DSCI 531

Defining Fairness

Kristina Lerman

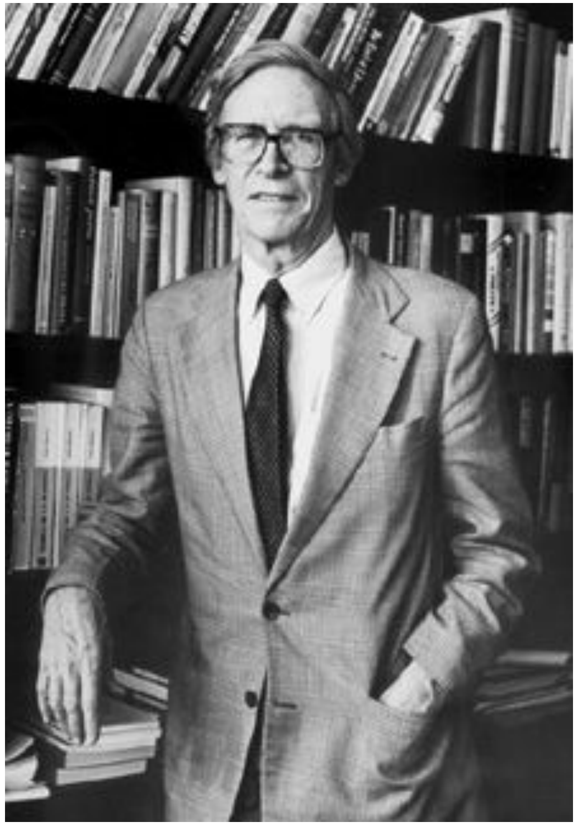
Spring 2025

What is fairness?

- Not one established definition, many from many different levels.
- Borrows from:
    - Social sciences
    - Philosophy
- In computer science, three macro definitions:
    - Group fairness
        - Treat different groups equally.
    - Individual fairness
        - Give similar predictions to similar individuals.
    - Subgroup fairness
        - Apply group fairness to a large collection of subgroups.

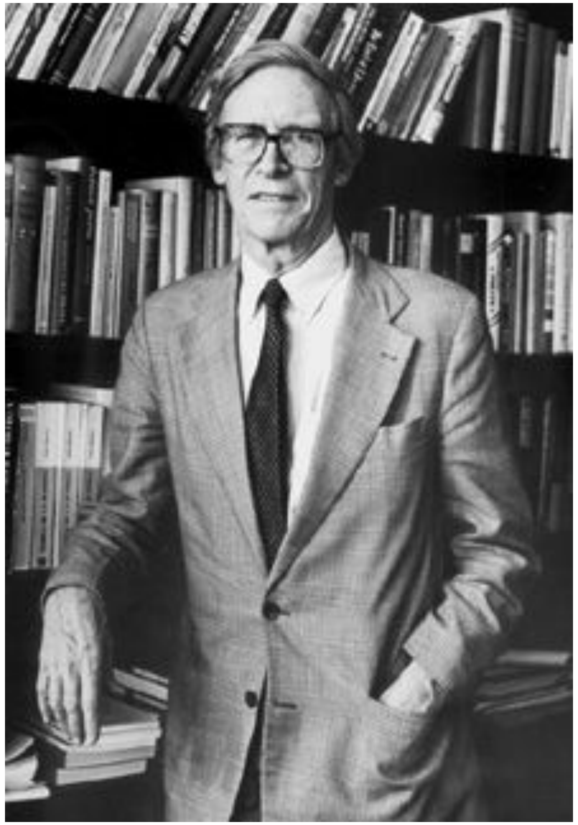
John Rawls (1921 – 2002)

- Developed theory of “Justice as Fairness”
- “Veil of ignorance“ thought experiment: Imagine you're designing a society, its rules and institutions, but you don't know what position you'll occupy in that society (your race, gender, social class, natural abilities, or any other personal characteristics).
- Rawls argued that from behind this veil of ignorance, rational people would choose principles that ensure a fair society for everyone, since they might end up in any position.



John Rawls (1921 – 2002)

- Developed theory of “Justice as Fairness”
- First principle: equal liberties (political system)
    - Everyone has equal right to the most extensive system of basic liberties compatible with similar liberties for others.
- Second principle: (law) Social and economic inequalities should be arranged so that
    - Equality of opportunity (first priority)
        - Everyone has a chance at an opportunity
    - Promote interests of least advantaged (second priority)
        - If everyone has an opportunity, greatest benefit for least-advantaged

