INFO 4390/5390 / CS 5382: Designing Fair Algorithms

Lecture 3: 2024-01-30

Pierson & Koenecke

Today’s lecture

- Review of confusion matrices
- COMPAS & fairness definitions
- Principle 1: Trade-offs between definitions
- Zooming out: broader questions about the COMPAS case study
- Principle 2: Be precise about what you mean by “bias”

Review from last time: confusion matrices

Model predicts...

True Value		Egg	Rock
	Egg	**True positive** Correct prediction	**False negative** **(kids' model predicts rock so it doesn't get thrown, but it's actually just an egg)**
	Rock	**False positive** **(kids' model predicts egg, so it'll get thrown, but it's actually a rock!)**	**True negative** Correct prediction

https://www.reddit.com/r/mildlyinteresting/comments/5uvn8r/i_collect_rocks_that_look_like_eggs/

We can use tp, fn, fp, and tn to describe other statistics!

	ŷ=1	ŷ=0
y=1	tp	fn
y=0	fp	tn

Accuracy = (tp + tn) / (tp + fn + fp + tn)

We can use tp, fn, fp, and tn to describe other statistics!

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Precision = tp / (tp + fp)

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Recall = tp / (tp + fn)

a.k.a. **True Positive Rate**

We can use tp, fn, fp, and tn to describe other statistics!

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Positive Rate = fp / (fp + tn)

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Recall = tp / (tp + fn)

a.k.a. **True Positive Rate**

What does it mean for a predictive algorithm to be fair?

Example: can you release a defendant?

- A defendant can be either:
    a. Kept in jail while they await trial
    b. Released prior to trial

- How does the judge decide?

Example: can you release a defendant?

- A defendant can be either:
  a. Kept in jail while they await trial
  b. Released prior to trial

- How does the judge decide? **Maybe by estimating how likely the defendant is to *recidivate***

  Recidivism: committing another crime

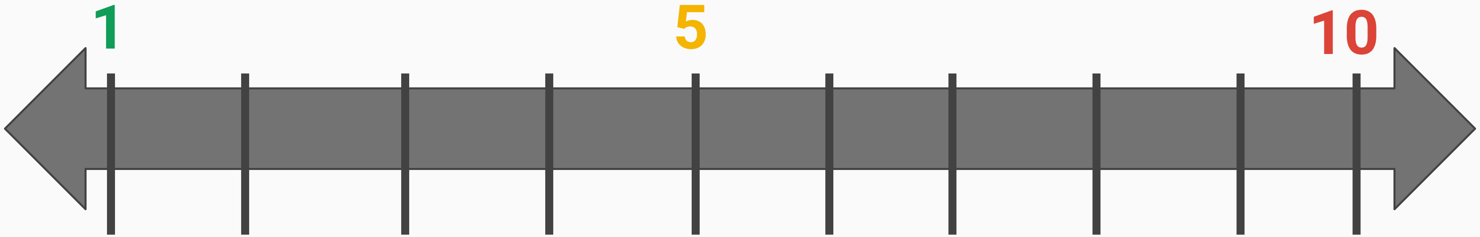
Example: can you release a defendant?

- This can be done by calculating a discrete “risk score”

[Risk of recidivism](#) (risk of committing another crime)

Example: can you release a defendant?

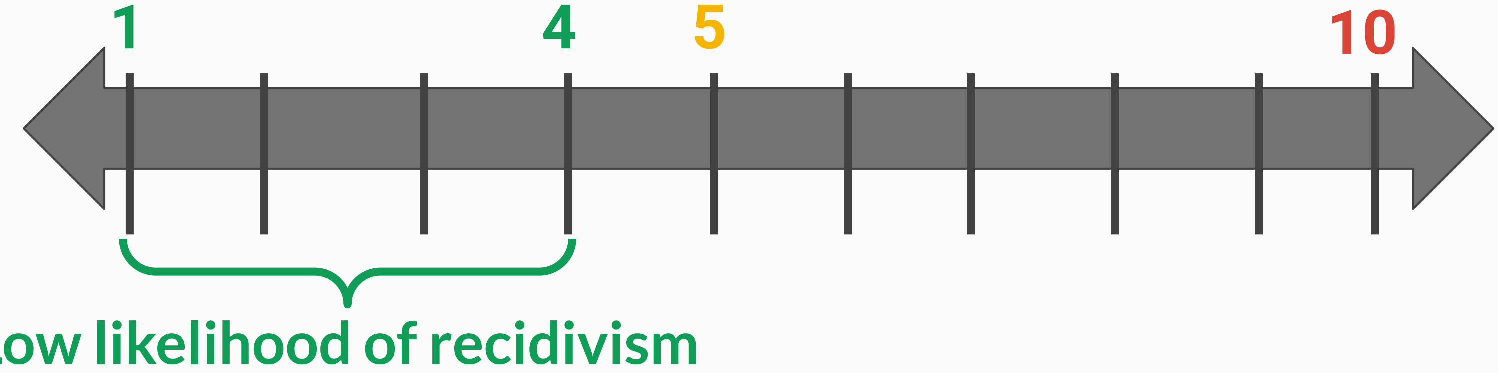
- This can be done by calculating a discrete “risk score”



[Risk of recidivism](#) (risk of committing another crime)

Example: can you release a defendant?

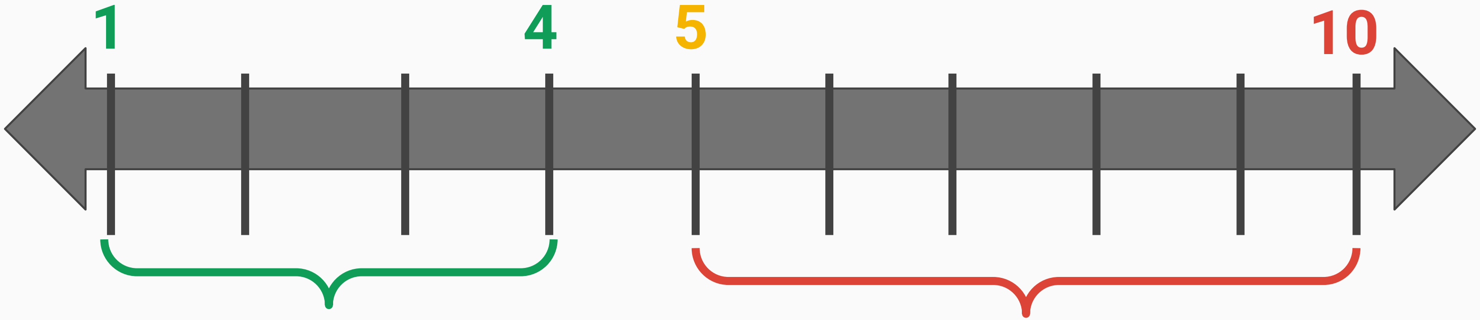
- This can be done by calculating a discrete "risk score"



Risk of recidivism (risk of committing another crime)

Example: can you release a defendant?

- This can be done by calculating a discrete “risk score”

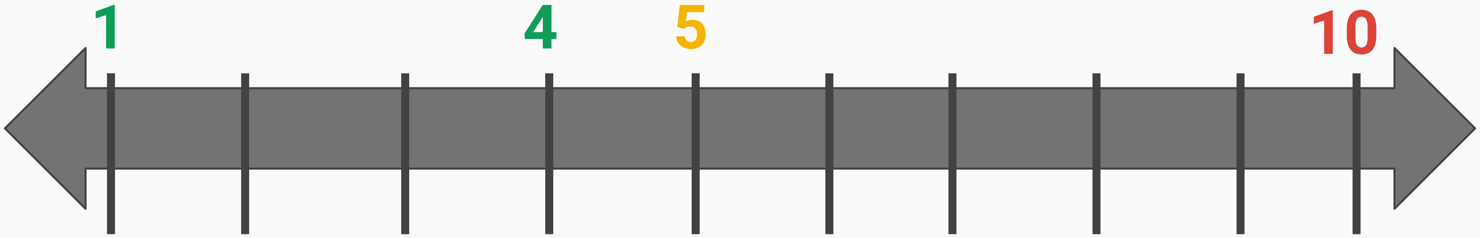


Medium/High likelihood of recidivism

Risk of recidivism (risk of committing another crime)

Example: can you release a defendant?

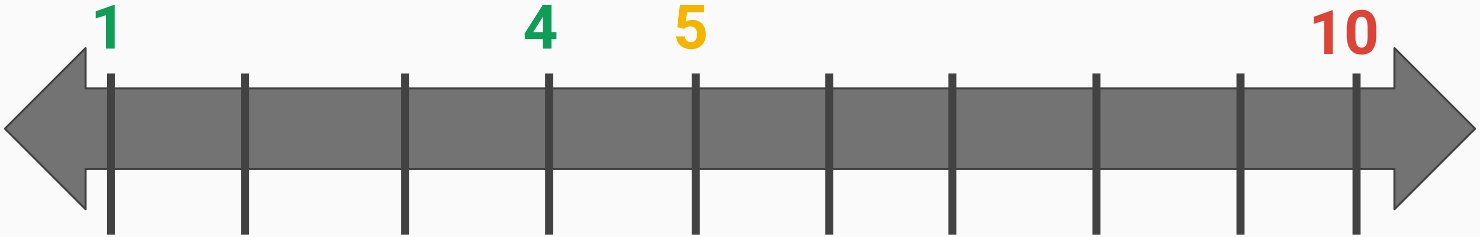
- <u>**How do we calculate a defendant's risk score?**</u>



- Any ideas?

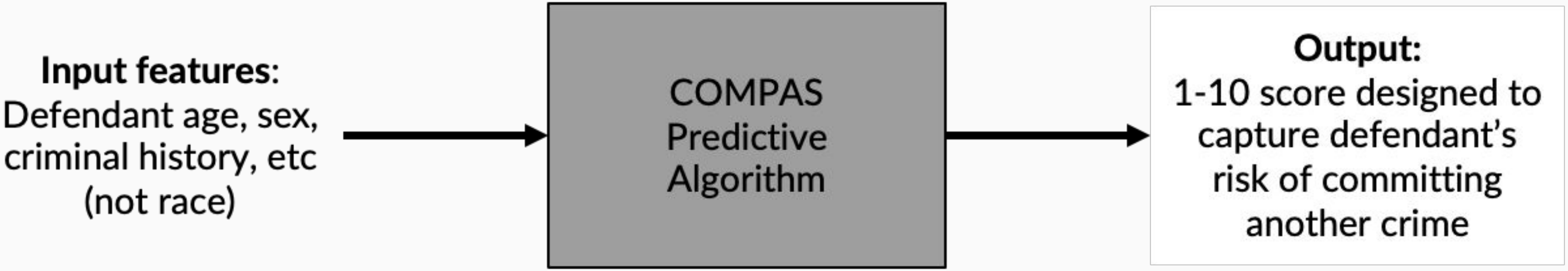
Example: can you release a defendant?

