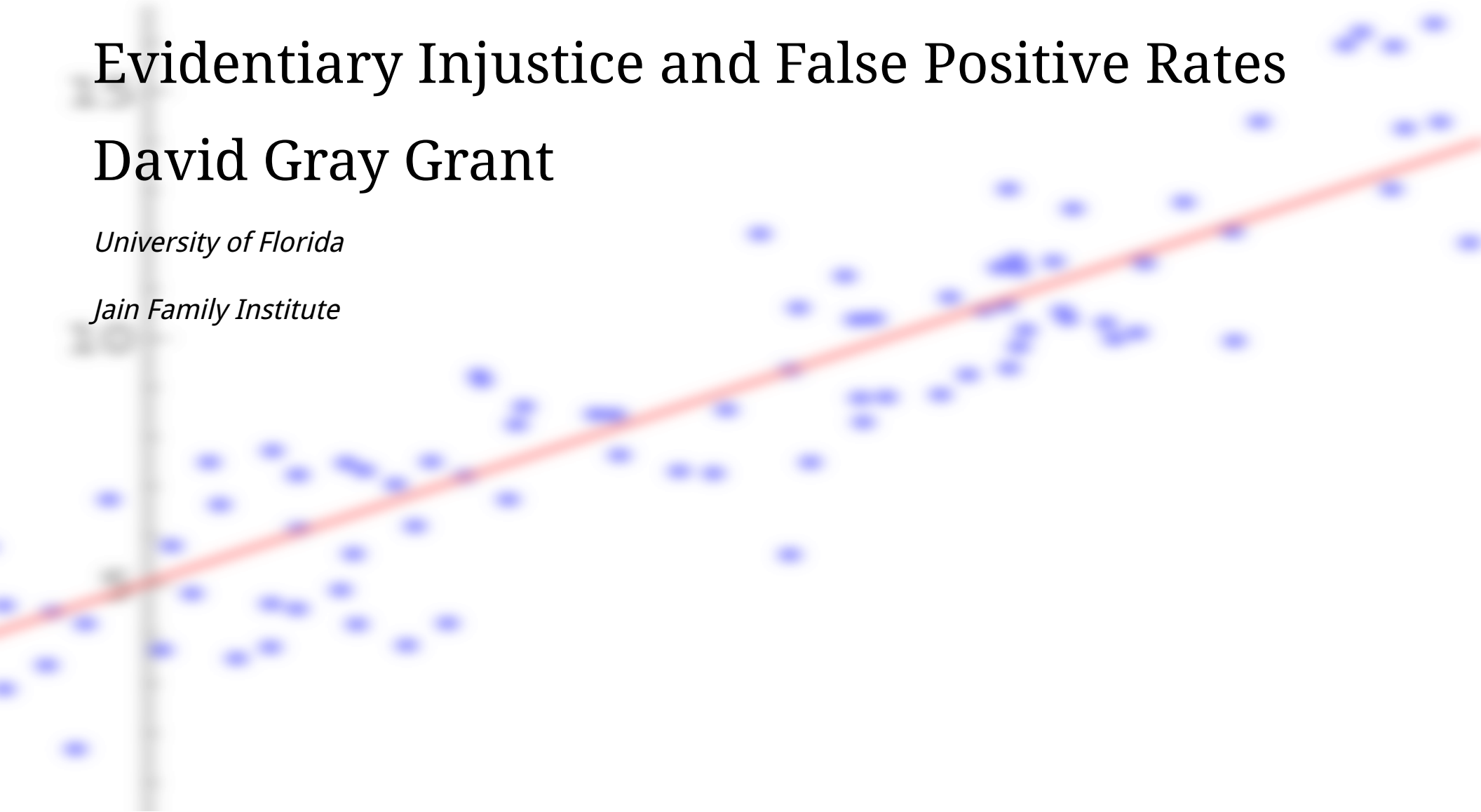
Evidentiary Injustice and False Positive Rates