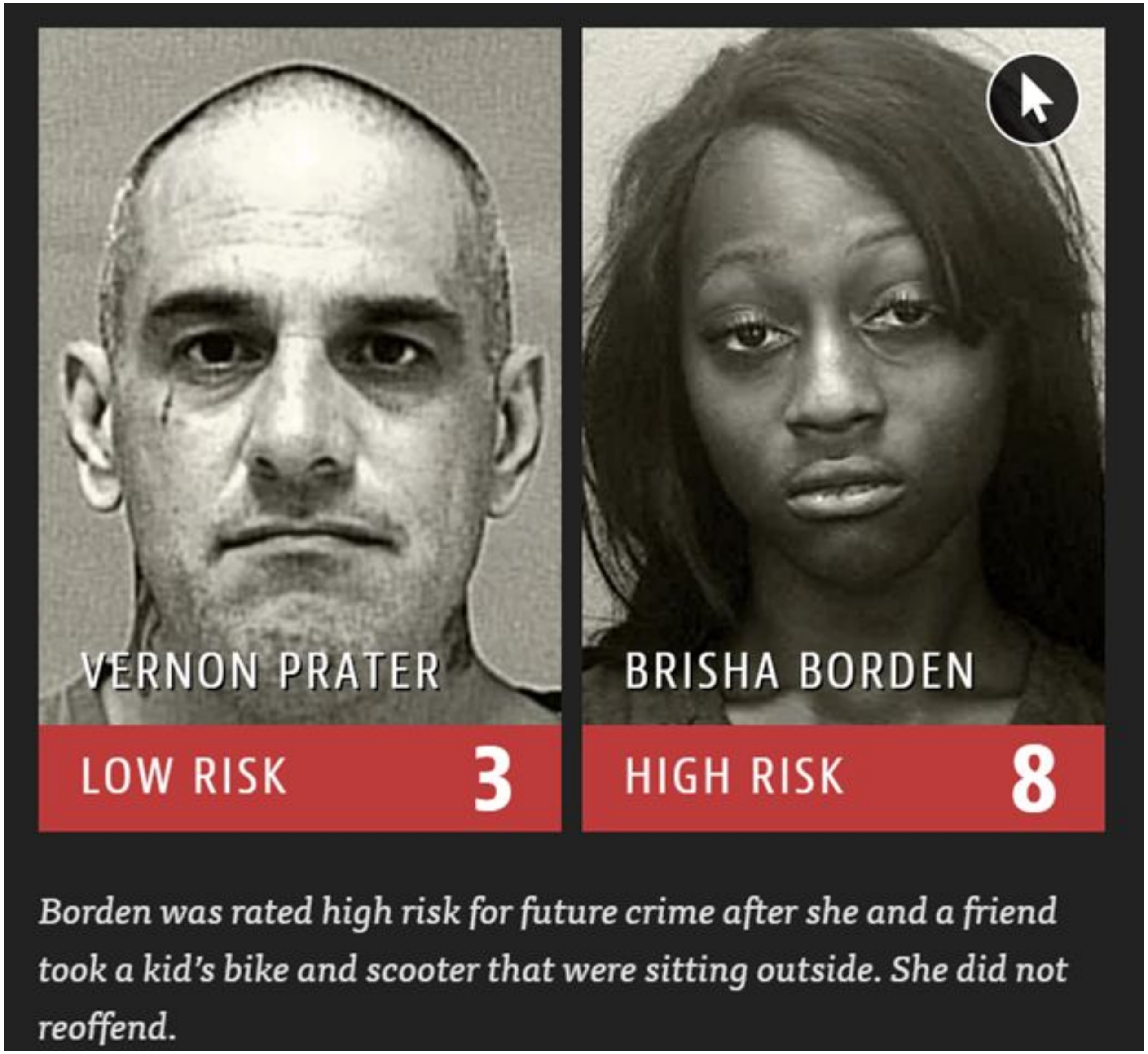


Rawls and AI fairness

- Veil of ignorance □ Following Rawls, we can think about designing AI algorithms without knowing which demographic groups we might belong to
- Encourages testing of AI across different demographic groups and use cases
- Difference Principle: Ensure that AI algorithms don't disproportionately harm disadvantaged groups
  - Any performance disparities between groups should only be accepted if they ultimately benefit the worst-off groups
  - When training models, we might prioritize minimizing error rates for disadvantaged groups over maximizing overall accuracy

Bias in automated criminal risk assessment

COMPAS tool systematically gives black defendants higher risk scores for future recidivism



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

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

ProPublica study

- Data: 7000 people arrested in Broward County, FL (2013-2014)
  - COMPAS algorithm predicts whether a defendant will commit a crime within next 2 years

Prior Offense	1 attempted burglary
LOW RISK	**3**
Subsequent Offenses	3 drug possessions

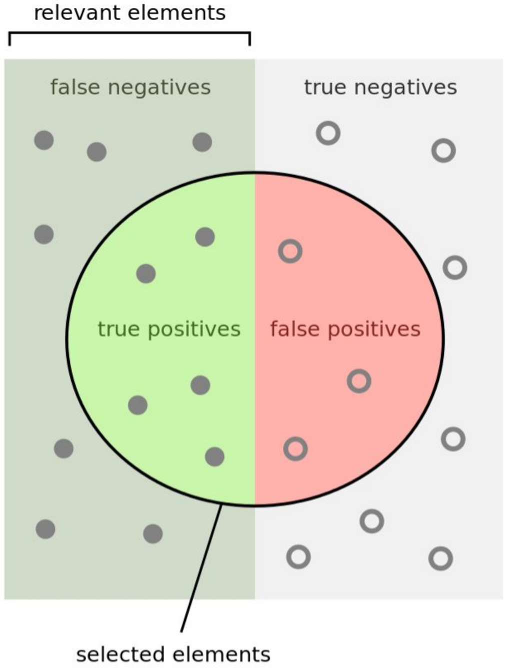
Prior Offense	1 resisting arrest without violence
HIGH RISK	**10**
Subsequent Offenses	None

	Did commit a crime	Did not commit a crime
High risk (will commit a crime)	True Positives	*False Positives*
Low risk (will not commit a crime)	*False Negatives*	True Negatives

Measuring classification performance

	Did commit a crime	Did not commit a crime
Will commit crime	True Positives (TP)	*False Positives* (FP)
Will not commit a crime	*False Negatives* (FN)	True Negatives (TN)

- Accuracy = Number of correct predictions/Total predictions
- Accuracy = TP + TN / (TP+TN+FP+FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- Sensitivity = TP rate = TP / (TP + FN)
- Specificity = TN rate = TN / (TN + FP)
- False positive rate = FP / (TN + FP) = 1 - Specificity