- <u>**How do we calculate a defendant's risk score?**</u>

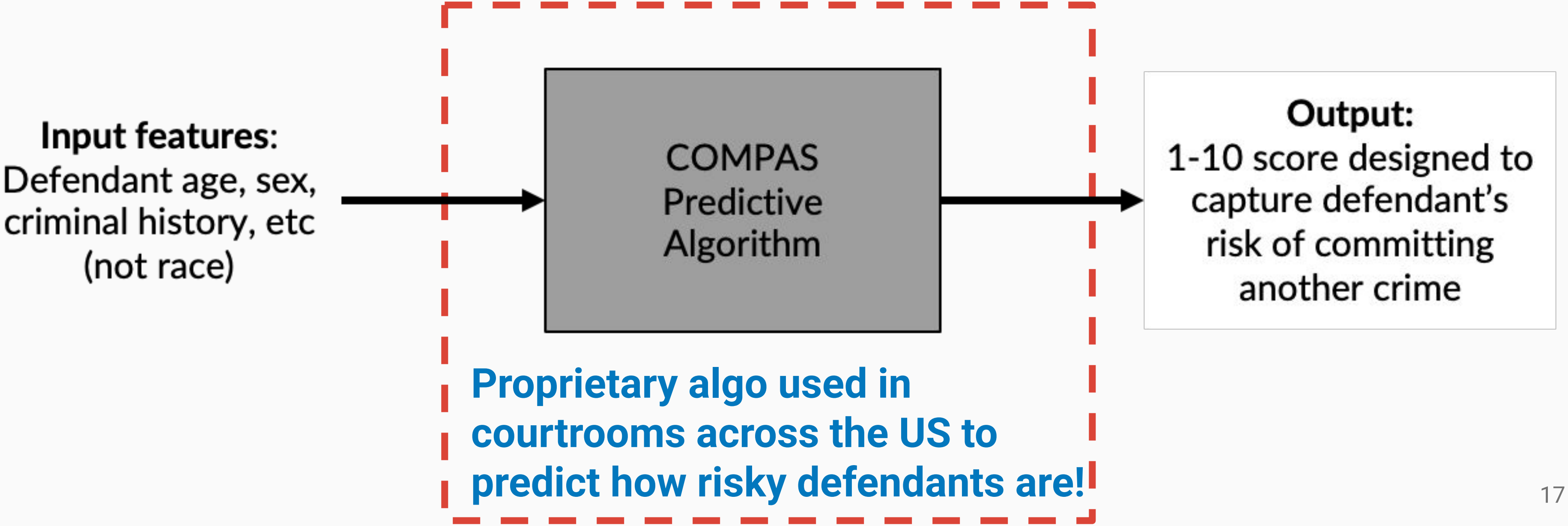


- **Variables like defendant's age, sex, criminal history, ...**

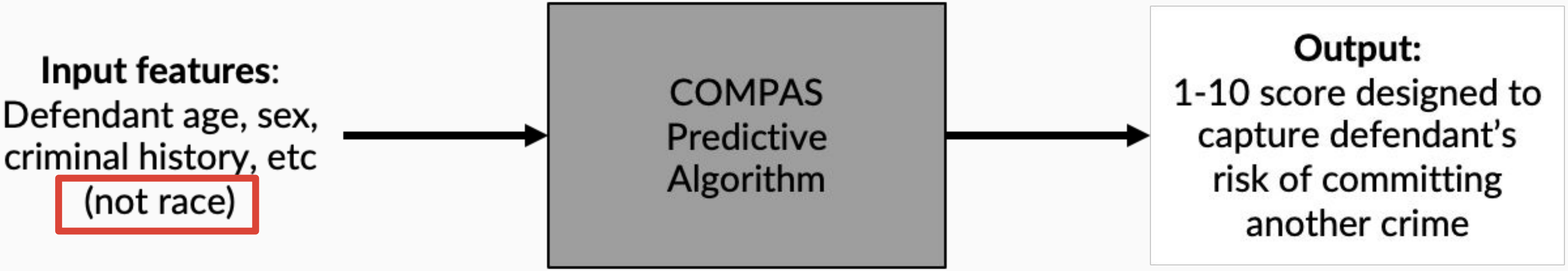
The COMPAS algorithm



The COMPAS algorithm



The COMPAS algorithm



Source: Angwin, Larson, Mattu, and Kirchner. “Machine Bias”. *ProPublica*, 2016.

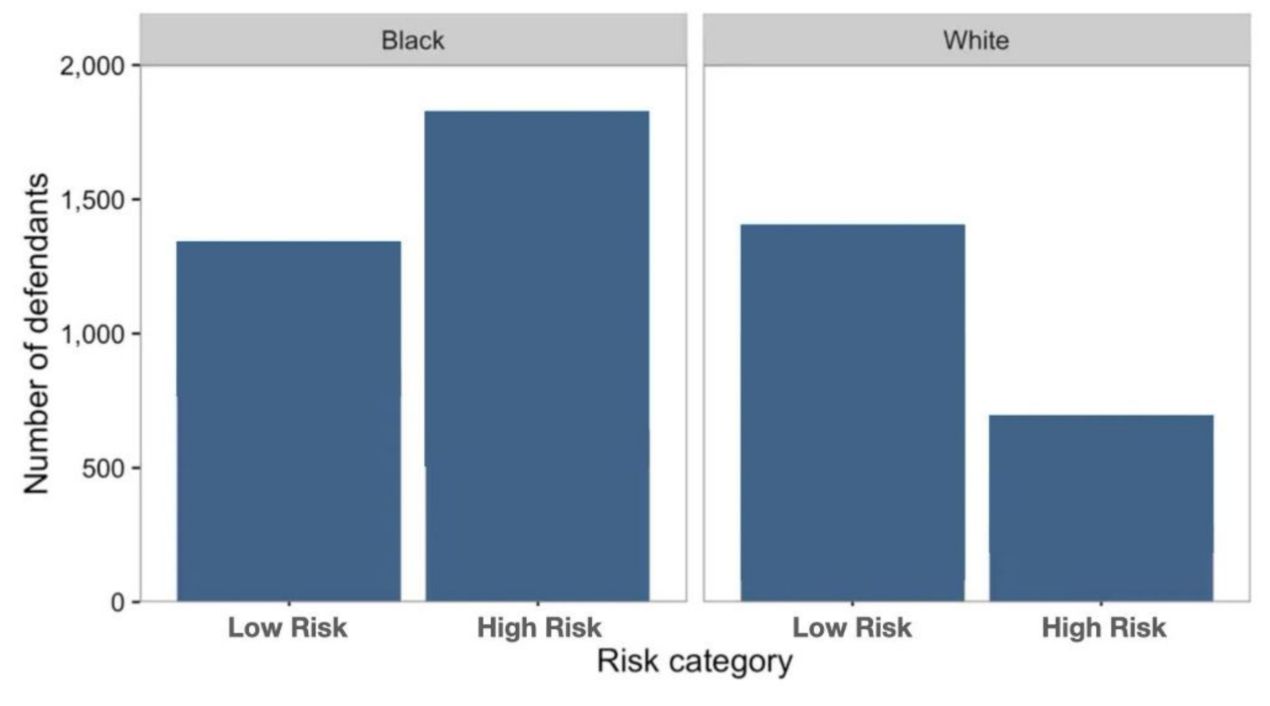
Is COMPAS fair?

ProPublica: No

Observation	**Fairness Principle**
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity

Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

Is COMPAS fair?



Is COMPAS fair?

ProPublica: No

Observation	**Fairness Principle**
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity

If race isn't an input, how come COMPAS can still results in this?

Is COMPAS fair?

ProPublica: No

Observation	**Fairness Principle**
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity

If race isn't an input, how come COMPAS can still results in this?

The algorithm inputs (e.g., prior arrests) are likely *correlated* with race!

In general, how <u>important</u> do you think it is for (any) algorithm to satisfy *statistical parity* (classifying equal fractions of each group as high risk)? ✋

Not very

Somewhat

Very

A related fairness definition: *demographic balance*

	ŷ=1	ŷ=0
y=1	tp	fn
y=0	fp	tn

Predicted Positive =

(tp + fp) / (tp + fn + fp + tn)

- Statistical parity: the % predicted positive (ŷ=1) is equal for **all groups**
  - $PP_{Black} = PP_{White}$

A related fairness definition: *demographic balance*

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Predicted Positive =

(tp + fp) / (tp + fn + fp + tn)

- Statistical parity: the % predicted positive ($\hat{y}=1$) is equal for **all groups**
  - $PP_{Black} = PP_{White}$
- Demographic balance: the % predicted positive ($\hat{y}=1$) for each group reflects its share **in the real world**
  - PP_{Black} = recidivism rate for Black defendants
  - PP_{White} = recidivism rate for White defendants

A related fairness definition: *demographic balance*

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Predicted Positive =

(tp + fp) / (tp + fn + fp + tn)

- Example: **predicting breast cancer**
  - There are sex differences: women's rates should be higher

A related fairness definition: *demographic balance*

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

Predicted Positive =
(tp + fp) / (tp + fn + fp + tn)

- Example: **predicting breast cancer**
  - There are sex differences: women's rates should be higher
- Demographic balance: the % predicted positive ($\hat{y}$=1) for each group reflects its share **<u>in the real world</u>**
  - PP_{Women} = % women with breast cancer
  - PP_{Men} = % men with breast cancer
  - $PP_{Women} > PP_{Men}$ doesn't satisfy statistical parity

Is COMPAS fair?

ProPublica: No

Observation	Fairness Principle
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity
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**

“The algorithm's false positive rates and false negative rates are not equal across races”

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Positive Rate =

fp / (fp + tn)

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Negative Rate =

tp / (tp + fn) = 1 - TPR

“The algorithm's false positive rates and false negative rates are not equal across races”

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Positive Rate =

fp / (fp + tn)

How likely someone who does *not* reoffend is to be falsely classified as high risk

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Negative Rate =

tp / (tp + fn) = 1 - TPR

“The algorithm's false positive rates and false negative rates are not equal across races”

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Positive Rate =

fp / (fp + tn)

How likely someone who does *not* reoffend is to be falsely classified as high risk

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	tp	fn
$y=0$	fp	tn

False Negative Rate =

tp / (tp + fn) = 1 - TPR

How likely someone who *does* reoffend is to be falsely classified as low risk

“The algorithm's false positive rates and false negative rates are not equal across races”

	ŷ=1	ŷ=0
y=1	tp	fn
y=0	fp	tn

False Positive Rate =

fp / (fp + tn)