David Gray Grant

University of Florida

Jain Family Institute



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

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

May 23, 2016

COMPAS

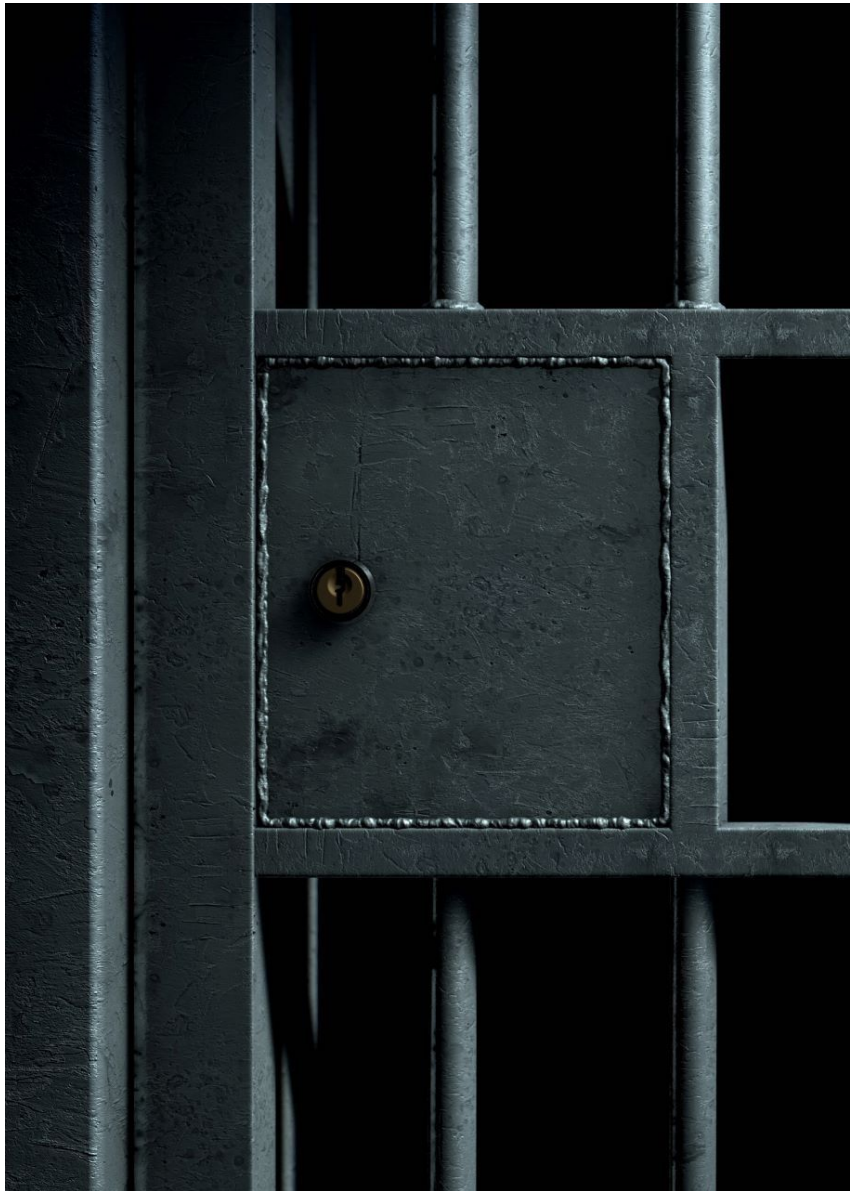
Recidivism prediction instrument

- Estimates **risk of committing a crime** within 2 years
- Outputs **recidivism risk score**
- Widely used in US criminal justice system

Pretrial detention

The justification for pretrial detention is **preventative**

- Detain **high risk** defendants to protect others
- Release **low risk** defendants



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

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

May 23, 2016

ProPublica's analysis

> "[We] turned up significant racial disparities ...
>
> In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.
>
> The formula was particularly likely to **falsely flag black defendants as future criminals**, wrongly labeling them this way at almost twice the rate as white defendants."*

* Angwin et al. (2016), "Machine Bias"

False positive rates

True positive = labeled "high risk," reoffends

False positive = labeled "high risk," does not reoffend

False positive rate = the probability that a randomly selected defendant that does not reoffend (**actual negative**) will be classified as "high risk"