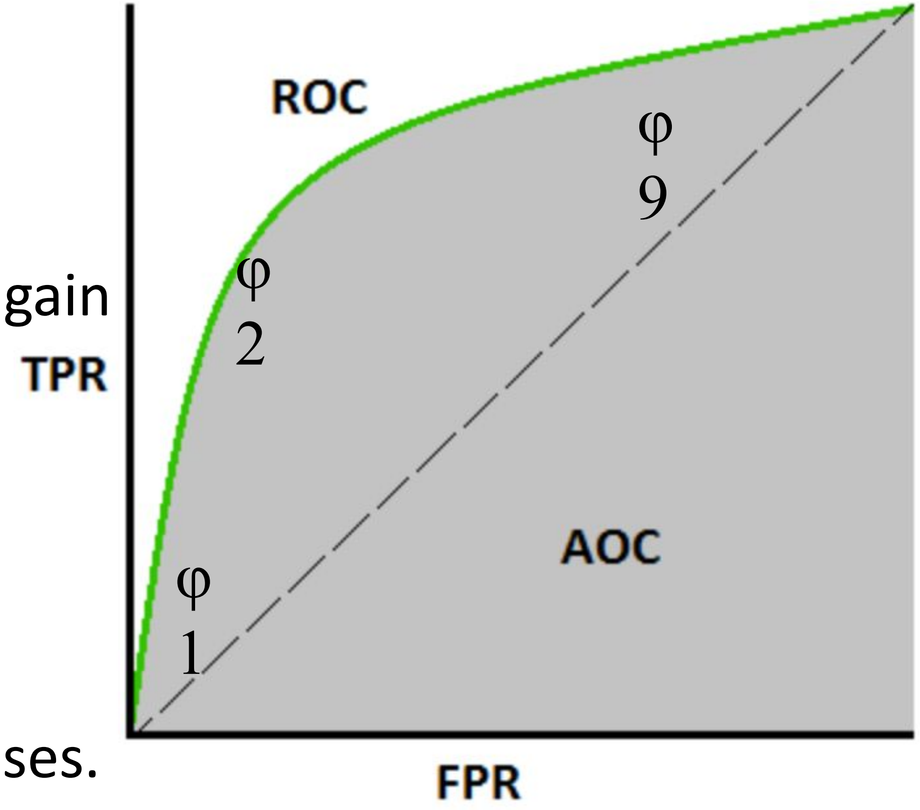


Measuring classification for non-binary outcomes

- COMPAS makes a prediction p(y) in [1, 10]

- Thresholding φ
  - If p(y) > φ then defendant will commit a crime again
  - Otherwise, defendant will not commit a crime
  - Converts to a binary prediction

- AUC – Area under ROC curve
  - How well a model can distinguish between classes.
  - **Probability** the model will score a randomly chosen positive class higher than a randomly chosen negative class



COMPAS performance

- Accuracy
  - 20% accuracy for violent crimes
  - 61% when all crimes are considered

- Asymmetry of mistakes: how do mistakes affect the least fortunate
  - False negatives
  - False positives

Racial disparities in COMPAS risk scores

Significant racial disparities: different types of mistakes for whites and blacks

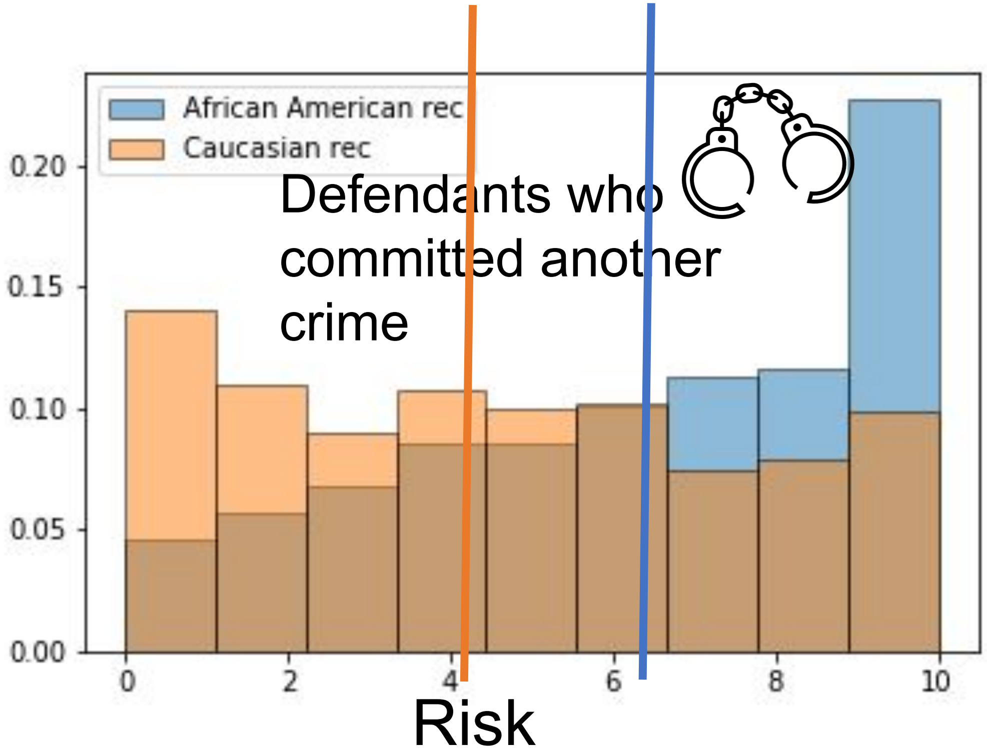
False Positive rate high for blacks:

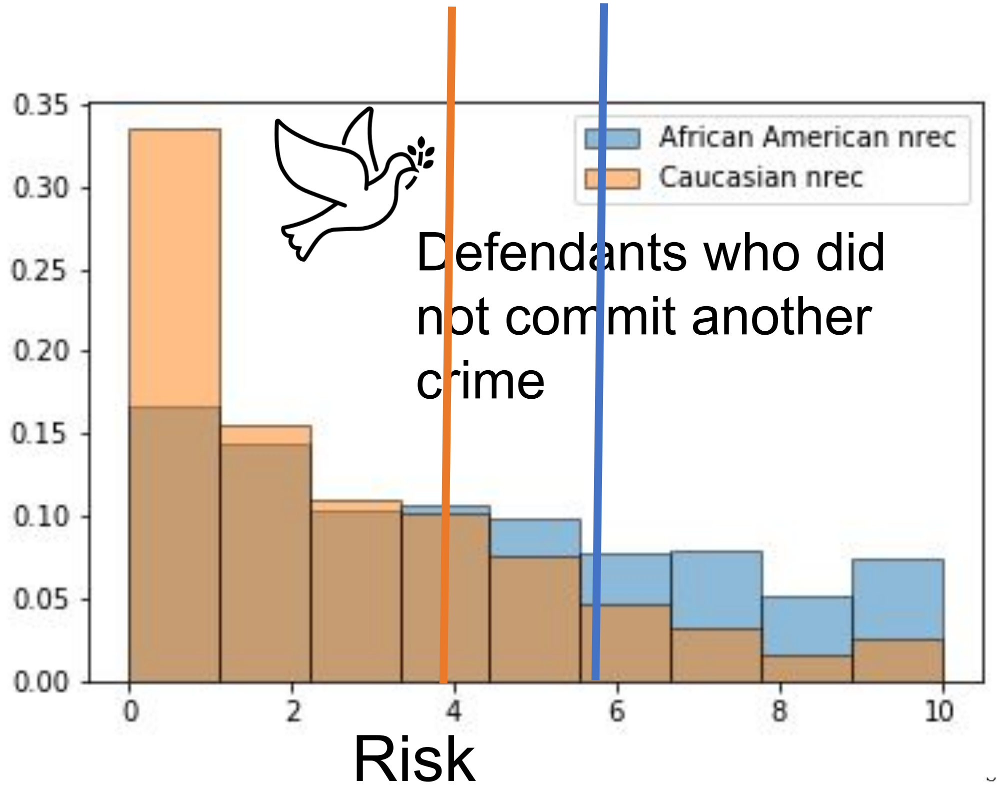
Algorithm falsely flagged black defendants as future criminals at 2X the rate for white defendants

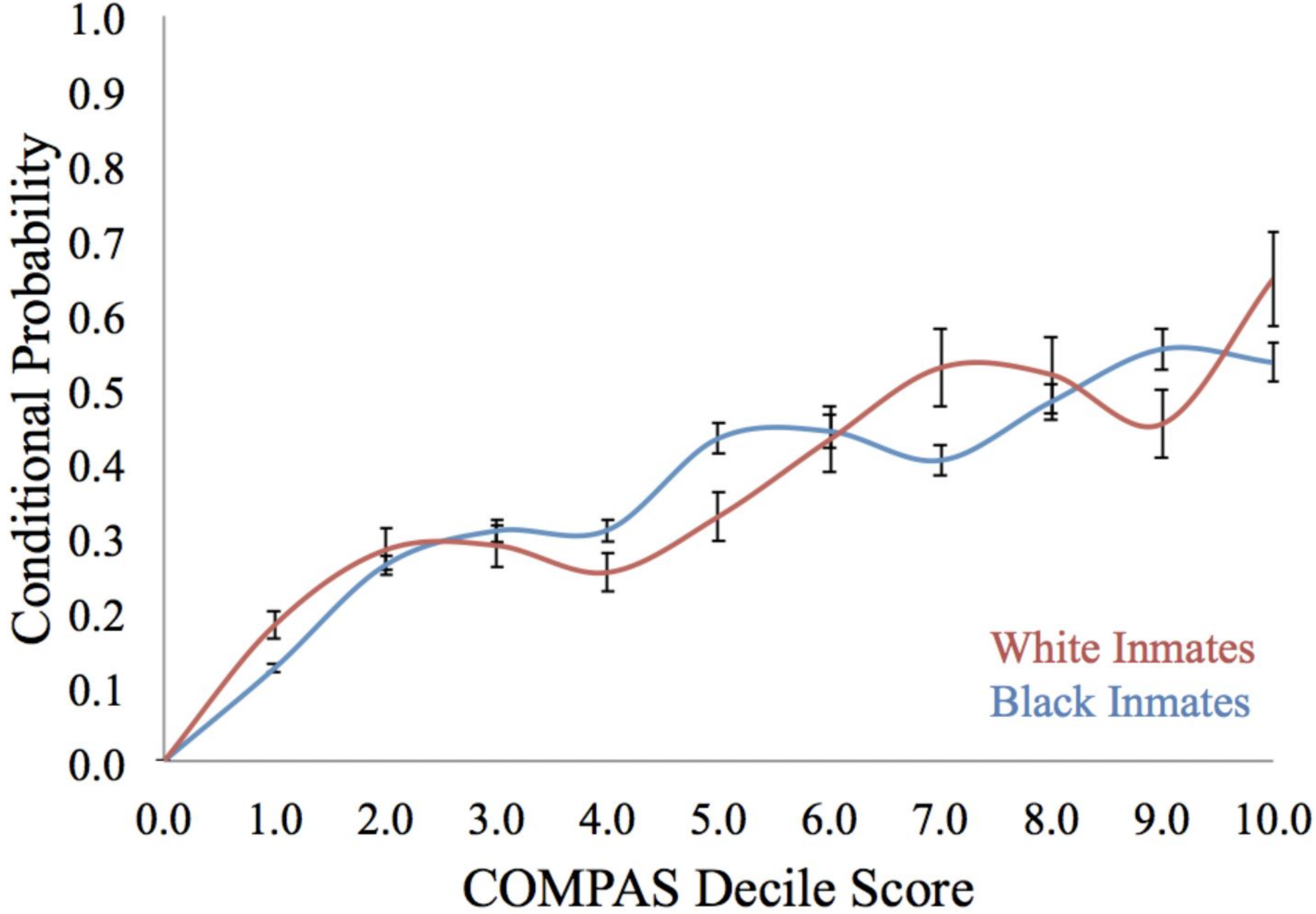
False Negative rate high for whites

White defendants were mislabeled as low risk more often than black defendants.

This disparity cannot be explained by prior crimes, the type of crimes, their age and gender


Defendants who committed another crime


Defendants who did not commit another crime



Wadsworth et al. 2018

Quantifying fairness

Quantifying fairness

- **AI fairness** focuses on ensuring that AI systems do not produce discriminatory or biased outcomes, especially when the systems affect different demographic groups.
- Metrics for AI fairness quantify fairness and help assess the impact of an AI model on different groups.