	ŷ=1	ŷ=0
y=1	tp	fn
y=0	fp	tn

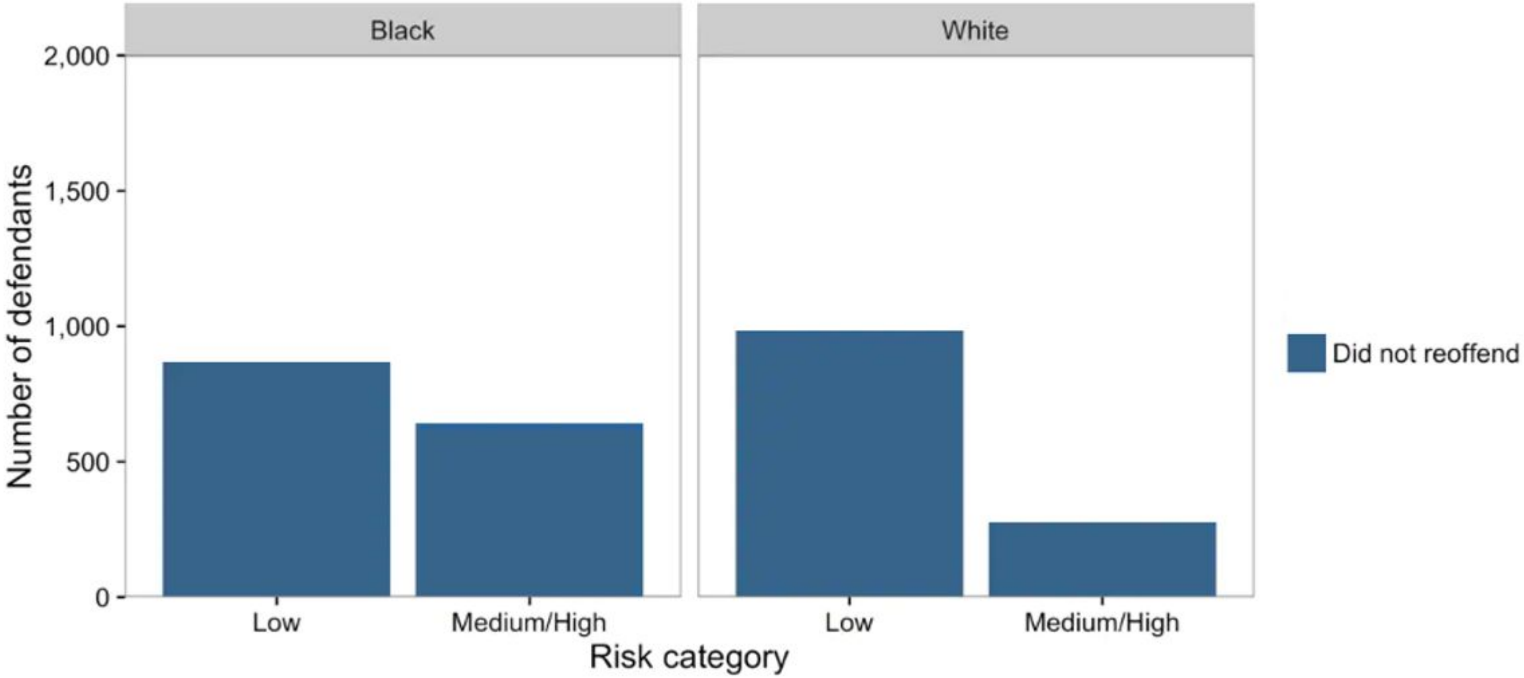
False Negative Rate =

tp / (tp + fn) = 1 - TPR

Predictive equality: $FPR_{Black} = FPR_{White}$ and $FNR_{Black} = FNR_{White}$

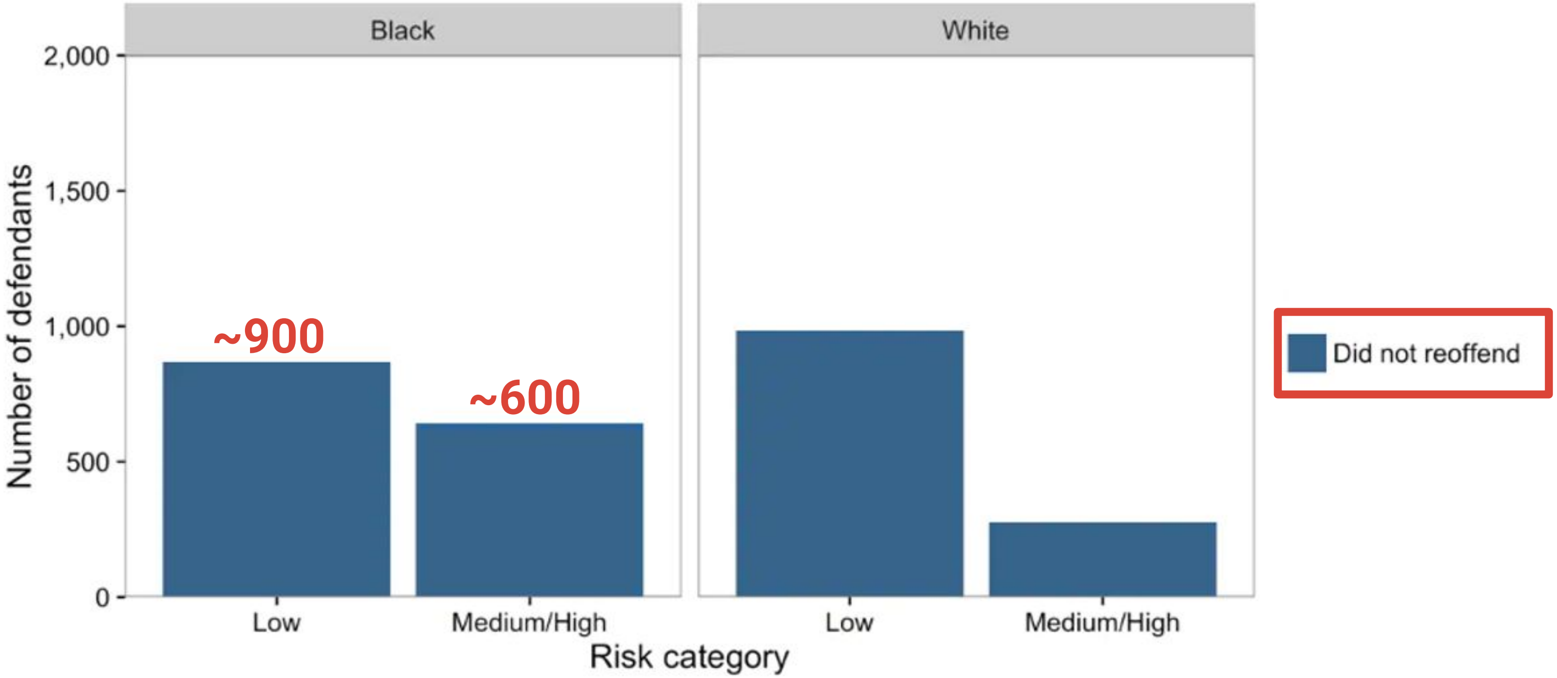
Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

Is COMPAS fair?



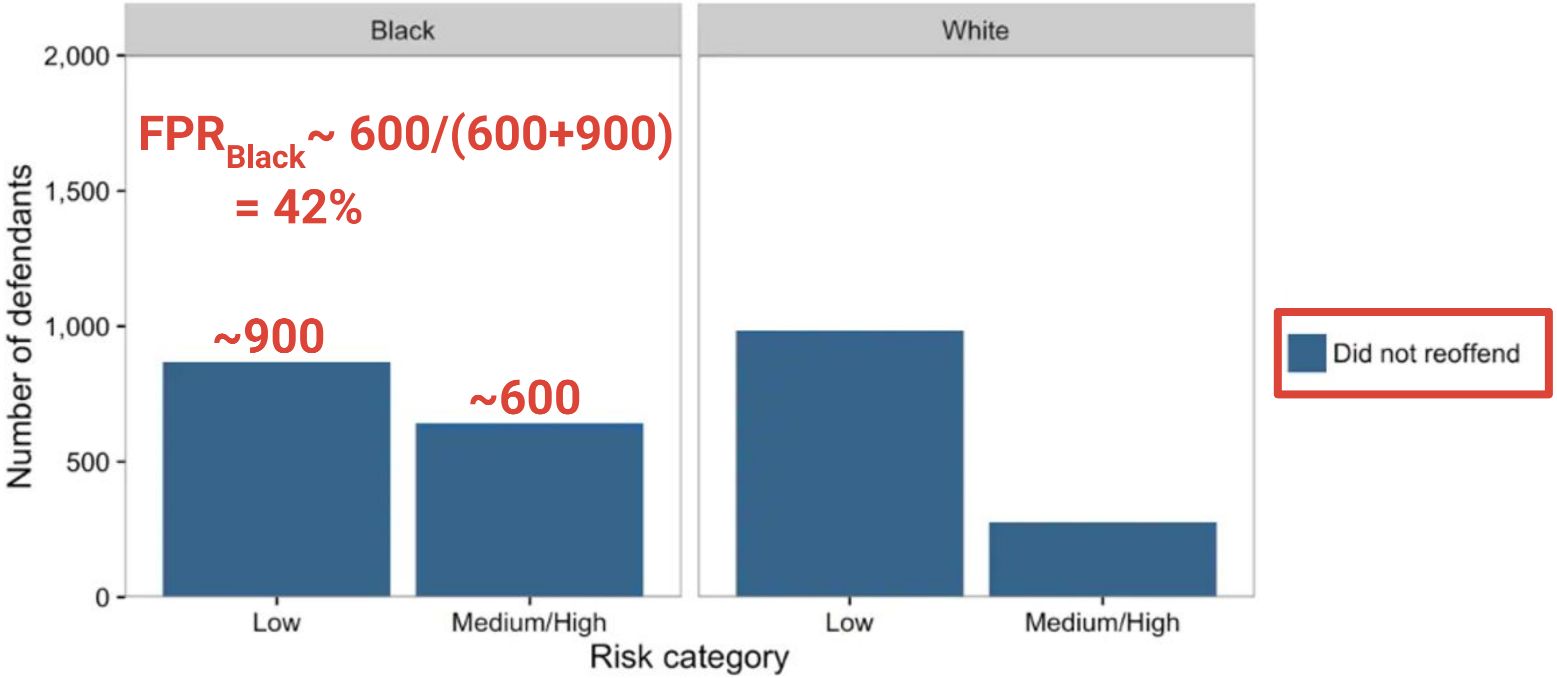
Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

Is COMPAS fair? False Positive Rates



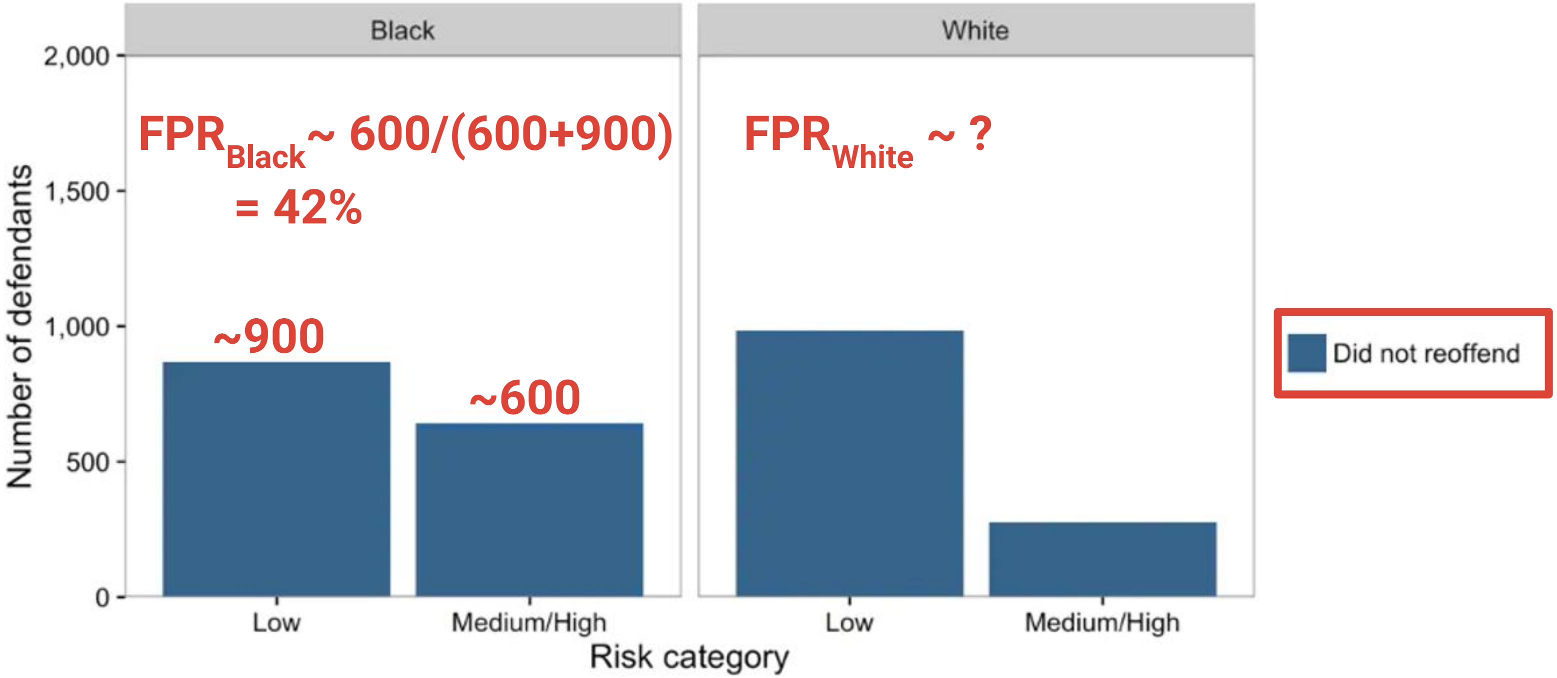
Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

Is COMPAS fair? False Positive Rates



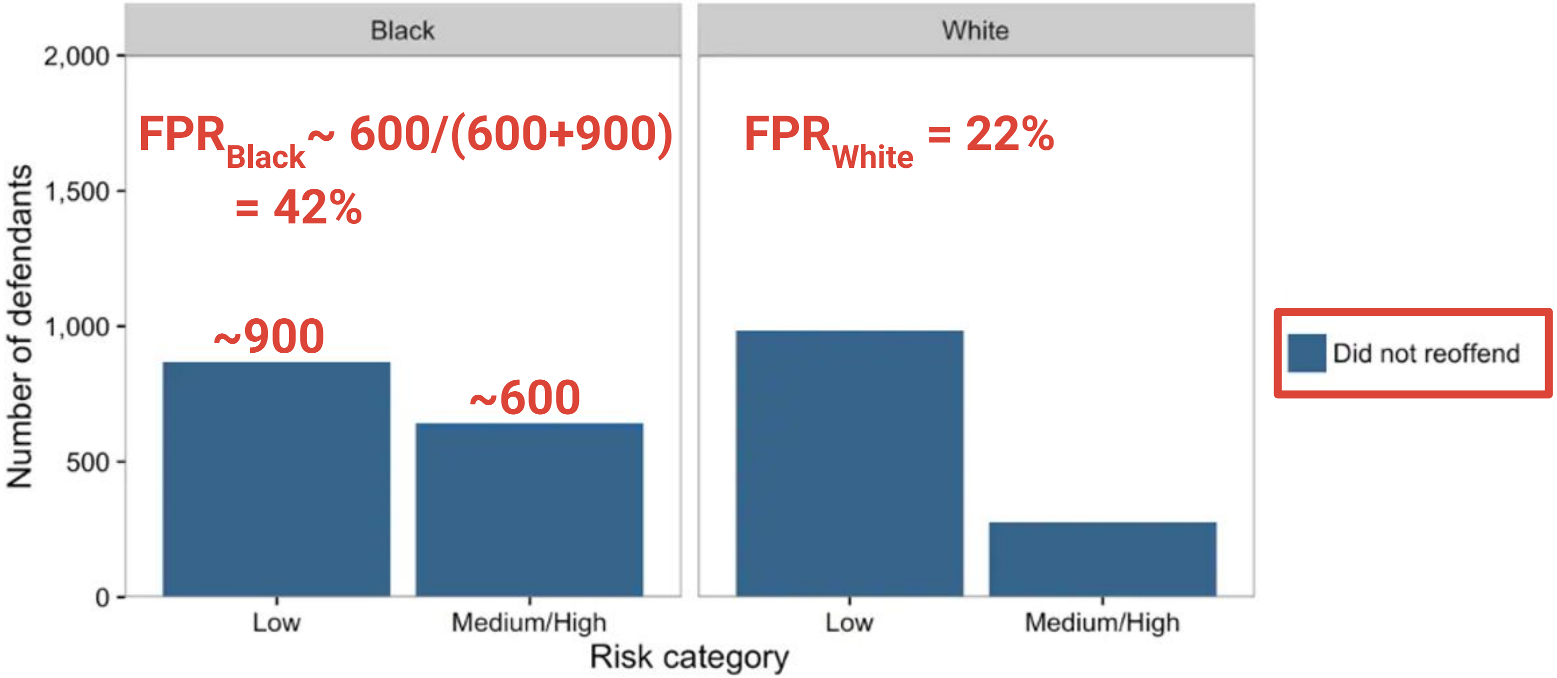
Source: Corbett-Davies, Pierson, Feller, and Goel. “A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.” *The Washington Post*, 2016.