False positive rates

True positive = labeled "high risk," reoffends

False positive = labeled "high risk," does not reoffend

False positive rate = the probability that a randomly selected defendant that does not reoffend (**actual negative**) will be classified as "high risk"

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%

False Positive Rate Equality

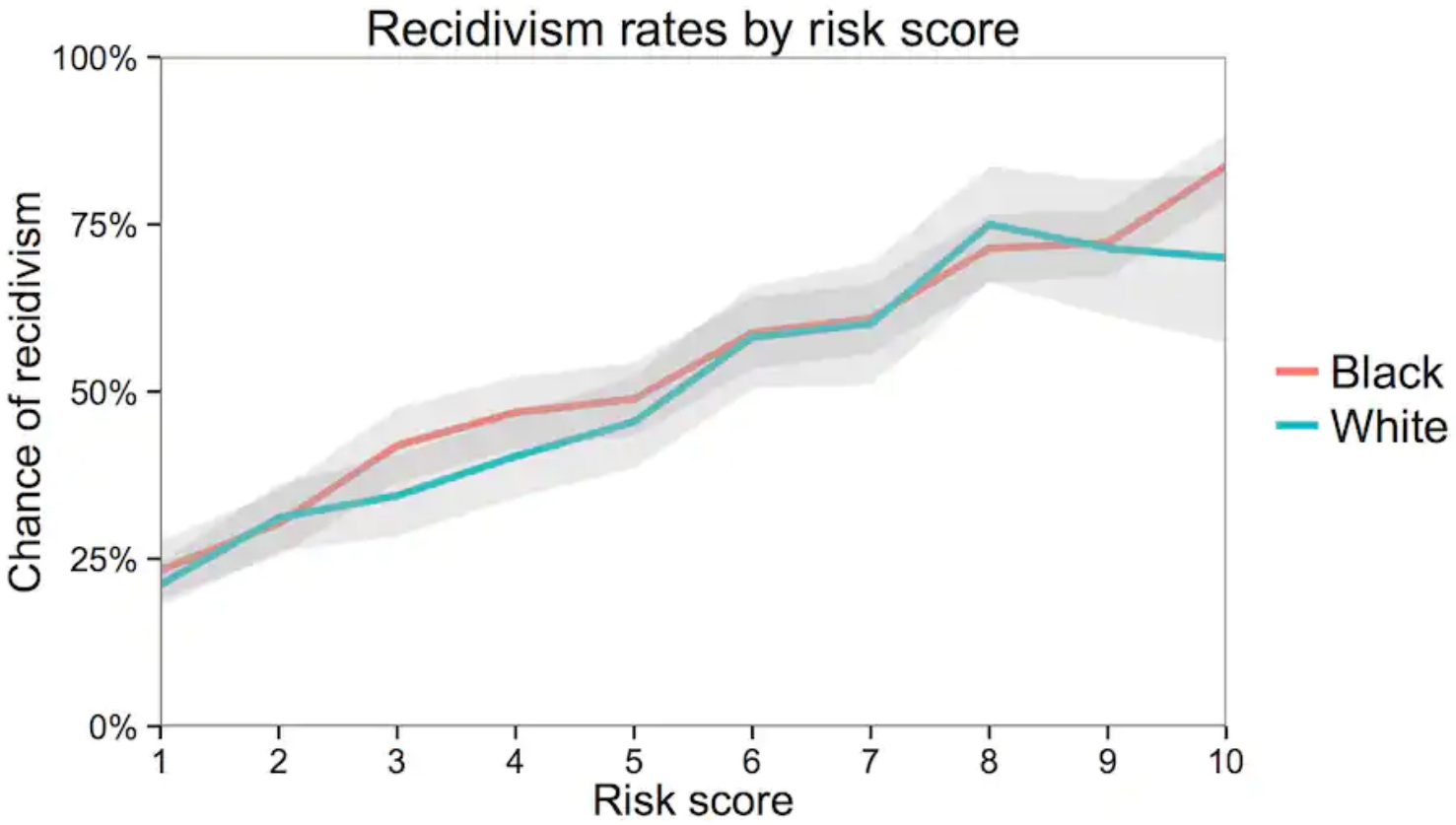
False Positive Rate Equality: the false positive rate for Black defendants is similar to the false positive rate for White defendants

