```
</body>
```

background-image: url(images/linearregression.png)

Evidentiary Injustice and False Positive Rates

David Gray Grant

.smaller[*University of Florida*

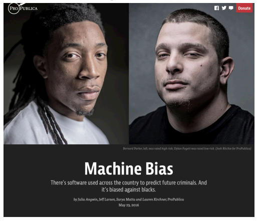
Jain Family Institute]

???

Thanks for having me here today, I'm excited to be meeting you all and looking forward to hearing your feedback on the project

---

class: img-centered, inverse



ProPublica

Donate

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

???

In May of 2016, ProPublica published an expose arguing that COMPAS, an algorithm that is used widely in the US criminal justice system, is unfairly biased against Black Americans.

---

class: img-right

COMPAS

Recidivism prediction instrument

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

???

COMPAS is a recidivism prediction instrument developed by Northpointe Inc.

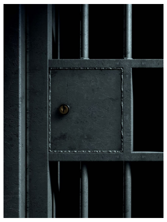
- The purpose of the algorithm is to estimate the likelihood that a person will commit a crime within the next two years
- The algorithm takes as input an intake survey completed by the defendant, along with the defendant's criminal record, and outputs a risk score between 1 and 10 representing the defendant's risk of recidivism
- Several states rely on COMPAS scores to make decisions about pretrial detention, parole, and sentencing.

The use case for COMPAS that I'll be focusing on is pretrial detention

---

class: img-right-full

Pretrial detention



The justification for pretrial detention is **preventative**

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

???

In pretrial detention, a defendant who has been charged with a crime -- but not yet convicted for that crime -- is imprisoned during the period leading up to their trial

The justification for pretrial detention is preventative

- The idea is that "high risk" defendants are likely to commit additional crimes against others if released, and so imprisoning them in the leadup to trial can be justified by appeal to the right that others have to

be protected from crime
- However, some defendants are unlikely to commit crimes if released, which means that detaining them pretrial isn't necessary for the protection of others, and so they should be released

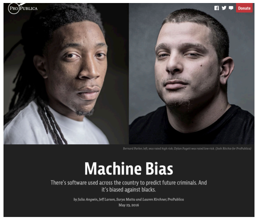
Given this justification for pretrial detention courts need to have some way to distinguish high risk defendants from low risk defendants, so that they can detain the high risk defendants and release the low risk ones

- That's where recidivism prediction instruments like COMPAS come in. They're supposed to help the state do a better job at distinguishing the high risk defendants from the low risk ones

Side note here: pretrial detention is also sometimes justified by appeal to the risk that the defendant will **fail to reappear** in court for their trial. I'm setting that justification aside for simplicity.

---

class: img-centered, inverse



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 argued that COMPAS scores are unfairly biased against Black defendants on the basis of a statistical analysis of the algorithm's performance.

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.".red[*]

.footer[.red[*] Angwin et al. (2016), "Machine Bias"]

???

Here's how they summarized their results.

> "[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."

In more technical terms, what ProPublica found was that the false positive rate for Black defendants was much higher than the false positive rate for white defendants

---

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"

???

A defendant is a **true positive** just in case they are labeled high risk, and they do go on to reoffend within two years, as predicted

A defendant is a **false positive** in the relevant sense if they are labeled high risk, but do not go on to reoffend within two years

The **false positive rate** is the probability that a randomly selected defendant that does not go on to reoffend—an **actual negative**—will be labeled "high risk," and so turn out to be a false positive

- If this is going by a little fast, don't worry, we'll come back to it

---

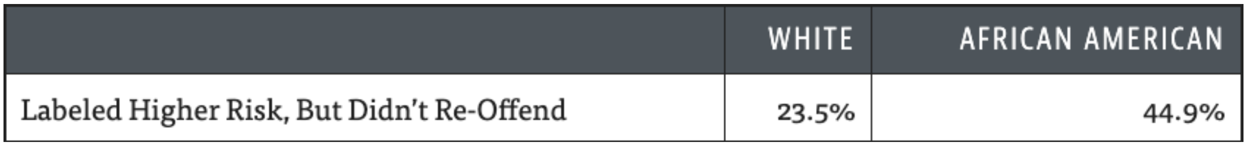
False positive rates

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"

.center[



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

]

???