Think, Pair, Share: Is FPR_{White} != FPR_{Black}?



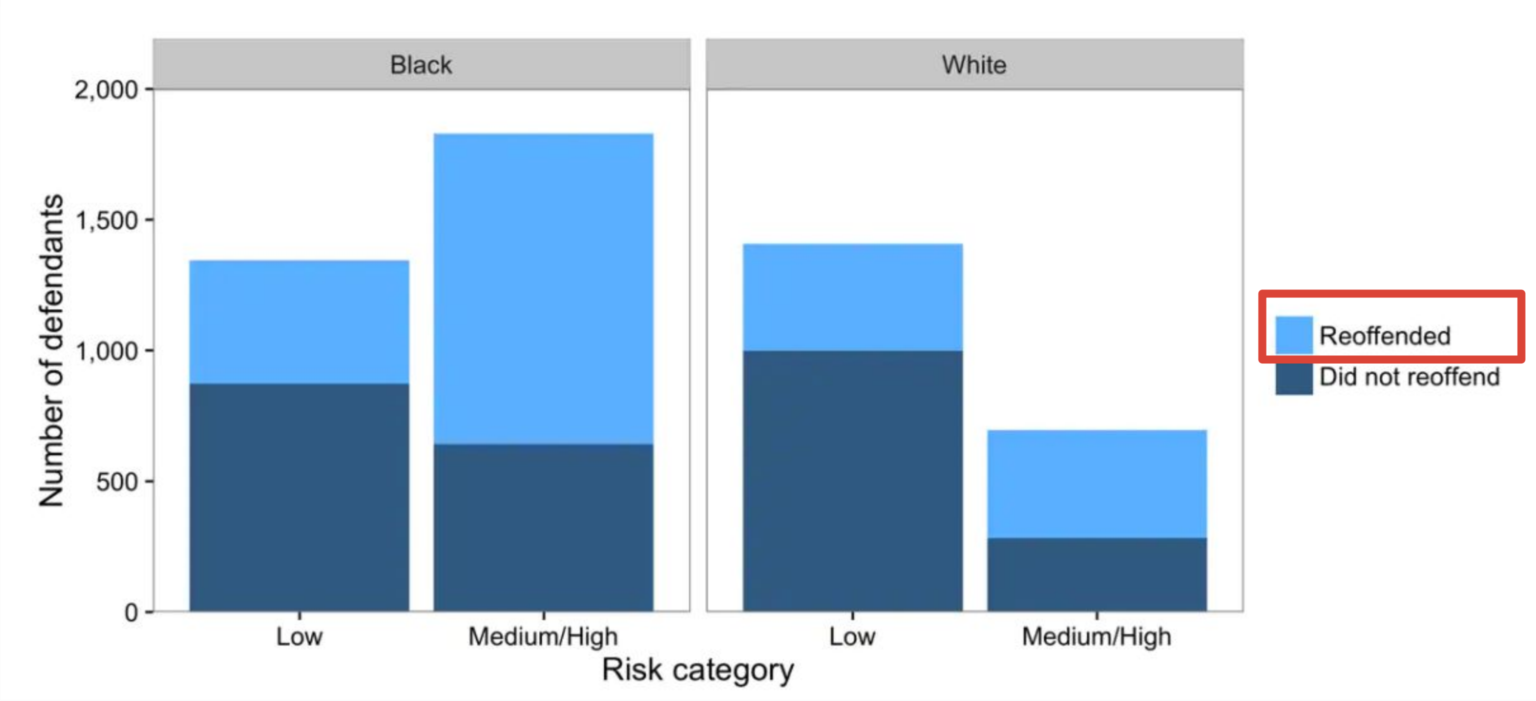
Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

$FPR_{Black} >> FPR_{White}$



Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

FNR are also different between groups (violating predictive equality)



In general, how <u>important</u> do you think it is for (any) algorithm to satisfy *predictive equality* (equal FPR and FNRs across groups)? ✋

Not very

Somewhat

Very

Is COMPAS fair?

ProPublica: No

Observation	**Fairness Principle**
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity
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

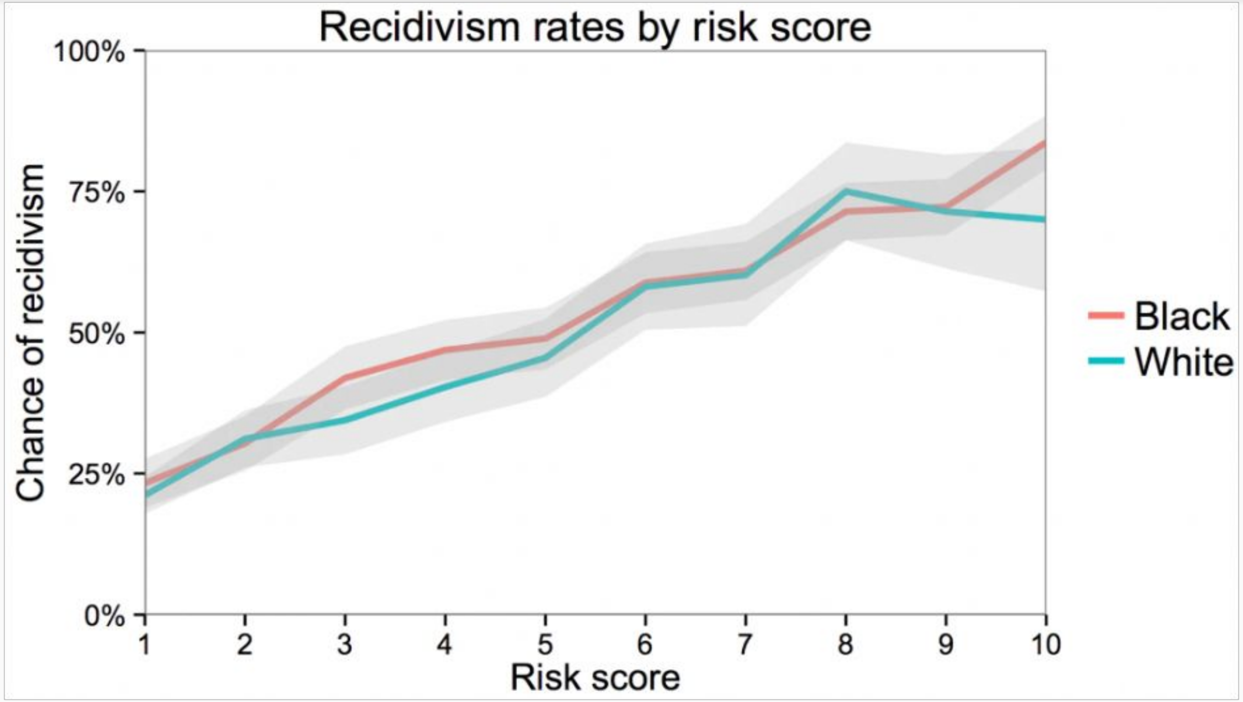
Is COMPAS fair?

NorthePointe: Yes

Observation	Fairness Principle
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity
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
Black defendants and white defendants with the same score are equally likely to reoffend.	**Calibration**

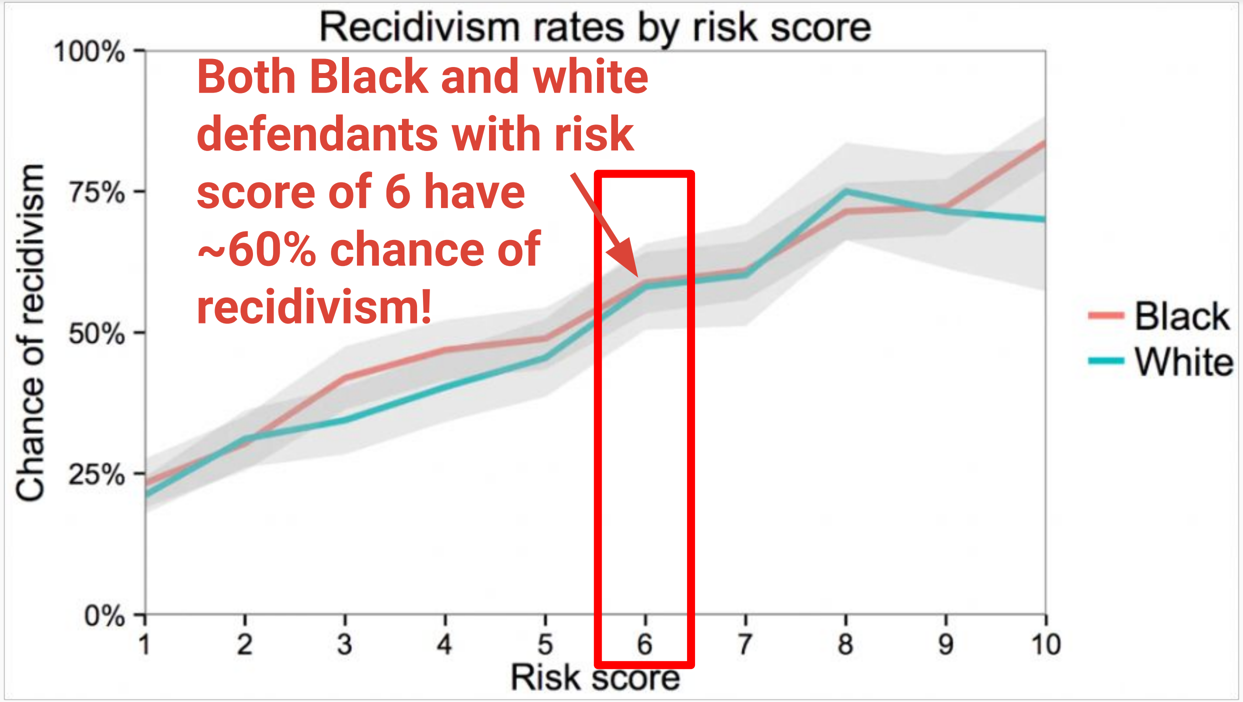
Source: Corbett-Davies, Pierson, Feller, and Goel. “A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.” *The Washington Post*, 2016.

Is COMPAS fair?