Notation

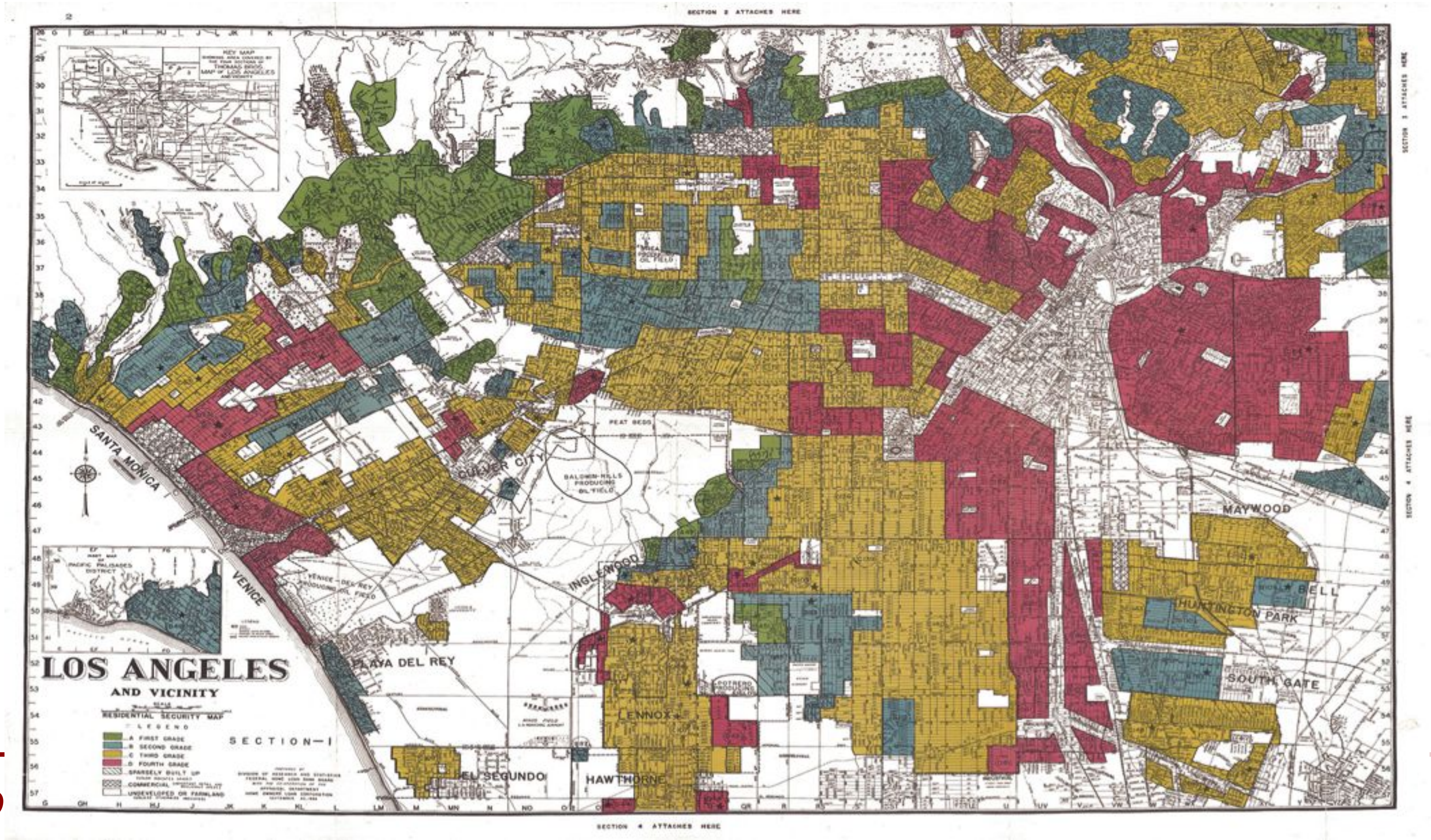
- A – set of protected attributes
- X – all other observable attributes
- U – set of latent attributes not observed
- Y – outcome to be predicted
- Ŷ – predictor, dependent on A, X, U.

Take 1: Fairness through Unawareness

- *An algorithm is fair so long as any* ***protected*** *attributes A are not explicitly used in the decision-making process.*
  - Any mapping Ŷ : X → Y that excludes A satisfies this (e.g., do not use race to evaluate loans)
  - However, it can be biased for many reasons.

- *A* – set of protected attributes
- *X* – all other observable attributes
- *U* – set of latent attributes not observed
- *Y* – outcome to be predicted
- Ŷ – predictor, dependent on A, X, U.

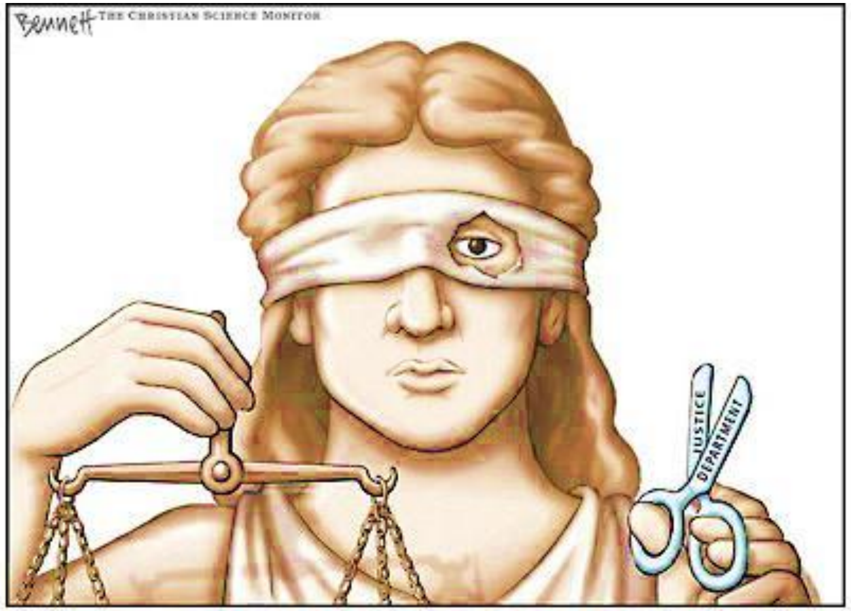
“Redlining” – discrimination by ~~race~~ zipcode



Lesson: Fairness is *task-specific*

Fairness requires understanding of classification task and protected groups

“Awareness”



Fairness through Privacy?