ProPublica's researchers found that Black defendants were nearly twice as likely to be false positives as white defendants, and took this to show that COMPAS is unfairly biased against Black defendants.

- In making this argument, ProPublica is often interpreted as appealing to a statistical criterion of fairness called "False Positive Rate Equality"

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

???

False Positive Rate Equality is satisfied when Black defendants and White defendants have similar false positive rates

ProPublica claimed that because COMPAS fails to satisfy False Positive Rate Equality, using it to make pretrial detention decisions is unfair to Black defendants

- ProPublica did not offer an explicit argument for this claim; instead they seem to have thought it was obviously true.
- And indeed researchers in a variety of disciplines have agreed with ProPublica's reasoning and concluded that COMPAS is unfairly biased against Black 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

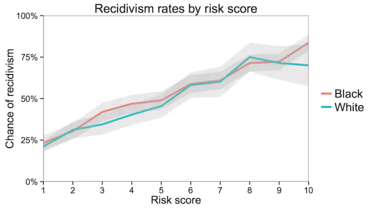
???

In response, the company that developed COMPAS -- Northpointe -- argued that COMPAS is *not* biased against Black defendants because it has a different statistical property, Equal Calibration

- Equal Calibration requires that defendants who receive the same risk score reoffend at similar rates, regardless of race
- For example, if Black defendants who receive a COMPAS score of 7 reoffend about 60% of the time, then white defendants who receive a COMPAS score of 7 should also reoffend about 60% of the time.
- Analyses that various groups have conducted do seem to back up Northpointe's claim that COMPAS satisfies equal calibration, at least as measured using rates of being charged with a new crime as a proxy for recidivism

---

class: center



???

Here's a chart from an independent analysis conducted by some researchers in the fair machine learning community, Corbett-Davies et al.

- As you can see, the chart seems to support the conclusion that COMPAS satisfies equal calibration
- Within a given COMPAS score, Black and white defendants were charged with new crimes at similar rates

Ok, so how is this supposed to show that COMPAS is not unfairly biased against Black defendants?

> "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.".red[*]

.footer[.red[*] Corbett-Davies et al. (2016)]

???
This is from the Corbett-Davies article

> "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."

So, since COMPAS scores represent the same level of recidivism risk for Black and white defendants, COMPAS scores treat relevantly similar Black and white defendants in similarly favorable ways, and so are not unfairly biased on the basis of race

---

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

???

A natural way of understanding the debate between ProPublica and Northpointe is to see them as endorsing different "requirements of fairness in recidivism prediction."

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

To say that something is "pro tanto" unfair is to say that it is unfair in some respect.

- For example, suppose you promise one employee a promotion, but then you get hit on the head and completely forget, and promise a different employee the same job. When it comes time to promote someone, there might be no way for you to be perfectly fair to both employees, since you made conflicting promises.
- As a result, your decision will be unfair to some extent no matter what you do, which is just to say that it will inevitably be at least pro tanto unfair

---

The COMPAS debate

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

???

So we can understand ProPublica as arguing that False Positive Rate Equality is a requirement of fairness, and that COMPAS is unfair because it violates it

- And we can see Northpointe as arguing that Equal Calibration is a requirement of fairness, but False Positive Rate Equality is not, and so that COMPAS is not unfairly biased against Black defendants

--

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

???

Other researchers subsequently demonstrated that if Black defendants reoffend at higher rates than white defendants, as does seem to be the case for at least some kinds of crimes, then it's impossible for any recidivism prediction instrument to satisfy both False Positive Rate Equality and Equal Calibration

- This led some people to claim that fairness in recidivism prediction is impossible under current conditions
- In the ensuing debate, it's generally conceded that Equal Calibration is a requirement of fairness, but there's disagreement about whether False Positive Rate Equality is

---

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!

.overlay[

]

???