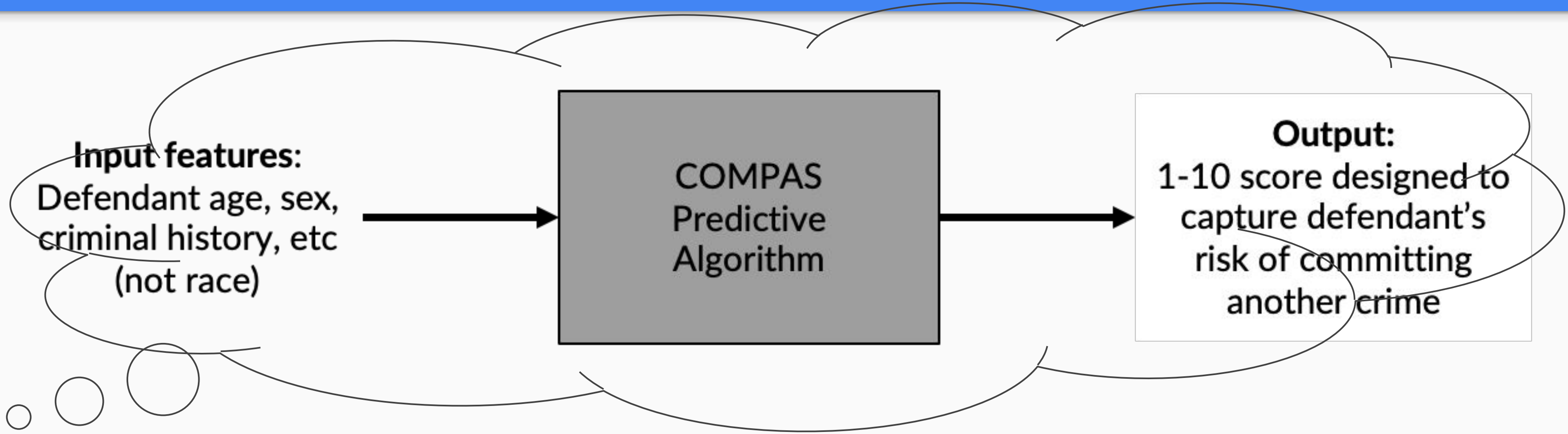


Source: Corbett-Davies, Pierson, Feller, and Goel. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 2016.

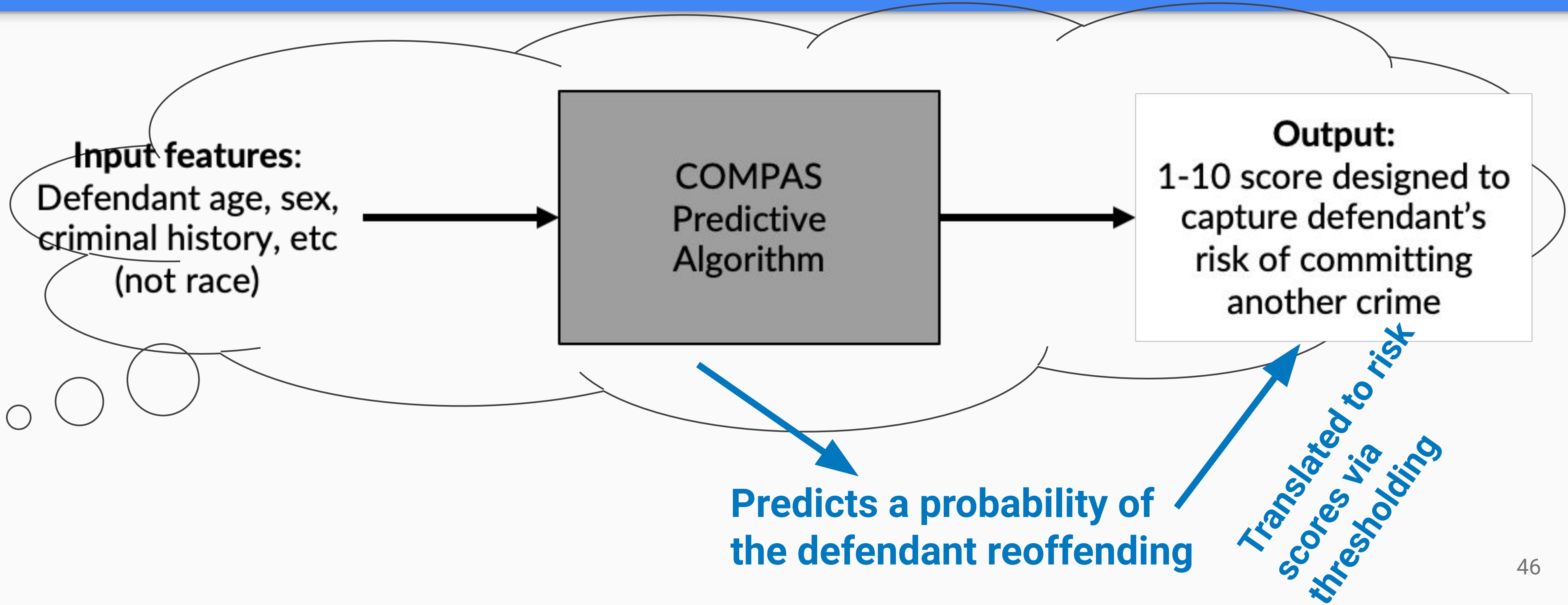
Calibration: the recidivism probability is the same between groups, conditional on risk score



The COMPAS algorithm

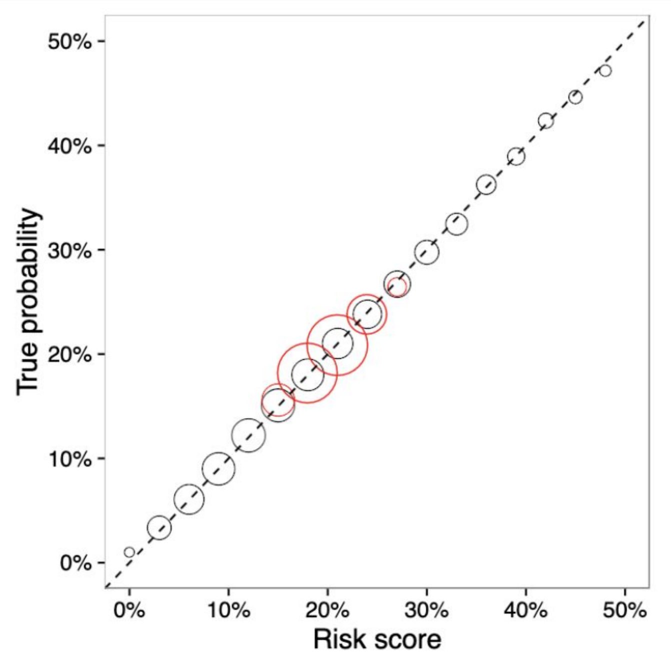


The COMPAS algorithm



Calibration with probabilistic scores

- Often when people say a risk score is calibrated, they're talking about risk models which actually output probabilities (e.g., "20%", not "6" for a defendant)
- In this case, *calibration* means that if you look at all people who get a 20% from the algorithm, 20% of them should actually reoffend
- "Scores mean the same thing for both groups" → i.e., the actual probability of an event



In general, how <u>important</u> do you think it is for (any) algorithm to be *calibrated* (for each group, the same fraction of people in each bin are y=1)? ✋

Not very

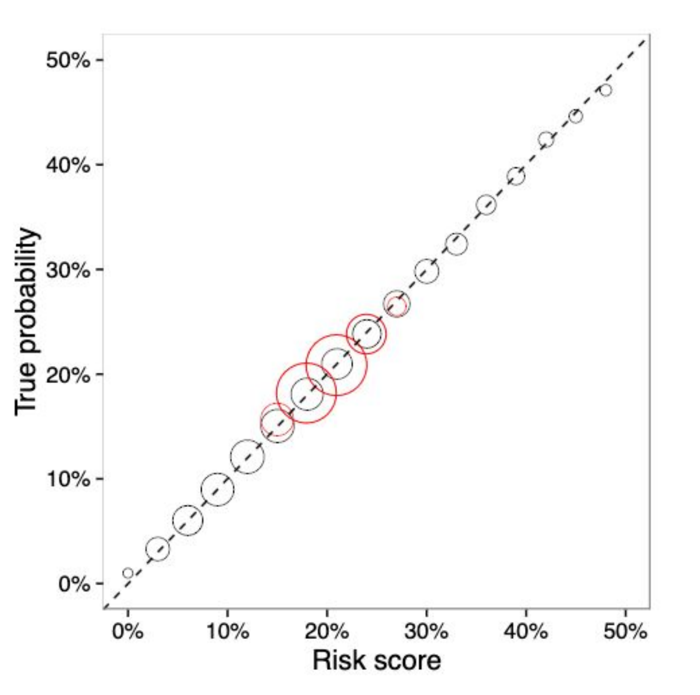
Somewhat

Very

Source: Corbett-Davies, Pierson, Feller, Goel, Huq. "Algorithmic decision making and the cost of fairness" *KDD, 2017.*

Calibration might be *necessary* but not *sufficient* for an algorithm to be fair

Is the algorithm that resulted in the red dots *calibrated?* →

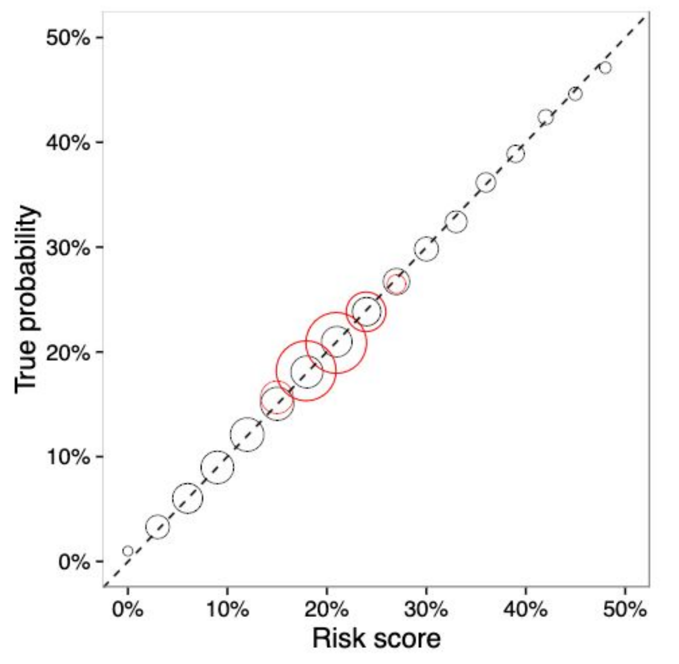


Source: Corbett-Davies, Pierson, Feller, Goel, Huq. "Algorithmic decision making and the cost of fairness" *KDD, 2017.*

Calibration might be *necessary* but not *sufficient* for an algorithm to be fair

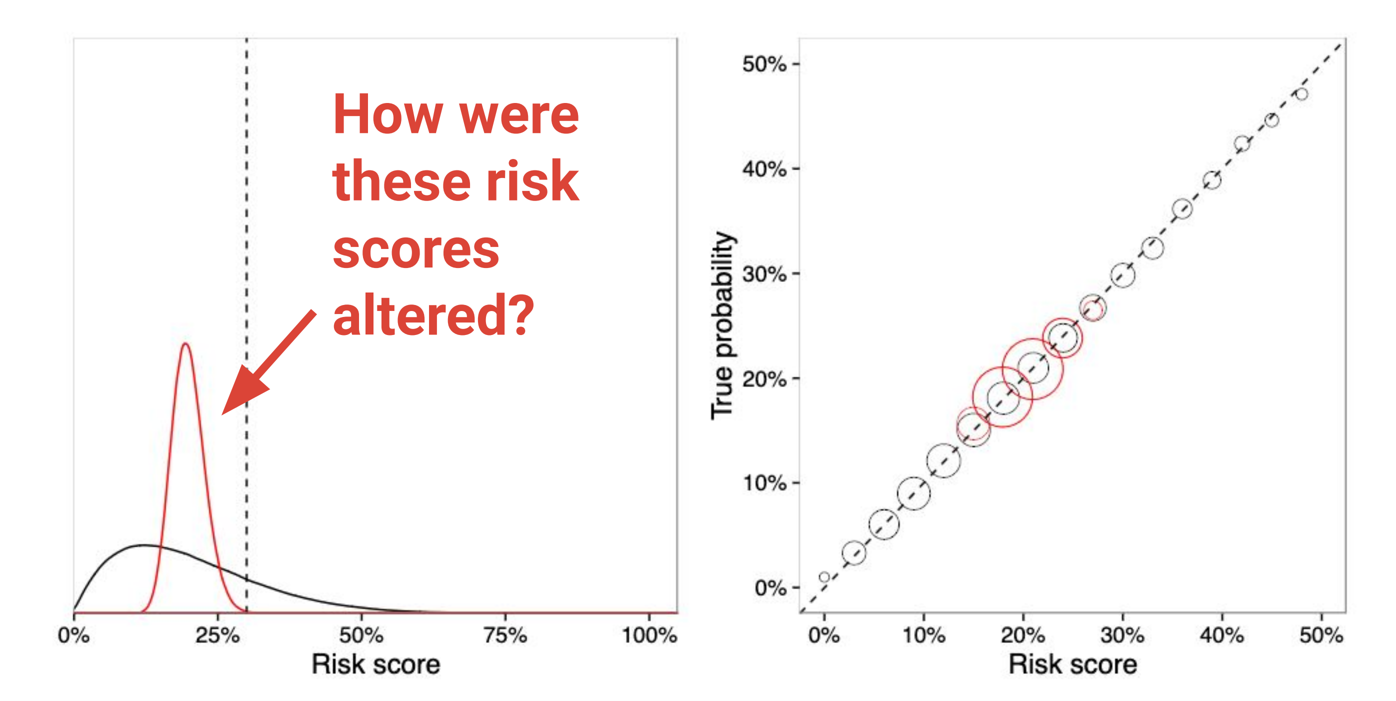
Is the algorithm that resulted in the red dots *calibrated?* →

Yes, but it was generated in a **deliberately** biased way, assigning *everyone in a group* a similar score!