“It's Not Privacy, and It's Not Fair”

Cynthia Dwork & Deirdre K. Mulligan. Stanford Law Review.

Privacy is no Panacea: Can’t hope to have privacy solve our fairness problems.

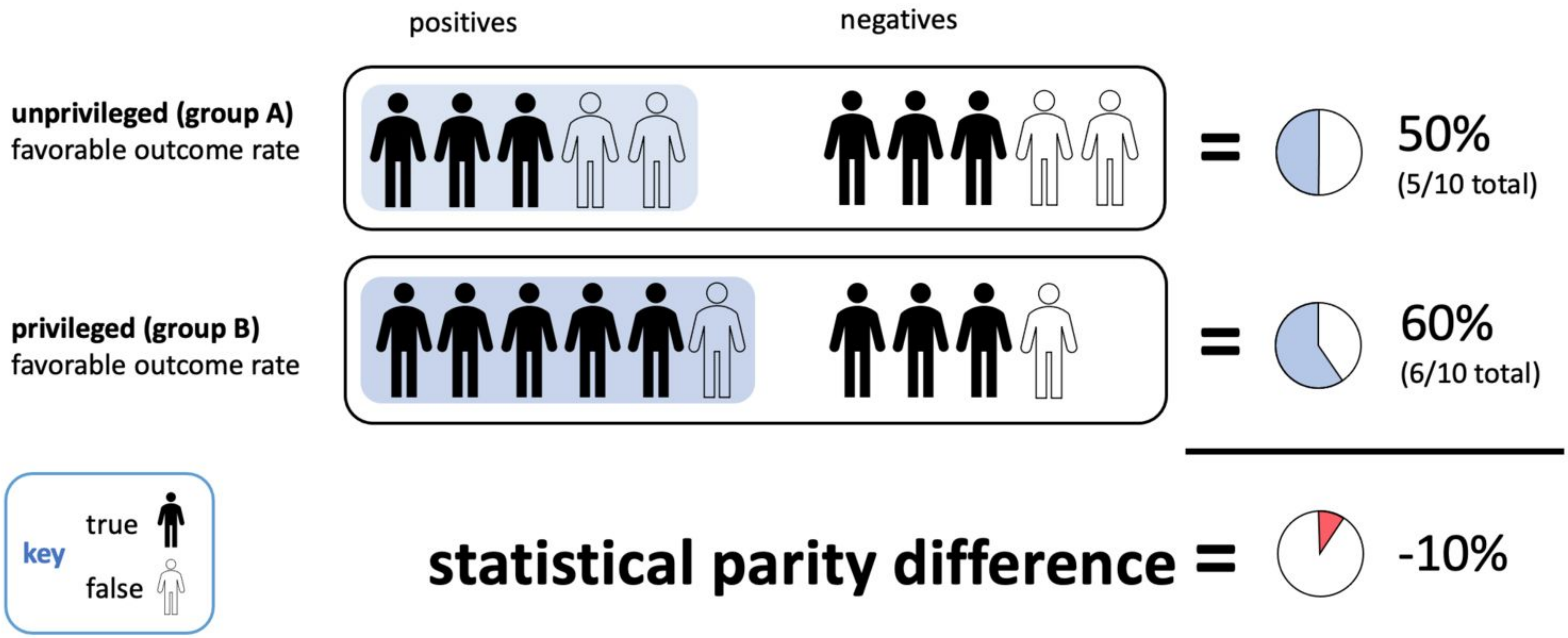
“At worst, **privacy solutions can hinder efforts to identify classifications that unintentionally produce objectionable outcomes**—for example, differential treatment that tracks race or gender—by limiting the availability of data about such attributes.“

Statistical Parity (Group fairness)

- Ensures that the positive outcome rate is the same across groups (e.g., gender, race)

$$P(\hat{Y} = 1 | A = a_1) = P(\hat{Y} = 1 | A = a_2)$$

- E.g., Fraction of people in group a1 getting credit is the same as in group a2.



positives

negatives

unprivileged (group A)
favorable outcome rate = 50% (5/10 total)

privileged (group B)
favorable outcome rate = 60% (6/10 total)