I think that none of this is true

- False Positive Rate Equality is not a requirement of fairness, but neither is Equal Calibration
- So the conflict between False Positive Rate Equality and Equal Calibration does not show that fairness in recidivism prediction is impossible

Today I'll be talking about the first claim, that False Positive Rate Equality is a requirement of fairness

class: img-right-full

Overview

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

???

Here's an overview of the talk

1. First, I'll reconstruct ProPublica's argument that COMPAS is unfairly biased against Black defendants
2. Second, I'll argue that that argument fails, because it rests on a misinterpretation of their statistical analysis
3. Third, I'll argue that COMPAS is unfairly biased against Black defendants after all, just not for the reason ProPublica cited

I'll start by reconstructing ProPublica's argument

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?

???

Here's ProPublica's finding again

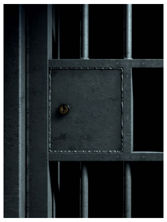
- “The formula was particularly likely to **falsely flag black defendants as future criminals**, wrongly labeling them this way at almost twice the rate as white defendants."
- ProPublica's claim here is that COMPAS is more likely to misclassify Black defendants as high risk than white defendants

- Why would that be unfair?

We can call the principle that seems to be in the background here disparate risk of error

class: img-right-full

Disparate Risk of Error



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

???

Disparate Risk of Error says that exposing Black defendants to a greater risk of mistaken detention than White defendants is unfair to Black defendants.

- If COMPAS misclassifies a low-risk defendant as high risk, then that makes it very likely that the judge will conclude that they are indeed high risk and detain them pretrial
- So if COMPAS is more likely to misclassify Black defendants as high risk, using COMPAS to make pretrial detention decisions will expose Black defendants to a greater risk of being mistakenly detained than white defendants
- Disparate Risk of Error says that that's unfair

This puts us in a position to reconstruct ProPublica's argument in the following way

class: smaller

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.

???
[Read out the argument]

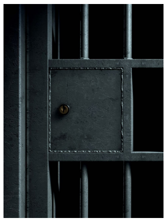
I call this the Argument from Disparate Risk

In the rest of this part of the talk, I'm going to explain why premise 3, Disparate Risk of Error, is plausible

- I'll start by saying more about what I mean by mistaken detention, and why it is morally significant

---

class: img-right-full



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

???

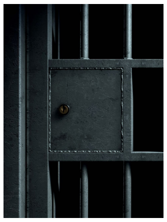
A defendant is **mistakenly detained** just in case they are objectively low risk, but judged to be high risk, and detained on that basis

So why am I calling this mistaken detention?

- As I mentioned, the justification for pretrial detention that's relevant here is a preventive one
- Even though the defendant hasn't been convicted of a crime, the thought is that detention can be justified in some cases if releasing the defendant would pose an unacceptably serious danger to others
- In cases of mistaken detention, the defendant is not objectively dangerous enough to others to justify detaining them on preventive grounds, so the objectively right thing to do -- the substantively fair thing to do -- is to release them
- That's why I'm calling mistaken detention a mistake -- it's a failure to treat the defendant in the way that we objectively ought to treat them

---

class: img-right-full



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

???

Now, even though it's substantively unfair, mistaken detention might nonetheless be procedurally fair

- Substantively unfair outcomes are not always procedurally unfair -- for example, it's possible for a procedurally fair criminal trial to nonetheless result in a mistaken conviction

However, I'm going to argue that whether a pretrial detention procedure treats defendants fairly depends on how it distributes the risk of mistaken detention across different groups of defendants

Specifically, I'll argue that exposing Black defendants to a greater risk of mistaken detention than white defendants is at least pro tanto unfair because it constitutes a form of procedural injustice I'll call evidentiary injustice

---

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

???

What I'm calling "evidentiary injustice" is a form of procedural injustice faced by members of marginalized groups

Social marginalization creates evidence that individuals in the marginalized group lack traits that society values and rewards, as well as evidence that they possess traits that society disvalues and penalizes

- In some cases, that evidence will be misleading, putting marginalized people who do in fact have the desired traits at a procedural disadvantage

---

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

???