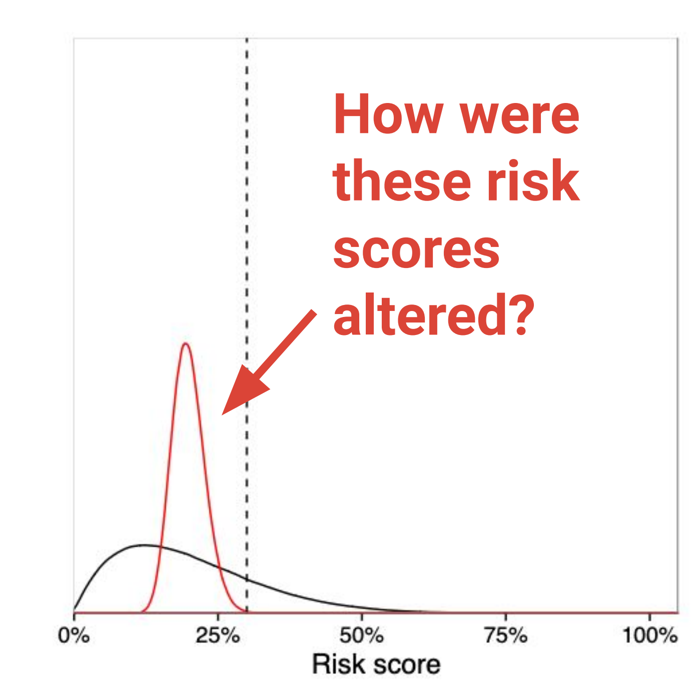
Source: Corbett-Davies, Pierson, Feller, Goel, Huq. "Algorithmic decision making and the cost of fairness" *KDD, 2017.*

Calibration might be *necessary* but not *sufficient* for an algorithm to be fair



Source: Corbett-Davies, Pierson, Feller, Goel, Huq. "Algorithmic decision making and the cost of fairness" *KDD, 2017.*

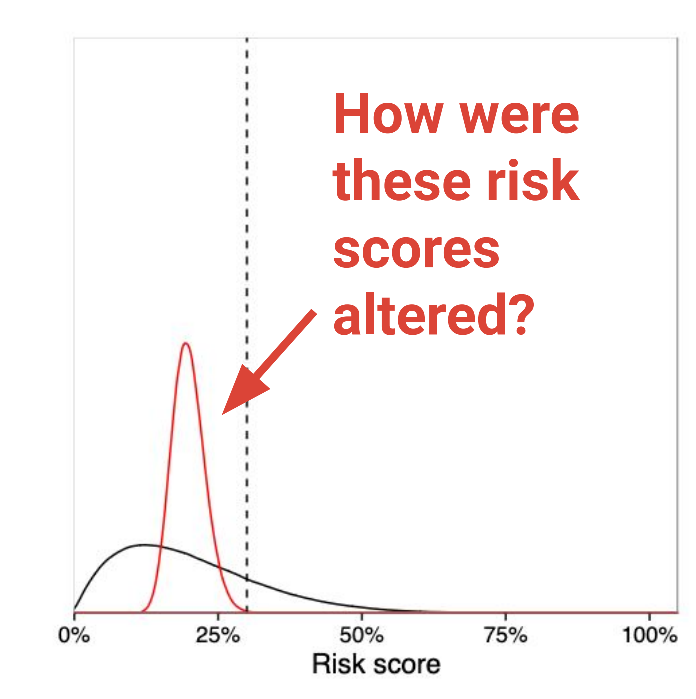
Calibration might be *necessary* but not *sufficient* for an algorithm to be fair



- Ex.: Deliberately bad algorithm predicts "10% chance of reoffending" for every Black defendant
- For white defendants, use a better algorithm which predicts different risks for each person
- Then threshold, e.g. "a person stays in jail if they're above a 5% chance of reoffending"

Source: Corbett-Davies, Pierson, Feller, Goel, Huq. "Algorithmic decision making and the cost of fairness" *KDD, 2017.*

Calibration might be *necessary* but not *sufficient* for an algorithm to be fair



- You can still be calibrated when deliberately ignoring relevant information for that group, so the risk scores are less informative
- Then, the decision-maker can apply a threshold that treats people with certain scores badly!

Is COMPAS fair?

NorthePointe: Yes

Observation	**Fairness Principle**
Black defendants are more likely than white defendants to be classified as high risk	Statistical parity
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
Black defendants and white defendants with the same score are equally likely to reoffend.	**Calibration**

Confusingly, all these things are referred to by multiple names

- Unfortunate consequence of a fast-moving field
- Statistical parity is also known as demographic parity or independence, related to notions of disparate impact
- Predictive equality is a.k.a. **equalized odds** or balance for the positive/negative class
  - Slightly weaker definition only requiring equal FNRs is called ***equal opportunity***
- Be sure to understand the definitions people are using, and be precise with your own language!

1 minute break!

Principle 1:
There are tradeoffs between fairness definitions

Sources: Kleinberg, Mullainathan, and Raghavan. “Inherent Tradeoffs in the Fair Determination of Risk Scores”. *ITCS*, 2017.
Chouldechova. “Fair prediction with disparate impact: a study of bias in recidivism prediction instruments”. *Big Data*, 2017.

The Bad News

- You can’t have all these fairness properties at the same time!
  - For example, if you have calibration… you can’t also have equal false positive/negative rates* (except in special cases)

Sources: Kleinberg, Mullainathan, and Raghavan. “Inherent Tradeoffs in the Fair Determination of Risk Scores”. *ITCS*, 2017.
Chouldechova. “Fair prediction with disparate impact: a study of bias in recidivism prediction instruments”. *Big Data*, 2017.

The Bad News

- You can’t have all these fairness properties at the same time!
  - For example, if you have calibration… you can’t also have equal false positive/negative rates* (except in special cases)
- Academics prove mathematically that it’s impossible to simultaneously satisfy these fairness definitions at the same time
  - *unless you have either perfect prediction or equal base rates (super rare in practice)
  - Kleinberg-Mullainathan-Raghavan 2016; Chouldechova 2016

Another example of fairness tradeoffs

Algorithmic decision making and the cost of fairness

Sam Corbett-Davies
Stanford University
scorbett@stanford.edu

Emma Pierson
Stanford University
emmap1@stanford.edu

Avi Feller
Univ. of California, Berkeley
afeller@berkeley.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

Aziz Huq
University of Chicago
huq@uchicago.edu

Setup

- Imagine you're a judge trying to decide whom to detain before trial
- Assumptions:
  - Pay some cost c for every defendant you detain
  - 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 detain?

Unconstrained by fairness: Apply a single threshold

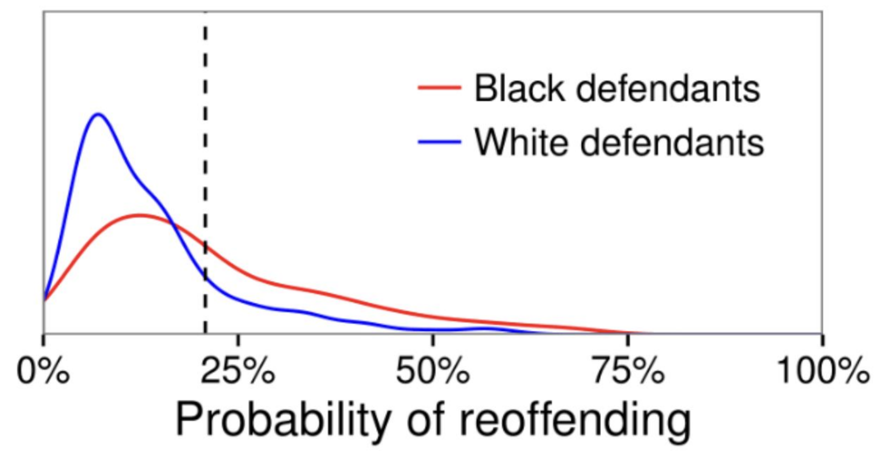
- Detain every defendant who's more likely than $p = c$ to commit another crime
- Apply a *single threshold* to all defendants
- What if you care about satisfying notions of fairness like statistical parity?

Constrained by fairness: Apply multiple thresholds

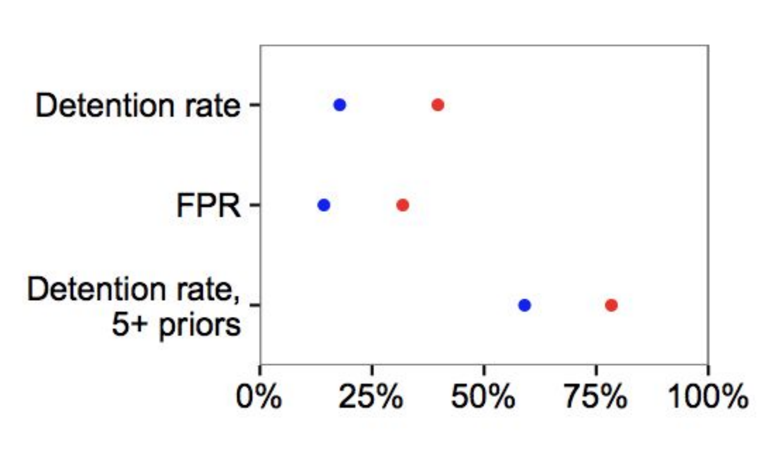
- If you want to satisfy statistical parity or predictive equality, your optimal behavior is to apply *multiple, group-specific thresholds*

Either way has downsides!

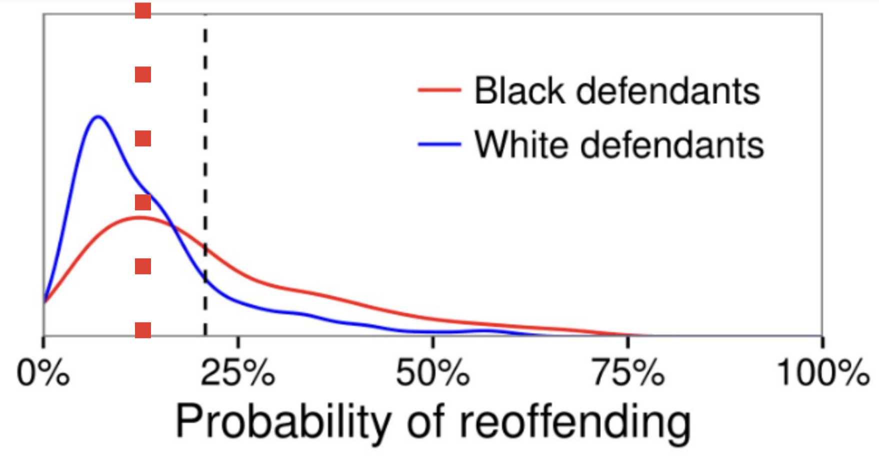
Different risk distributions → different detention rates when sharing a single threshold!



Single threshold



Either way has downsides!



What if we use multiple thresholds? Then we're holding people to different standards… is this legally/ethically okay?

Either way has downsides!