Northpointe's response

Equal Calibration: for each risk score *s*, Black defendants who receive *s* reoffend at the same rate as White defendants who receive *s*

E.g., among defendants who receive a COMPAS score of 7:

- ~60% of Black defendants reoffend
- ~60% of white defendants reoffend


Recidivism rates by risk score
100%
75%
50%
25%
0%
Chance of recidivism
1
2
3
4
5
6
7
8
9
10
Risk score
Black
White

> "Northpointe contends [that COMPAS scores] are indeed fair because **scores mean essentially the same thing** regardless of the defendant's race. Among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended. Consequently, Northpointe argues, when judges see a defendant's risk score, they need not consider the defendant's race when interpreting it."*

* Corbett-Davies et al. (2016)

Fairness requirements

Property P is a **requirement of fairness in recidivism prediction** just in case using a method that lacks P to make preventive detention decisions would be pro tanto unfair

- **Pro tanto unfair** = unfair in some respect

The COMPAS debate

- **ProPublica:** False Positive Rate Equality is a requirement of fairness
- **Northpointe:** Equal Calibration is a requirement of fairness

The COMPAS debate

- **ProPublica:** False Positive Rate Equality is a requirement of fairness
- **Northpointe:** Equal Calibration is a requirement of fairness
- **Impossibility result:** if base rates differ, False Positive Rate Equality are Equal Calibration are incompatible—so fairness is impossible!

The COMPAS debate

- **ProPublica:** False Positive Rate Equality is a requirement of fairness
- **Northpointe:** Equal Calibration is a requirement of fairness
- **Impossibility result:** if base rates differ, False Positive Rate Equality are Equal Calibration are incompatible—so fairness is impossible!