I'll give two examples of evidentiary injustice

- First, loan applications. Social marginalization causes poverty, and indicators of poverty are also indicators that a loan applicant is likely to default. So even if a marginalized person is in fact at low risk of defaulting on a loan, the fact that they are marginalized means it's more likely that there will be misleading evidence that they are at high risk of default
    - And that in turn means that members of marginalized groups who in fact qualify for loans -- in the sense that they are objectively likely to repay them -- are less likely to be recognized as being qualified
- Second, job applications. Social marginalization deprives people of opportunities to demonstrate that they are "talented and motivated," such as access to prestigious schools, internship opportunities, tutoring assistance, and so on. And that in turn means thatqualified applicants who belong to marginalized groups will be less likely to be classified as qualified and hired

---

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

???

This phenomenon is pervasive, and that the upshot is that members of marginalized groups that **in fact** qualify for more favorable treatment by society's standards are less likely to receive that treatment

- So on top of the unfair disadvantages they already face in acquiring the traits that society values and rewards -- a failure of what is sometimes called **substantive opportunity** -- they are subjected to a pervasive form of **procedural** injustice that exacerbates those unfair disadvantages
- I'm calling this kind of procedural injustice "evidentiary injustice"

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

???

To recap, evidentiary injustice is a form of procedural injustice that occurs when qualified members of a marginalized group are at an **unfair disadvantage** because their marginalized status makes it harder for them to **demonstrate** that they are qualified -- that they have the features used to justify more favorable treatment in context

- This is a kind of procedural injustice, because the substantively fair outcome is for the qualified person to receive the relevant benefit, but the procedure used to allocate the benefit puts them at an unfair procedural disadvantage relative to others

That's what evidentiary injustice is, now I'll explain how it's relevant in context

We can appeal to the phenomenon of evidentiary injustice to explain why Disparate Risk seems plausible

---

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

???

In the context of pretrial detention, Defendants **qualify** for more favorable treatment (namely release) just in case they are at **objectively low risk** of recidivism

It follows that evidentiary injustice occurs in the context of pretrial detention when objectively low-risk defendants from marginalized groups, such as low-risk black defendants, are more likely to be misclassified as high risk than others

Since evidentiary injustice is a form of procedural injustice, it follows that exposing Black defendants to a greater risk of mistaken detention than White defendants is a form of procedural injustice against Black defendants, and so is unfair to Black defendants

This gives us an argument for the principle I'm calling Disparate Risk of Error

---

class: smaller

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.

???

The reason that exposing Black defendants to a higher risk of mistaken detention than white defendants is unfair is that it constitutes a form of evidentiary injustice against Black defendants

- And that is at least pro tanto unfair to Black defendants, even if it turns out that it's the morally best option on balance

---

class: img-right-full

Overview

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

???

That completes my reconstruction of ProPublica's argument that using COMPAS is unfair because it has a higher false positive rate for Black defendants than white defendants

Next, I'm going to argue that ProPublica's argument fails, because their analysis fails to show that the first premise of the argument is true

---

class: smaller

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.

???

So far, I've argued that we can make sense of ProPublica's concern by appealing to the principle of Disparate Risk, which says that exposing Black defendants to a greater risk of mistaken detention is unfair

- I've also argued that the principle of Disparate Risk is plausible, because exposing Black defendants to a greater risk of mistaken detention is a form of evidentiary injustice

I'm going to turn now to premise 1

- According to premise one, 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 as high risk than White defendants

I'm going to argue that this premise is false

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

???

What ProPublica actually found was that COMPAS' false positive rate was higher for Black defendants than white defendants

- What they took that to show was that COMPAS is more likely to incorrectly classify Black defendants as high risk than white defendants

In other words, ProPublica seems to have made the following assumption:

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

.center[**If** (1) is true, **then** (2) must also be true]

???

Necessarily, if COMPAS' false positive rate is higher for Black defendants than white defendants, then COMPAS is more likely to misclassify Black defendants as high risk

That sounds tautological -- it sounds like it's just a logical truth -- so it's easy to see why ProPublica would have assumed it

- But surprisingly, it turns out to be false
- It's not true in general that if a recidivism prediction method has a higher false positive rate for Black defendants, it will also be more likely to misclassify Black defendants as high risk

To see why this assumption is false, consider first what it means to say that a defendant is misclassified as high risk

---

False positives and misclassification

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

False positive = classified as high risk, no recidivism

???

