified as high risk.	*Predictive equality* (equal false positive rates across groups)
Algorithm does not use race	*Fairness through blindness* (very weak notion of fairness -- why?)
Black defendants and white defendants with the same score are equally likely to reoffend.	*Calibration* (scores mean the same thing for both groups)

Is COMPAS algorithm fair?

Northpointe: Yes



Is COMPAS algorithm fair?

Academics: AAAAARRRGGGG

- Prove mathematically: it’s generally impossible to satisfy all these definitions of fairness at the same time.
- If:
  - two groups have different probabilities of recidivism
  - your risk scores don’t perfectly predict recidivism
  - risk scores are calibrated (Northpointe’s fairness requirement)
- Then:
  - other fairness requirements (eg, ProPublica’s) will not be satisfied
  - eg, cannot have calibration, equal false positive rates across groups, and equal false negative rates across groups all at the same time
- Often a conflict between **maximizing accuracy** and **minimizing disparities**

Our paper

- imagine you're a judge trying to decide whom to jail before trial.

Assumptions:

- Pay some cost *c* for every defendant you jail
- Pay a cost of 1 for every defendant you free who commits another crime.
- Each defendant has some probability *p* of committing another crime

Whom should you jail?

Unconstrained by fairness: apply a single threshold

Jail every defendant who’s more likely than $p = c$ to commit another crime.

Apply a *single threshold* to all defendants

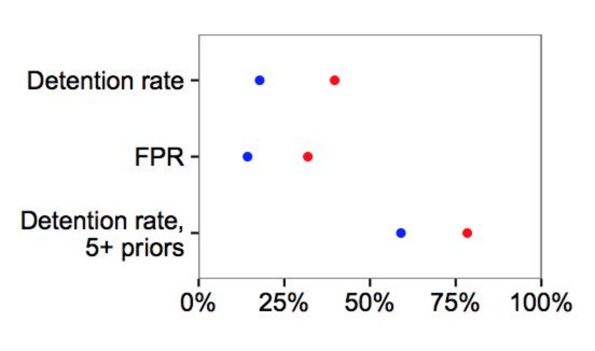
What if you care about satisfying notions of fairness like statistical parity?

Constrained by fairness: apply multiple thresholds

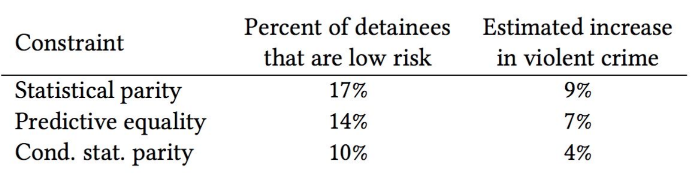
If you want to satisfy statistical parity or predictive equality, your optimal behavior is to apply *multiple, group-specific thresholds*.

Either way has downsides!

Single threshold



Multiple thresholds



Constraint	Percent of detainees that are low risk	Estimated increase in violent crime
Statistical parity	17%	9%
Predictive equality	14%	7%
Cond. stat. parity	10%	4%

Questioning assumptions: utility function

1. *Group* utility function (eg, want a diverse class)
2. Individual-specific costs (eg, jailing a defendant with children)
3. Long-term costs / benefits (eg, preferentially lending to minorities)
4. State has an obligation to repair disparities it helped create?
5. Are you measuring what you want to measure?

“Biased data”

Important to be precise about what you mean!

If the *predictors* are biased, we could potentially correct for this

- eg, if *past arrests* predict future crime less well for black than white defendants, our algorithm could weight arrests less heavily for black defendants

What if the *outcome* is biased? Bigger problem.

- If we’re trying to predict drug crime, but all we have is drug arrests, we have a problem
- Violent crime data may be less likely to be biased

What do we do?!

What do you think?

What do we do?!

simplistic answer: stop using algorithms. These problems apply to all decision-makers, and there is actually evidence algorithms can be less biased than human judges

better answers:

- require **algorithmic transparency**
- **simple** (or at least **interpretable**) algorithms
- can we make decisions that are less costly to get wrong?
- can we explore longer-term ways to reduce societal inequities?

Please take a short survey!