none of this is true

Overview

1. ProPublica's argument
2. ProPublica's argument fails
3. COMPAS is unfairly biased anyway

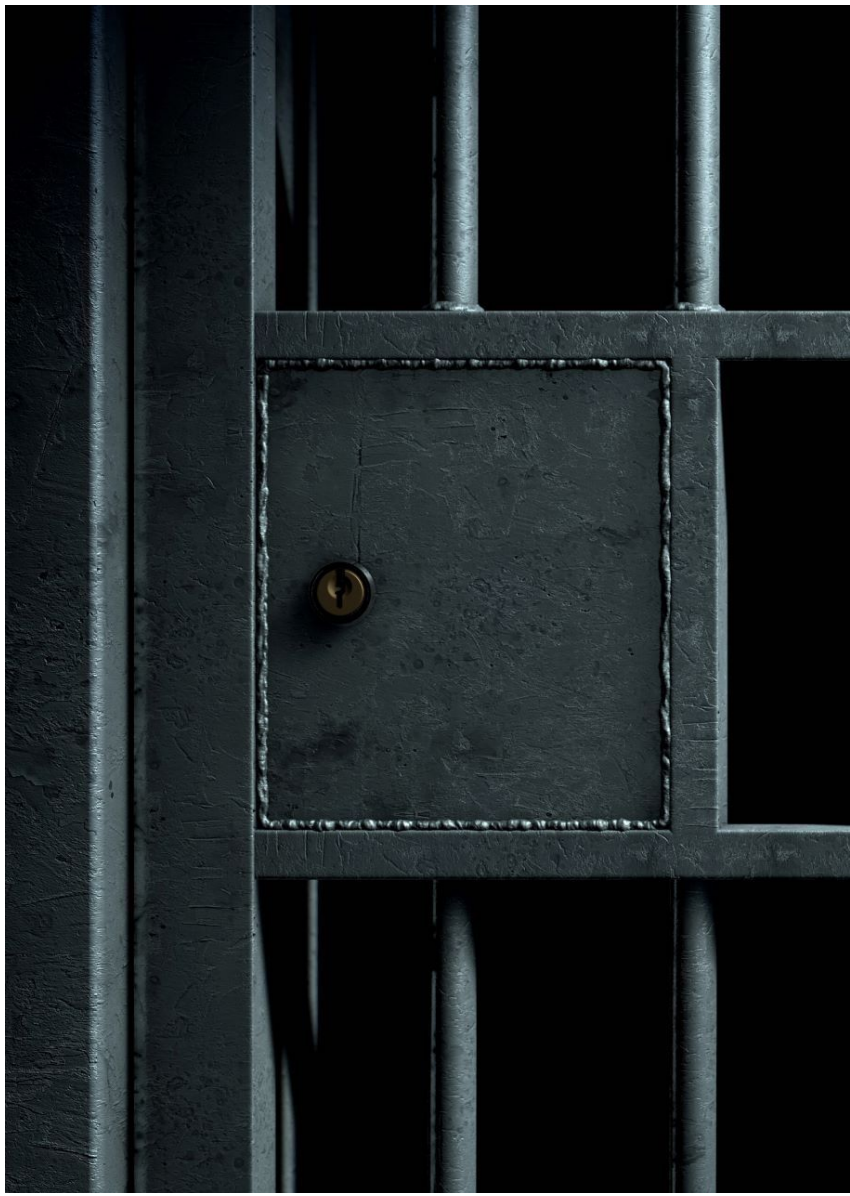
ProPublica's finding

> "The formula was particularly likely to **falsely flag black defendants as future criminals**, wrongly labeling them this way at almost twice the rate as white defendants."

Why would this be unfair?

Disparate Risk of Error

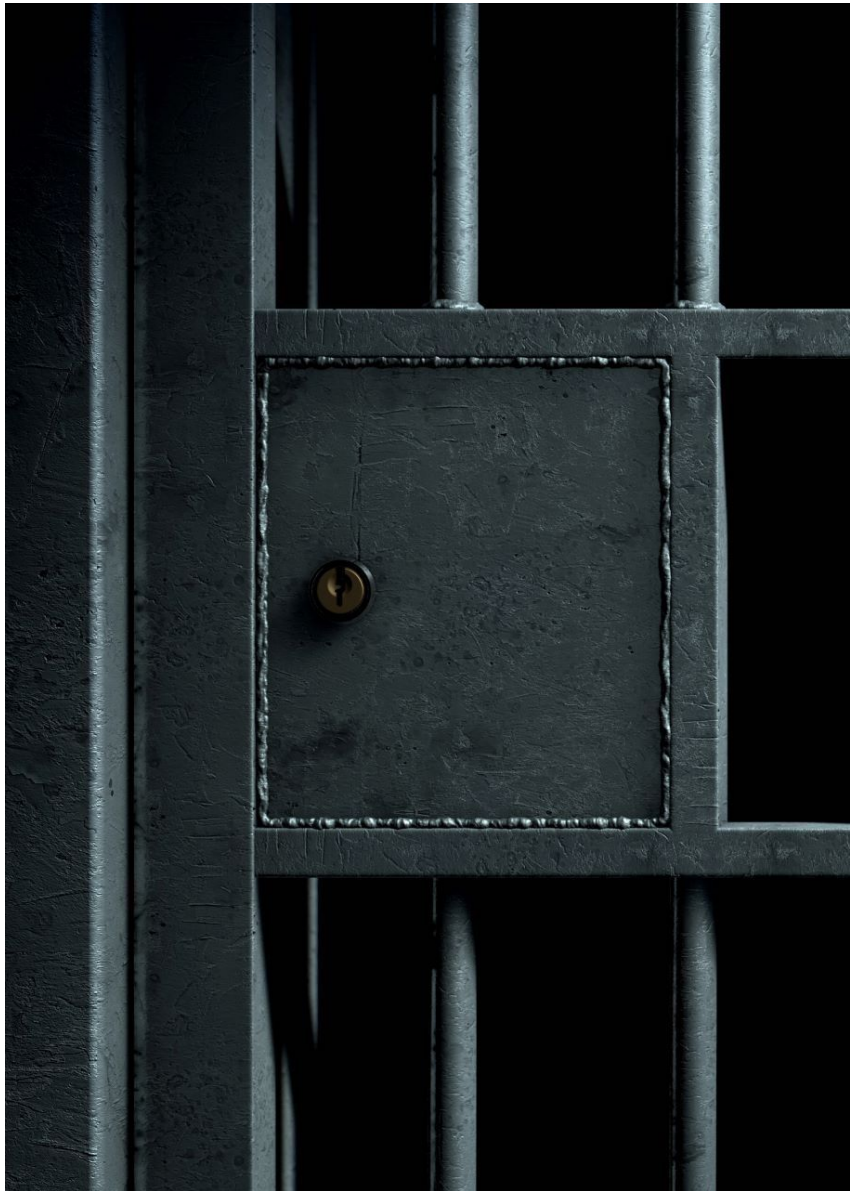
Exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants