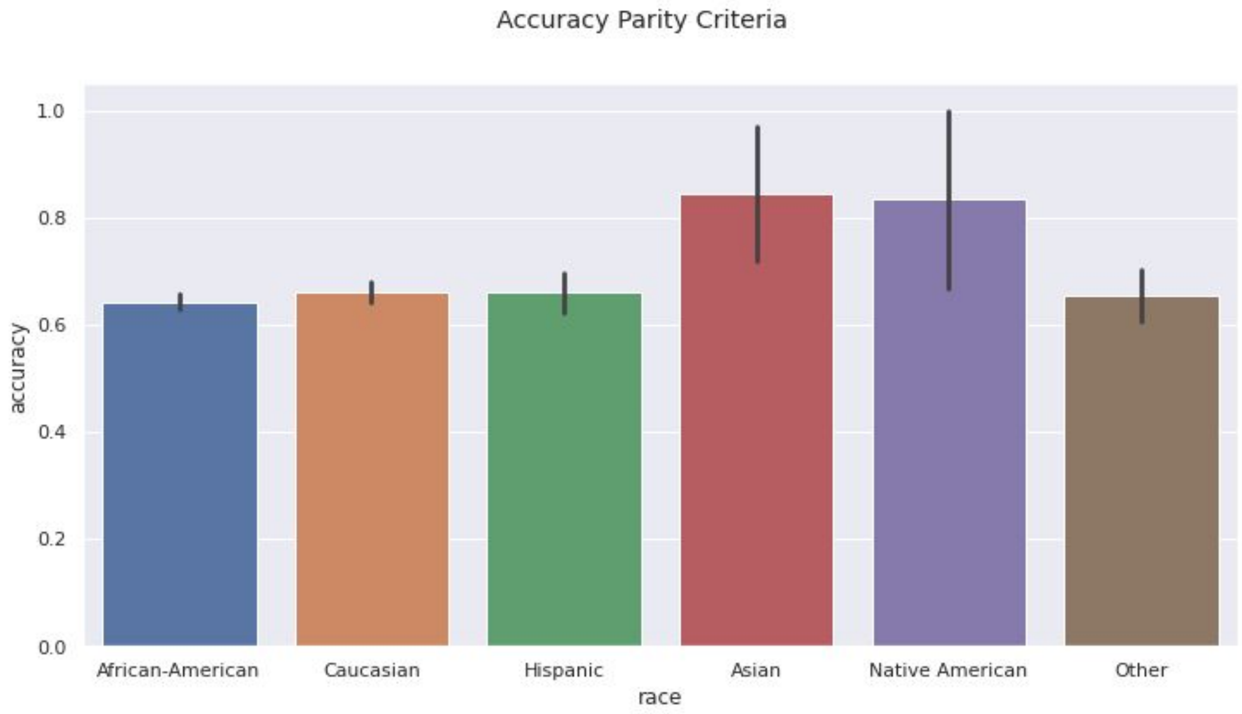
key
true
false

statistical parity difference = -10%

Group Level Graphic Credit: https://www.xyonix.com/blog/how-to-detect-and-mitigate-harmful-societal-bias-in-your-organizations-ai

Accuracy parity

- The rates of accurate predictions are the same among groups, within the standard of error.
  - Northpointe used this criteria when developing the COMPAS model. The model arrives at the *right* prediction 60% of the time.



Predictive Parity

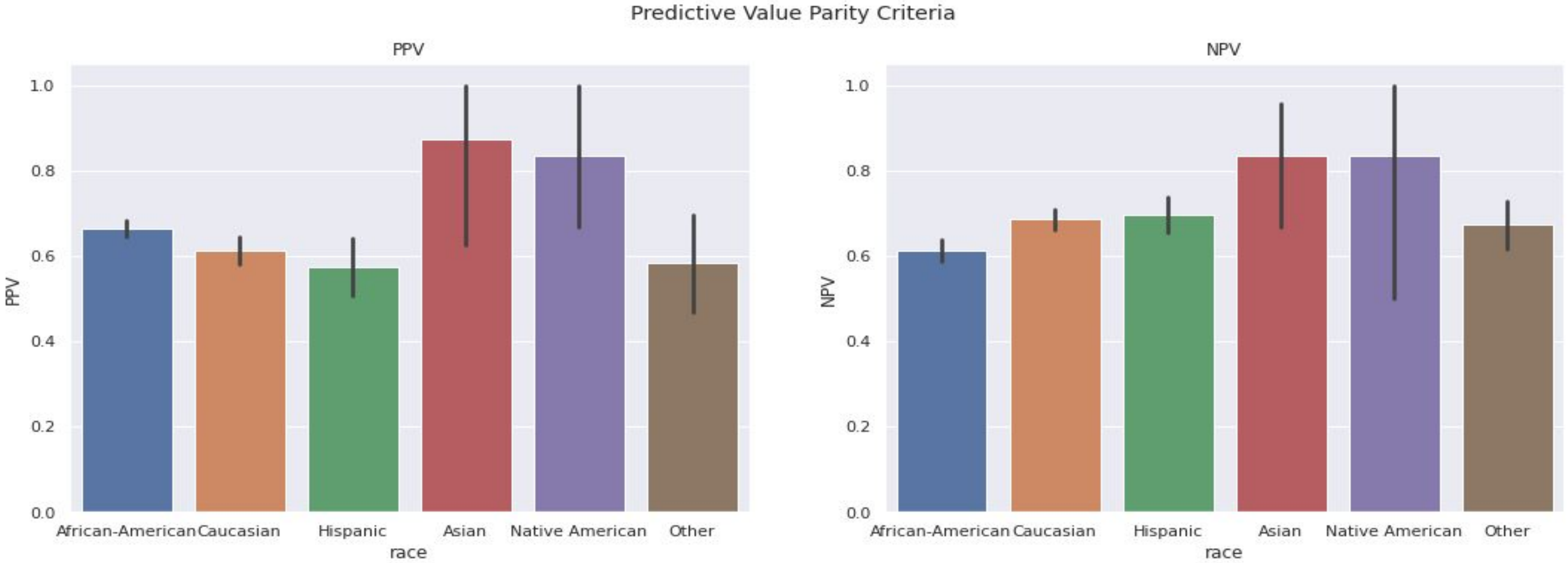
Ensures that the precision (positive predictive value) is the same across groups.

$$P(Y=1|\hat{Y}=1, A=a_1) = P(Y=1|\hat{Y}=1, A=a_2)$$

Similar to Equality of Opportunity. Just change Y and Y^.

- i.e., conditioned on predictions, outcomes should be similar across groups
- e.g., In a loan approval system, the proportion of approved loans that default should be the same for different income groups.

Predictive Parity. COMPAS largely maintains consistent rates of recidivism across groups



Among those labeled high risk by COMPAS, approximately 60% re-offended

Among those labeled low risk by COMPAS, approximately 60% did not re-offend.