And, with multiple thresholds, here we see:

→ more low-risk individuals are detained
→ violent crime goes up

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%

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities "right"?**
    a. Is the outcome we're predicting biased?
        i. The historical statistics of who has offended/reoffended, but there's bias in *who gets arrested*, *where police are stationed*, *etc.*
    b. Could we collect more data?

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities “right”?**
2. **Could we make a different decision?**
    a. Maybe we don’t have to hold people in jail.

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities “right”?**
2. **Could we make a different decision?**
3. **Are costs the same for all defendants?**
    a. Is it more costly to hold a single parent in jail?

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities “right”?**
2. **Could we make a different decision?**
3. **Are costs the same for all defendants?**
4. **Can defendants even be classified into only one group?**
    a. Intersectionality is harder to capture (what if one defendant is a member of multiple protected groups?)

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities “right”?**
2. **Could we make a different decision?**
3. **Are costs the same for all defendants?**
4. **Can defendants even be classified into only one group?**
5. **Do we care about more than immediate costs?**
    a. What about long-term impacts?

Question: the math in this paper is (hopefully) right given the assumptions. But how might we push back on the assumptions?

1. **Are the probabilities "right"?**
2. **Could we make a different decision?**
3. **Are costs the same for all defendants?**
4. **Can defendants even be classified into only one group?**
5. **Do we care about more than immediate costs?**
6. **Some decisions aren't just a property of the individual.**
    a. E.g. we don't just want a college class of excellent trombone players

Okay, so we can’t satisfy all fairness definitions at once. How do we choose which is most relevant?

Source: Arvind Narayanan, https://www.youtube.com/watch?v=jIXIuYdnyyk

It might depend on the stakeholders!

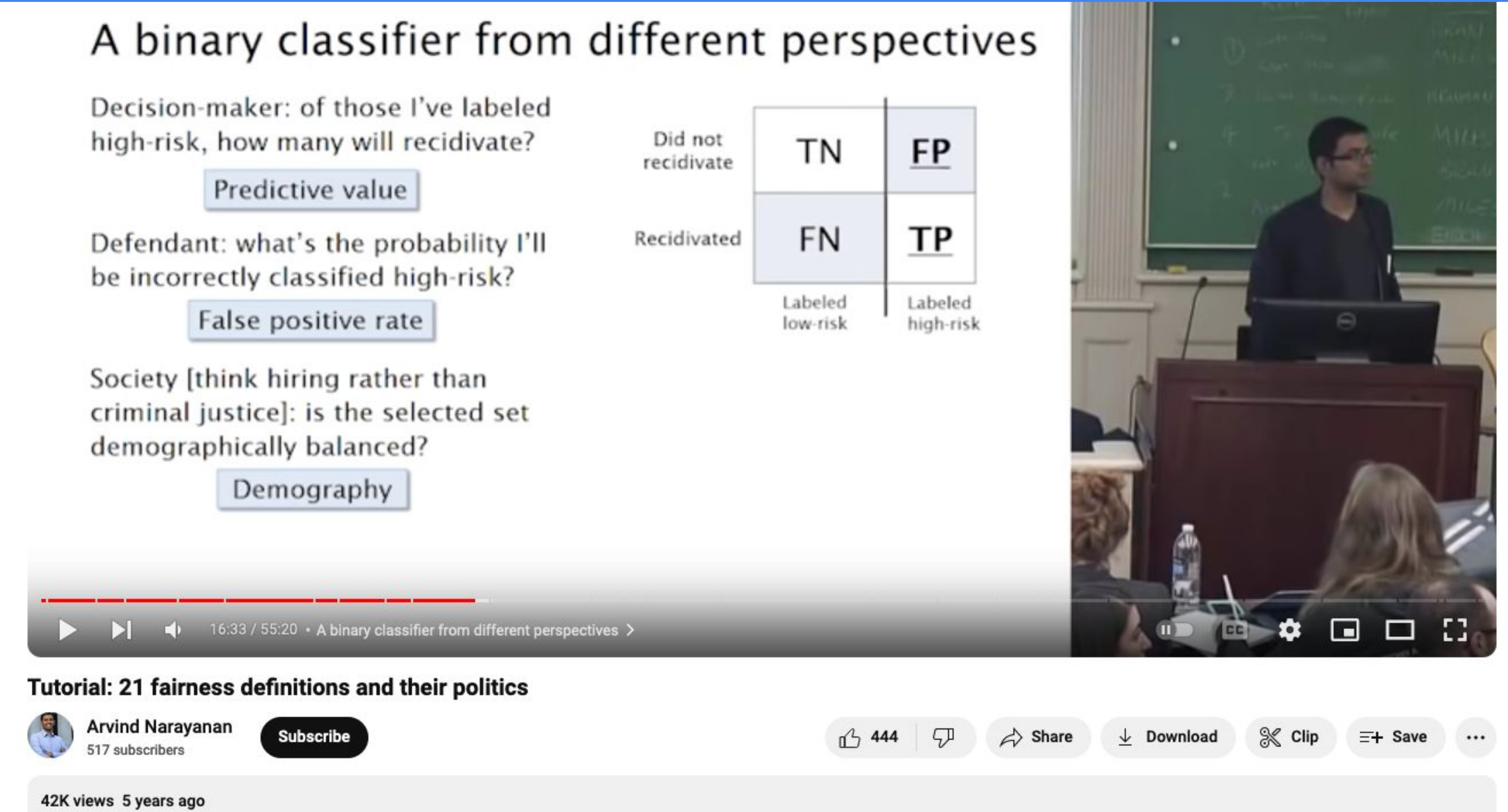


Image Sources: UpToDate, American College of Radiology.

Choosing a fairness definition: Understand what decisions are being made!

Assessment	Management	Likelihood of cancer
Category 0: Incomplete – Need additional imaging evaluation and/or prior mammograms for comparison	Recall for additional imaging and/or comparison with prior examination(s)	N/A
Category 1: Negative	Routine mammography screening	Essentially 0% likelihood of malignancy
Category 2: Benign	Routine mammography screening	Essentially 0% likelihood of malignancy

Image Sources: UpToDate, American College of Radiology.

Choosing a fairness definition: Understand what decisions are being made!

Assessment	Management	Likelihood of cancer
Category 0: Incomplete – Need additional imaging evaluation and/or prior mammograms for comparison	Recall for additional imaging and/or comparison with prior examination(s)	N/A
Category 1: Negative	Routine mammography screening	Essentially 0% likelihood of malignancy
Category 2: Benign	Routine mammography screening	Essentially 0% likelihood of malignancy
Category 3: Probably benign	Short-interval (6-month) follow-up or continued surveillance mammography	>0 but ≤2% likelihood of malignancy

Image Sources: UpToDate, American College of Radiology.

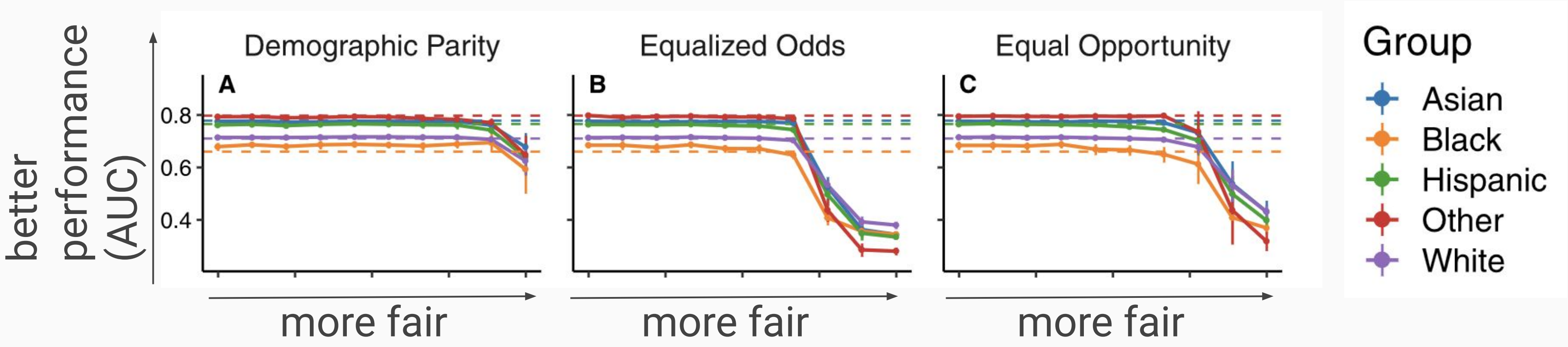
Choosing a fairness definition: Understand what decisions are being made!

Assessment	Management	Likelihood of cancer
Category 0: Incomplete – Need additional imaging evaluation and/or prior mammograms for comparison	Recall for additional imaging and/or comparison with prior examination(s)	N/A
Category 1: Negative	Routine mammography screening	Essentially 0% likelihood of malignancy
Category 2: Benign	Routine mammography screening	Essentially 0% likelihood of malignancy
Category 3: Probably benign	Short-interval (6-month) follow-up or continued surveillance mammography	>0 but ≤2% likelihood of malignancy
Category 4: Suspicious	Tissue diagnosis*	>2 but <95% likelihood of malignancy
Category 4A: Low suspicion for malignancy		>2 to ≤10% likelihood of malignancy
Category 4B: Moderate suspicion for malignancy		>10 to ≤50% likelihood of malignancy
Category 4C: High suspicion for malignancy		>50 to <95% likelihood of malignancy

> 2% threshold → biopsy, because false negatives are really bad!

Source: Pfohl, Foryciarz, and Shah. "An empirical characterization of fair machine learning for clinical risk prediction." Journal of Biomedical Informatics, 2021.

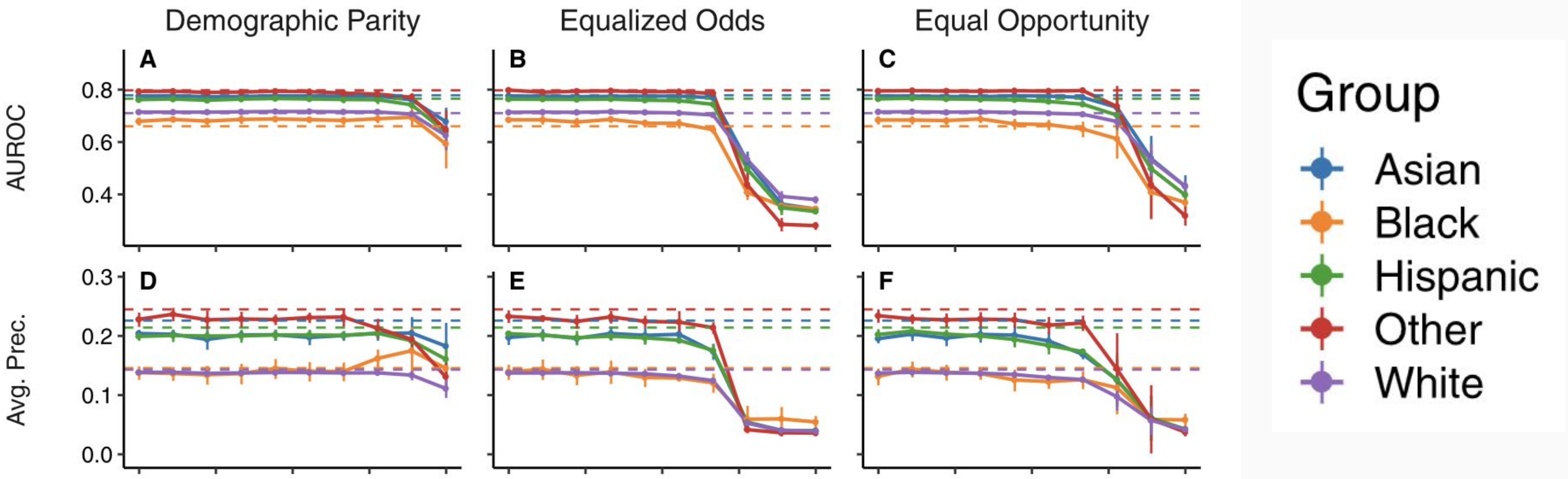
Be careful when applying *generic* "debiasing" methods to *nuanced* use cases



In some (but not all!) contexts – e.g., this research on clinical risk prediction – models optimizing for different fairness metrics can correspond to worse model predictions for patients

Source: Pfohl, Foryciarz, and Shah. "An empirical characterization of fair machine learning for clinical risk prediction." Journal of Biomedical Informatics, 2021.