A defendant is misclassified as high risk in the sense that is morally relevant just in case the defendant is actually at low risk of recidivism but is mistakenly judged to be at high risk of recidivism

- On the other hand, a defendant is a false positive just in case the defendant is classified as at high risk of recidivism, but the defendant does not in fact commit a crime within two years -- so the prediction that they are likely to recidivate does not come to fruition

Now, the first point to make here is that a defendant who is a false positive in this sense need not have been *misclassified* as high risk

---

False positives and misclassification

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

False positive = no recidivism, classified as high risk

.center[false positive ≠ misclassified as high risk]

???

If COMPAS classifies a defendant as at high risk of recidivism, the mere fact that they do not subsequently recidivate does not show that COMPAS's estimate was incorrect.

- After all, a defendant can be strongly disposed to commit crimes, but nonetheless fail to do so
- For instance, if we assume that a defendant has an 95% chance of recidivism, that means that they will fail to commit a new crime about 5% of the time
- The mere fact that they don't go on to recidivate doesn't in itself show that they weren't very likely to do so
- Things that are likely to happen fail to happen all the time.

This casts doubt on ProPublica's inference from "COMPAS's false positive rate for Black defendants is higher" to "COMPAS is more likely to misclassify Black defendants 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

.center[**If** (1) is true, **then** (2) must also be true]

???

Nonetheless, you might be tempted to think that this assumption is plausible anyway

- If the false positive rate is higher for Black defendants, *surely* that must mean that Black defendants are at least *more likely* to be misclassified as high risk

However, I'm going to argue that this line of reasoning is mistaken

---

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

???

As I mentioned earlier, the false positive rate for a group is the probability that a randomly selected member of the group who does not commit a new crime within two years will be labeled "high risk"

- It's calculated using the formula on the screen
- You take the total number of false positives for the group, and divide it by the total number of actual negatives
- False positives are defendants that are labeled high risk and don't reoffend
- True negatives are defendants that are labeled low risk and don't reoffend
- The actual negatives are all of the defendants in the group that did not reoffend, which includes both the false positives and the true negatives
- One thing to notice is that the false positive rate only depends on how the defendants who do not go on to reoffend are classified

With that in mind, here's an example that shows that false positive rates and rates of misclassification can come apart dramatically

---

class: small

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

???

Suppose that there are two guilds of assassins, the Professionals and the Hobbyists.

- Each assassin has made the following deal with the devil: the devil will use a random number generator between 0 and 100 to determine whether they kill within the next two years (by either inducing them to kill or preventing them from killing)

The terms of the deal, however, are different for Professionals and Hobbyists.

- If the assassin is a Professional, the deal is that the devil will induce them to kill with a 90% probability and prevent them from killing with a 10% probability
- For Hobbyists, these odds are reversed. The devil will ensure that there is a 10% probability that each Hobbyist will kill, and a 90% probability that they won't

This is just an evocative way to ensure that the objective probability that each Professional will kill is 90%, and the objective probability that each Hobbyist will kill is 10%.

---

class: small

background-image: url(images/spyvspy-plain.png)

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H
WH	100	10
BH	10	100

???

Suppose also that some assassins wear Black hats and some assassins wear white hats

- Among the white hats, 100 are Professionals and only 10 are Hobbyists
- Among the Black hats, those numbers are reversed; only 10 are Professionals and 100 are Hobbyists

Now suppose that you are a pretrial hearing judge, and all of these assassins are going to appear before you in court

- Fortunately, you have a perfectly reliable way to estimate recidivism risk -- God will helpfully tell you the objective probability that each assassin will kill within two years, based on whether they are a Professional or a Hobbyist
- If an assassin is 50% likely to kill or greater, you will classify them as at "high risk" of killing and detain them, and otherwise you will release them
- Obviously this is not a very realistic example, but bear with me

This gives us all the information we need to estimate the false positive rate of the recidivism prediction method that you're going to use, which is to classify all Professionals as high risk and all Hobbyists as low risk

---

class: small

background-image: url(images/spyvspy.png)

The Assassins

Professionals kill 90% of the time

Hobbyists kill 10% of the time

	P	H
WH	100	10
BH	10	100

???