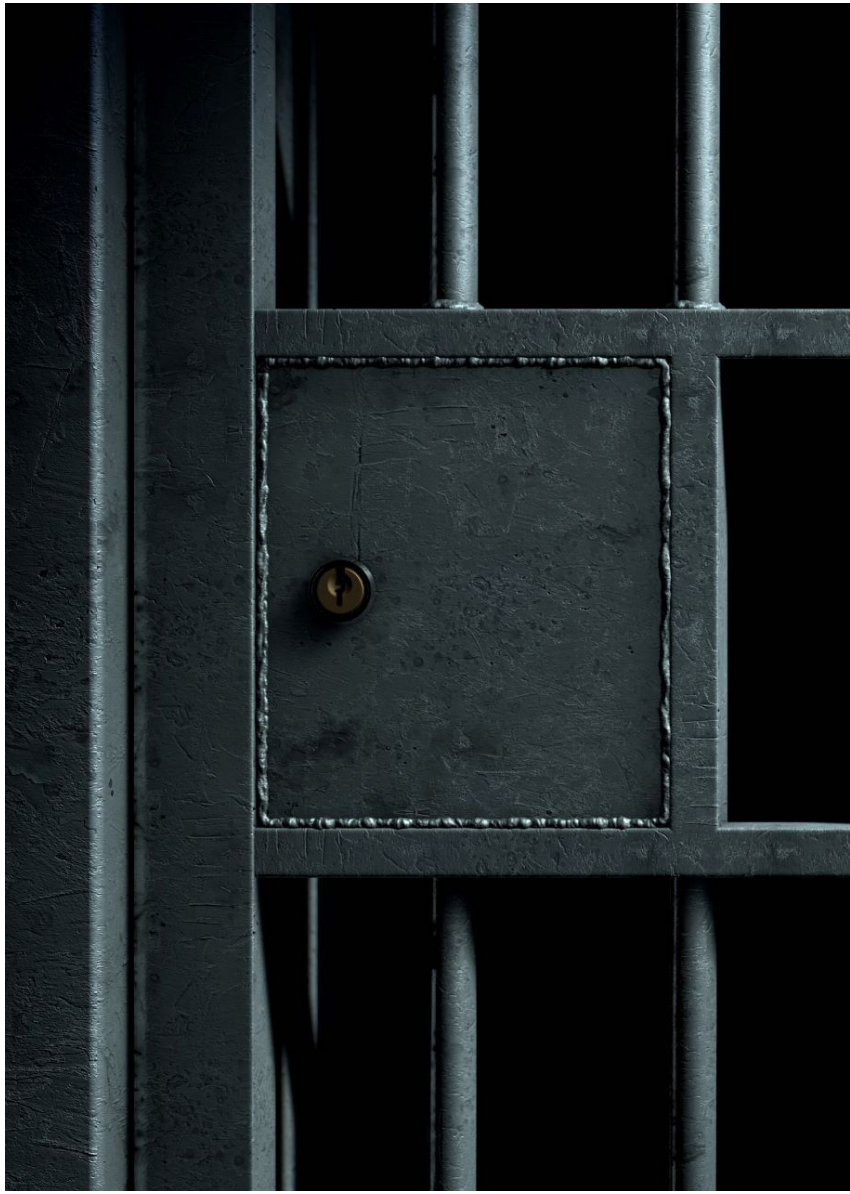
The Argument from Disparate Risk

1. If a recidivism prediction method has a higher false positive rate for Black defendants than White defendants, then that recidivism prediction method will be more likely to misclassify Black defendants than White defendants as high risk.
2. If a recidivism prediction method is more likely to misclassify Black defendants than White defendants as high risk, then using it to make pretrial detention decisions will expose Black defendants to a greater risk of mistaken detention than White defendants.
3. **Disparate Risk of Error.** Exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants.
4. Therefore, using a recidivism prediction method that has a higher false positive rate for Black defendants than White defendants to make pretrial detention decisions is unfair to Black defendants.