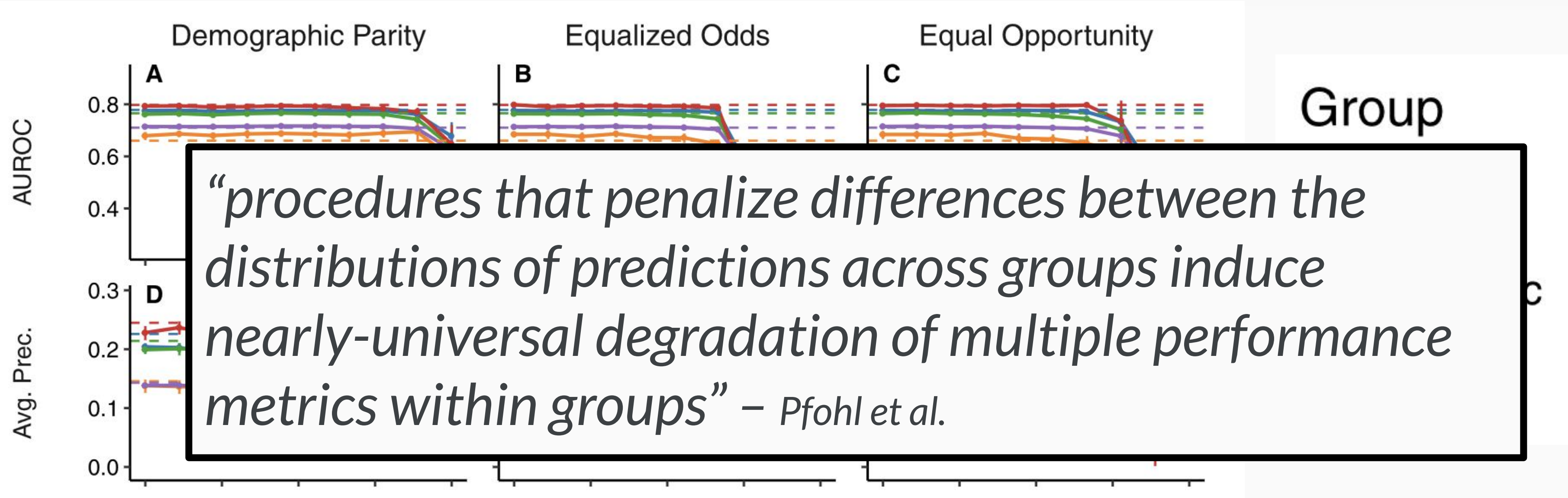
Be wary of generic "debiasing" methods



Note: it's not the case that making algorithms fairer always harms performance! It's just that *some* ways of making algorithms "fairer" don't always make much sense.

Source: Pfohl, Foryciarz, and Shah. "An empirical characterization of fair machine learning for clinical risk prediction." Journal of Biomedical Informatics, 2021.

Be wary of generic "debiasing" methods



"procedures that penalize differences between the distributions of predictions across groups induce nearly-universal degradation of multiple performance metrics within groups" – *Pfohl et al.*

Note: it's not the case that making algorithms fairer always harms performance! It's just that *some* ways of making algorithms "fairer" don't always make much sense.

Zooming out: broader questions about the COMPAS case study

Why race?

- In general, we often examine fairness with respect to *sensitive features* like race, gender, age, or socioeconomic status because we know there is bias along these dimensions (in our contexts)

- We want to avoid worsening existing inequality

Should algorithms be used in criminal justice at all?

Think, Pair, Share

What are some of the pros and cons of using algorithms in the criminal justice system?

Source: https://criminaljustice.tooltrack.org/tool/16627

Regardless, COMPAS is still being used

RISK ASSESSMENT
TOOL DATABASE
AT BERKMAN KLEIN CENTER

BACK TO ALL TOOLS

States Used

Rollover states to see highlights

Selected state
Jurisdictions using COMPAS

CALIFORNIA

Counties Using

5

Stage of Intended Use

Probation, Pretrial release, Other

Stages of Use	Number of Jurisdictions
Pretrial release	3
Other	1
Probation	1

View California »

Principle 2: Be precise about what you mean by bias!

Source: Arvind Narayanan, https://www.youtube.com/watch?v=jIXIuYdnyyk

There are many definitions of bias – be precise about what you mean!

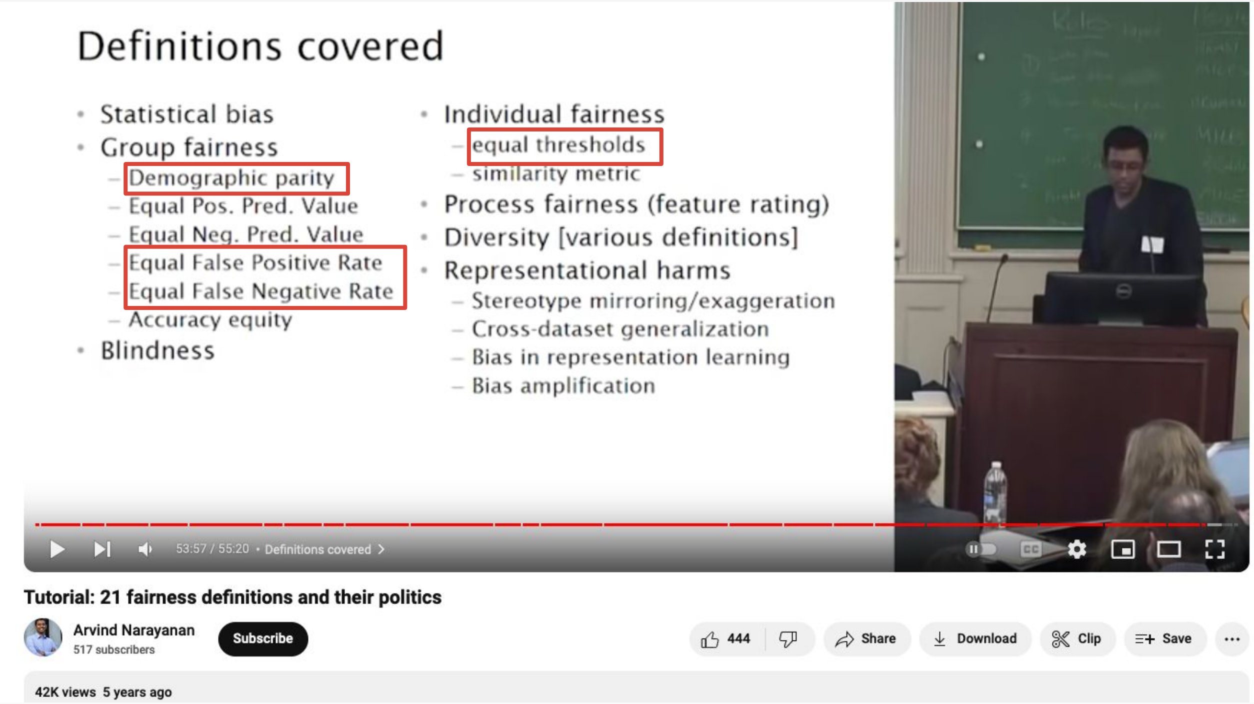


Image sources: *The New York Times, The Washington Post, The Guardian.*

The Washington Post
Democracy Dies in Darkness

INNOVATIONS

Twitter drops automated image-cropping tool after determining it was biased

The New York Times

Dealing With Bias in Artificial Intelligence

The New York Times

We Teach A.I. Systems Everything, Including Our Biases

The New York Times

Using A.I. to Find Bias in A.I.

AI expert calls for end to UK use of 'racially biased' algorithms

The Washington Post
Democracy Dies in Darkness

BUSINESS

OpenAI Project Risks Bias Without More Scrutiny

Hypothetical example: course recommendation algorithm is less likely to recommend computer science courses to women

Hypothetical example: course recommendation algorithm is less likely to recommend computer science courses to women

Here, we refer to two gender groups – men and women– for the sake of simplicity. Of course, there are many more gender identities which makes the problem more nuanced than presented here.

Possible causes of bias

Possible reason 1: women weren’t allowed to take computer science classes until recently

Possible causes of bias

Possible reason 2: professors are biased about which students they let into their classes

Possible causes of bias

Possible reason 3: we aren’t collecting the features we need to predict which classes women will take

Possible causes of bias

Possible reason 4: we're fitting the same model for students of all genders, and it doesn't work as well for women.

Possible causes of bias

Possible reason 5: we're doing recommendations based on which classes people *sign up for*, but should use what classes they complete/enjoy.

Possible causes of bias

Possible reason 6: we’re only training the algorithm on people who *take* computer science classes, and it’s biased when we run it on everyone else.

Possible causes of bias

Possible reason 7: maybe women truly don’t want to sign up for your computer classes due to larger social biases.

Research finding: STEM career ads less likely to be recommended to women

"Empirically, however, fewer women saw the [STEM career] ad than men. This happened because **younger women** are a prized demographic and are **more expensive** to show ads to."

Anja Lambrecht, Catherine Tucker. "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads" (2019)

Lecture takeaways

- There are many definitions of fairness, and you cannot (in general) achieve all of them at the same time

Lecture takeaways

- There are many definitions of fairness, and you cannot (in general) achieve all of them at the same time
- Don't give up on an algorithm just because it doesn't satisfy a particular fairness property if that isn't relevant to the decision
- Consider which fairness definitions are relevant to the task, consulting with stakeholders and domain experts

Lecture takeaways

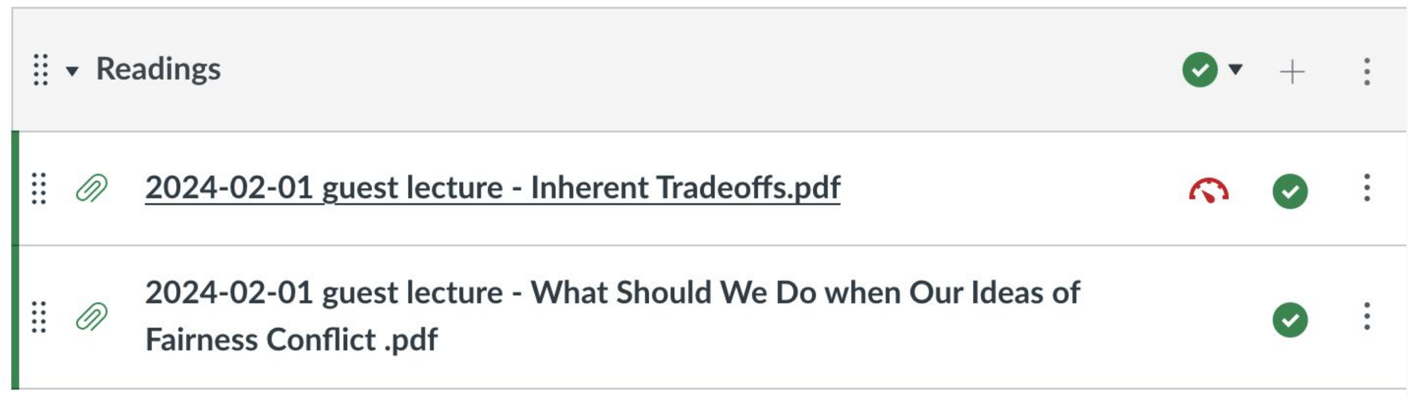
- There are many definitions of fairness, and you cannot (in general) achieve all of them at the same time
- Don't give up on an algorithm just because it doesn't satisfy a particular fairness property if that isn't relevant to the decision
- Consider which fairness definitions are relevant to the task, consulting with stakeholders and domain experts
- Don't seek to fulfill fairness properties which don't make sense for your use case, and be careful using automatic "debiasing" tools.

Lecture takeaways

- There are many definitions of fairness, and you cannot (in general) achieve all of them at the same time
- Don't give up on an algorithm just because it doesn't satisfy a particular fairness property if that isn't relevant to the decision
- Consider which fairness definitions are relevant to the task, consulting with stakeholders and domain experts
- Don't seek to fulfill fairness properties which don't make sense for your use case, and be careful using automatic "debiasing" tools.
- Be precise about what you mean by "bias" because you may be speaking to people with many different backgrounds, these topics are charged and easily misunderstood, and different types of bias imply different solutions

On Canvas: reading for Feb 1 guest lecture

Prof. Manish Raghavan



On Canvas: HW and Project Phase 1

Homework & Project Files

- HW1_DueFeb8.zip
- INFO4390 Phase 1 Rubric (due Feb 20).pdf