Here's the formula for calculating the false positive rate

- So the first thing we need to do is to calculate the number of false positives for each group, and the number of true negatives for each group
- An assassin is a false positive if you predict that they are high risk, but they don't go on to kill
- An assassin is a true negative if you predict that they are low risk, and they don't kill
- So the false positive rate only depends on how we classify the assassins that don't go on to kill

---

class: small

background-image: url(images/spyvspy.png)

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

???

Let's take the white hats first

- How many false positives should we expect to see among the white hats? Well, only Professionals can be false positives, since only they get classified as high risk
- They kill 90% of the time, which means they don't kill 10% of the time, so 10% of them will be false positives, which works out to 10 expected false positives for the white hats
- What about true negatives? Only Hobbyists can be true negatives, since you will classify all of them as low risk. About 10% of them will end up killing, which means 90% of them won't kill, so we should expect to see 9 true negatives among the white hats

Now the black hats

- 10 Black hats are professionals, and 10% of them will be false positives, so we should expect to see about 1 false positive among the Black hats
- 100 Black hats are Hobbyists, and 90% of them will be true negatives, which gives us about 90 true negatives for the white hats

Now we just plug that into our formula, which gives us the following false positive rates

---

class: small

background-image: url(images/spyvspy.png)

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%

???
The expected false positive rate for the white hats, as you can see, is 10/19, or about 53%

- By contrast, the expected false positive rate for the Black hats is 1/91, or about 1%
- So the false positive rate for the white hats is more than 50 times the false positive rate for the black hats

---

class: small

background-image: url(images/spyvspy.png)

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%

???

Recall, though, that your method for estimating recidivism risk is perfectly accurate -- God just tells you each assassin's objective probability of killing -- which means that you never misclassify anyone who is actually low risk as high risk

- All of the Professionals are correctly classified as high risk, since they are indeed objectively very likely to kill
- All the Hobbyists are correctly classified as low risk, since they are objectively unlikely to kill -- relative to the Professionals, anyway

This example shows that using a method that has a higher false positive rate for one group than another need not expose the group with the higher false positive rate to a greater risk of being misclassified as high risk

- But that's just to say that the assumption ProPublica made in interpreting their results is false

---

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

.center[**If** (1) is true, **then** (2) must also be true]

???

We can't conclude from the fact that COMPAS has a higher false positive rate for Black defendants that COMPAS is more likely to misclassify Black defendants as high risk

- If Black defendants are more likely to reoffend than white defendants -- as many researchers believe is true in the case of the population studied by ProPublica -- then we would expect to see more false positives among the Black defendants
- And we should expect to see that even if we assume that COMPAS is just as likely to misclassify a white defendant as high risk as a Black defendant

This in turn means that ProPublica's analysis does not establish that the first premise of the argument from disparate risk is true

---

class: smaller

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.

???

Now, if I'm right, then something like this argument is what ultimately explains why False Positive Rate Equality seems like a requirement of fairness in recidivism prediction

- The thought is that if a recidivism prediction method has a higher false *postive* rate for Black defendants, then it will necessarily be more likely to misclassify Black defendants as high risk
- And that in turn will result in exposing Black defendants to a greater risk of mistaken detention, which is a form of evidentiary injustice

But as we've seen, that need not be the case

- If base rates of recidivism differ -- as they plausibly do for Black and White defendants -- then a method that has a higher false positive rate for Black defendants need not be more likely to misclassify Black defendants as high risk

In turn, this means that False Positive Rate Equality is not a requirement of fairness, because using a recidivism prediction method that violates it will not always be unfair

---

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**

???