A defendant is **mistakenly detained** just in case they are objectively low risk, but judged to be high risk, and detained on that basis



A defendant is **mistakenly detained** just in case they are objectively low risk, but judged to be high risk, and detained on that basis



Evidentiary injustice

Form of **procedural injustice** faced by members of **marginalized groups**

- Social marginalization generates evidence that individuals lack features that society rewards
- Relying on that evidence disadvantages marginalized individuals who have the relevant features
- Phenomenon is pervasive, exacerbates preexisting disdvantage

Evidentiary injustice

Form of **procedural injustice** faced by members of **marginalized groups**

- Social marginalization generates evidence that individuals lack features that society rewards
- Relying on that evidence disadvantages marginalized individuals who have the relevant features
- Phenomenon is pervasive, exacerbates preexisting disdvantage

Evidentiary injustice

Form of **procedural injustice** faced by members of **marginalized groups**

- Social marginalization generates evidence that individuals lack features that society rewards
- Relying on that evidence disadvantages marginalized individuals who have the relevant features
- Phenomenon is pervasive, exacerbates preexisting disdvantage

Evidentiary injustice

A form of procedural injustice that occurs when members of a marginalized group that are **qualified** to receive favorable treatment are at an **unfair disadvantage** because their marginalized status makes them less likely to be recognized as qualified

- **qualified** = actually possess the features used to justify favorable treatment

Evidentiary injustice

In the context of pretrial detention:

- Defendants **qualify** for more favorable treatment (release) just in case they are **objectively low risk**
- So, evidentiary injustice occurs when **objectively low-risk** defendants from marginalized groups are more likely to be **misclassified as high risk** than others

The Argument from Disparate Risk

1. If a recidivism prediction method has a higher false positive rate for Black defendants than White defendants, then that recidivism prediction method will be more likely to misclassify Black defendants than White defendants as high risk.
2. If a recidivism prediction method is more likely to misclassify Black defendants than White defendants as high risk, then using it to make pretrial detention decisions will expose Black defendants to a greater risk of mistaken detention than White defendants.
3. **Disparate Risk of Error. Exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants.**
4. Therefore, using a recidivism prediction method that has a higher false positive rate for Black defendants than White defendants to make pretrial detention decisions is unfair to Black defendants.

Overview

1. ProPublica's argument
2. **ProPublica's argument fails**
3. COMPAS is unfairly biased anyway

The Argument from Disparate Risk

1. **If a recidivism prediction method has a higher false positive rate for Black defendants than White defendants, then that recidivism prediction method will be more likely to misclassify Black defendants than White defendants as high risk.**
2. If a recidivism prediction method is more likely to misclassify Black defendants than White defendants as high risk, then using it to make pretrial detention decisions will expose Black defendants to a greater risk of mistaken detention than White defendants.
3. Disparate Risk of Error. Exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants.
4. Therefore, using a recidivism prediction method that has a higher false positive rate for Black defendants than White defendants to make pretrial detention decisions is unfair to Black defendants.

ProPublica's analysis

1. COMPAS's **false positive rate** is higher for Black defendants than white defendants

2. Therefore, Black defendants are more likely to be **misclassified as high risk** by COMPAS than white defendants

ProPublica's analysis

1. COMPAS's **false positive rate** is higher for Black defendants than white defendants

2. Therefore, Black defendants are more likely to be **misclassified as high risk** by COMPAS than white defendants

Assumption

If (1) is true, **then** (2) must also be true

False positives and misclassification

Misclassified as high risk = classified as high risk, actually low risk

False positive = classified as high risk, no recidivism

False positives and misclassification

Misclassified as high risk = low risk, misclassified as high risk

False positive = no recidivism, classified as high risk

false positive ≠ misclassified as high risk

ProPublica's analysis

1. COMPAS's **false positive rate** is higher for Black defendants than white defendants

2. Therefore, Black defendants are more likely to be **misclassified as high risk** by COMPAS than white defendants

Assumption

If (1) is true, **then** (2) must also be true

False positive rate

The probability that a randomly selected defendant who does not reoffend within two years will be labeled "high risk"

$$FPR = \frac{\text{false positives}}{\text{actual negatives}} = \frac{FP}{FP + TN}$$

FP = false positive = labeled low risk, no recidivism

TN = true negative = labeled low risk, no recidivism

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H
WH	100	10
BH	10	100

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H
WH	100	10
BH	10	100

$$FPR = \frac{FP}{FP + TN}$$

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H	FP	TN
WH	100	10	10	9
BH	10	100	1	90

$$FPR = \frac{FP}{FP + TN}$$

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H	FP	TN	FPR
WH	100	10	10	9	53%
BH	10	100	1	90	1%

$$FPR = \frac{FP}{FP + TN}$$

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H	FP	TN	FPR
WH	100	10	10	9	53%
BH	10	100	1	90	1%

$$FPR = \frac{FP}{FP + TN}$$

ProPublica's analysis

1. COMPAS's **false positive rate** is higher for Black defendants than white defendants

2. Therefore, Black defendants are more likely to be **misclassified as high risk** by COMPAS than white defendants

Assumption

If (1) is true, **then** (2) must also be true

The Argument from Disparate Risk

1. **If a recidivism prediction method has a higher false positive rate for Black defendants than White defendants, then that recidivism prediction method will be more likely to misclassify Black defendants than White defendants as high risk.**
2. If a recidivism prediction method is more likely to misclassify Black defendants than White defendants as high risk, then using it to make pretrial detention decisions will expose Black defendants to a greater risk of mistaken detention than White defendants.
3. Disparate Risk of Error. Exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants.
4. Therefore, using a recidivism prediction method that has a higher false positive rate for Black defendants than White defendants to make pretrial detention decisions is unfair to Black defendants.

ProPublica's mistake

ProPublica conflated two senses in which a defendant can be a "false positive":

1. **False positive$_1$** = labeled high risk, but **does not reoffend**

2. **False positive$_2$** = labeled high risk, but **objectively low risk**

Overview

1. ProPublica's argument
2. ProPublica's argument fails
3. **COMPAS is unfairly biased anyway**

Practitioners Guide to COMPAS

August 17, 2012

NORTHPOINTE

Features used by COMPAS

- Age at first arrest
- Prior arrest history
- Residential status
- Employment status
- Employment history
- Substance abuse
- Criminal associates
- Failure to complete high school
- Lack of job skills
- Access to only minimum wage jobs

4.2.19 Vocation/Education

Another of the "big five" risk factors for crime and recidivism prediction in the Gendreau et al. (1996) meta-analysis is labeled "social achievement." This concept is an amalgam of educational attainment, vocational skills, job opportunities, a record of stable employment, good income, and, more generally, the level of legitimate economic opportunity. Basically, persons with more social capital have higher "life chances" than other persons who may have very restricted success opportunities (Hagan, 1998; Coleman, 1990).

Evidentiary injustice and COMPAS

A form of procedural injustice that occurs when members of a marginalized group that are **qualified** to receive favorable treatment are at an **unfair disadvantage** because their marginalized status makes it harder for them to demonstrate that they are qualified

Evidentiary injustice and COMPAS

A form of procedural injustice that occurs when members of a marginalized group that are **qualified** to receive favorable treatment are at an **unfair disadvantage** because their marginalized status makes it harder for them to demonstrate that they are qualified

> In the context of Violent Recidivism Risk, if you are young, unemployed and have an early age-at-first-arrest and a history of supervision failure, you could score medium or high on the Violence Risk Scale even though you never had a violent offense arrest.