What went wrong? ProPublica's mistake was to conflate two senses in which a defendant can be a "false positive"

- A defendant is a false positive in ProPublica's sense just in case they are labeled high risk, but do not go on to commit a crime within two years
- A defendant is a false positive in the sense that is **morally relevant** just in case they are labeled high risk, but are in fact at objectively low risk of recidivism

Initially, it seems like a defendant's risk of being a false positive in the first sense corresponds to their risk of being misclassified as high risk, and so their risk of being mistakenly detained

But that's not true

It's only a defendant's risk of being a false positive in the second sense that corresponds to their risk of mistaken detention

- Unfortunately, there is no straightforward way to measure that risk

class: img-right-full

Overview

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

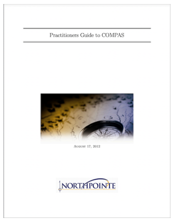
???

It might seem like I've validated Northpointe's side of the debate by showing that COMPAS is not unfairly biased against Black defendants

- However, that doesn't follow from anything I've said so far -- all I've shown is that COMPAS is not unfairly biased for the reasons identified by ProPublica
- In fact, developing ProPublica's argument more carefully helps us to see that there's a different reason to think that COMPAS is unfairly biased against Black defendants

---

class: fullbleed, center, inverse



Practitioners Guide to COMPAS

August 17, 2012

NORTHPOINTE

???

Someone at Northpointe helpfully posted their guide to how to use COMPAS for criminal justice professionals on their website

- It makes for pretty interesting reading class: col-2

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

???

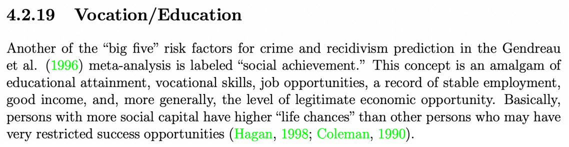
Here are some of the features that Northpointe says COMPAS uses to predict recidivism risk

- Notice anything interesting?
- Obviously, these are all markers of social marginalization -- the people that are most likely to possess the features that COMPAS treats as evidence of high recidivism risk are people who are socially marginalized in some way, whether that's due to poverty, or discrimination, or some other form of socioeconomic disadvantage

Interestingly, Northpointe seems to be well aware of this fact. Here's a passage from their guide to using COMPAS

---

class: center



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).

???

Another of the "big five" risk factors for crime and recidivism prediction 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

- Now, one upshot of this is that if we use COMPAS to make pretrial detention decisions, that will put objectively low risk defendants who are members of marginalized groups, such as Black defendants, in a worse position to demonstrate that they ought to be released, which is the substantively fair outcome

- But that means that using COMPAS to make pretrial detention decision puts members of marginalized groups that in fact qualify for more favorable treatment at an unfair disadvantage relative to non-marginalized defendants
- And that's just to say that we should expect basing pretrial detention decisions on COMPAS scores to lead to evidentiary injustice

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

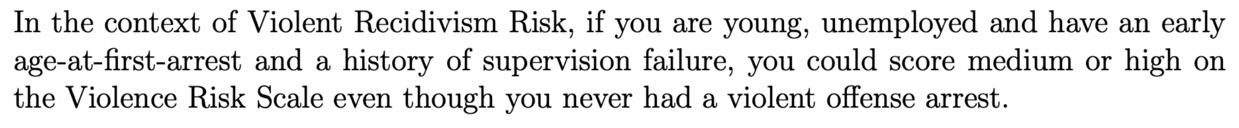
???

So even though I've argued that using COMPAS is not unfairly biased against Black defendants because it violates False Positive Rate Equality, unpacking the reasoning behind that argument helps us to see that using COMPAS is unfair to Black defendants for a different reason

Another quote from Northpointe's guide to using COMPAS puts the point quite vividly

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.

???

"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 Violent Risk Scale even though you never had a violent offense arrest."

- It follows from this that even if a defendant has never done anything to harm others, and even if they pose no danger to others, they can nonetheless be scored at high risk of recidivism—-largely because they have a collection of features that tend to be correlated with, and caused by, unjust forms of socioeconomic disadvantage
- And that seems like a clear reason to think that using COMPAS is at least pro tanto unfair to members

of socially marginalized groups
- Thanks