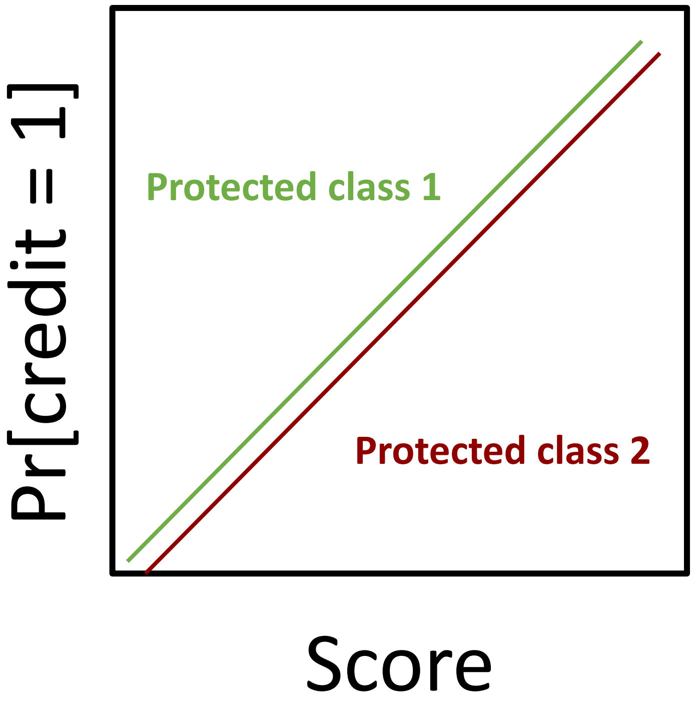
https://afraenkel.github.io/fairness-book/content/08-compas-2.html

Pretty fair (by this definition)

Conditional Statistical Parity

- $P(d = 1 \mid L = l, A = m) = P(d = 1 \mid L = l, A = f)$
- *Difference:* Considering these factors, protected and unprotected instances should have the same probability of success.
- L must be legitimate.
- Imagine the credit example, legitimate:
  - Credit history, employment, amount requested.

Ideal statistical parity



Equalized Odds

- Definition: $\hat{Y} \perp A | Y$

在已知真实结果 *Y* 的情况下，预测结果 *Y*^不应受到受保护属性 *A*（比如种族、性别等）的影响。

- Prediction does not provide information about *A* beyond what Y already does.
- $P(\hat{Y} | Y = y, A = m) = P(\hat{Y} | Y = y, A = f)$ Y can be 1or 0. meaning should be true both cases
- Protects against accuracy disparity.

- *A* – set of protected attributes
- *X* – all other observable attributes
- *U* – set of latent attributes not observed
- *Y* – outcome to be predicted
- $\hat{Y}$ – predictor, dependent on A, X, U.

Equalized Odds

Ensures that the model's prediction is equally accurate across groups.

- True positive rate (TPR) and false positive rate (FPR) should be equal across groups

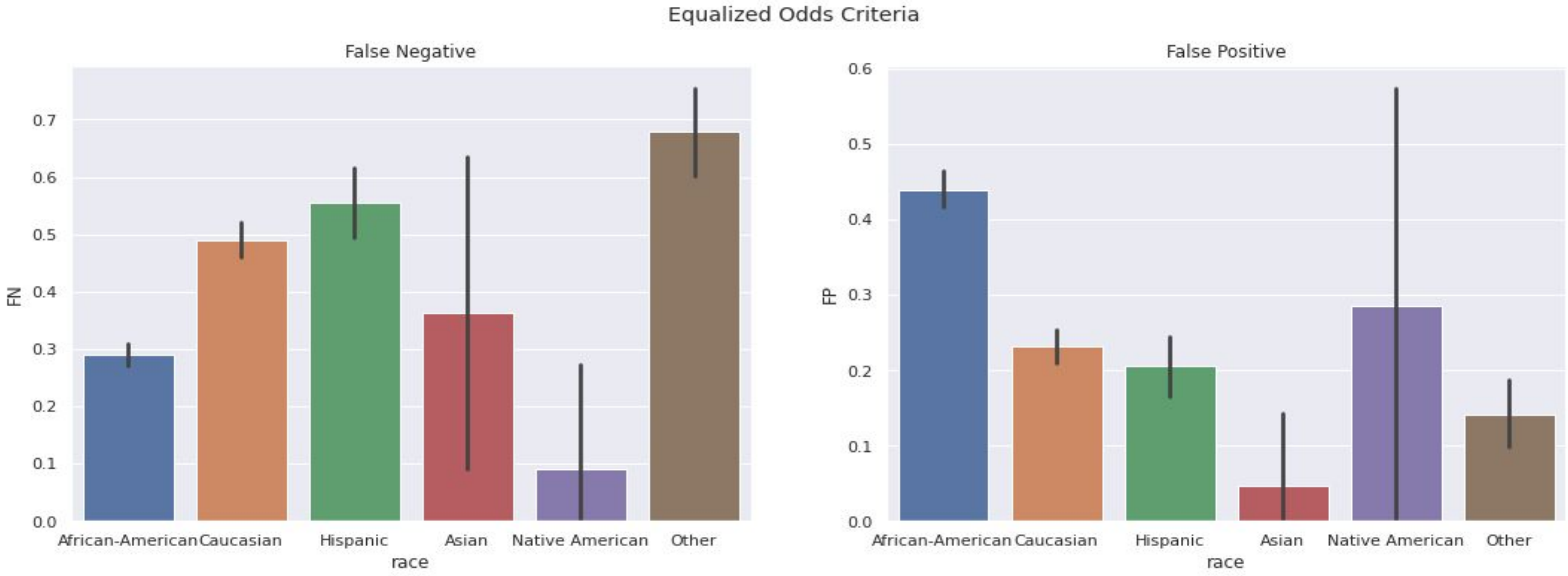
$$P(\hat{Y} = 1 | Y = 1, A = a_1) = P(\hat{Y} = 1 | Y = 1, A = a_2)$$

$$P(\hat{Y} = 1 | Y = 0, A = a_1) = P(\hat{Y} = 1 | Y = 0, A = a_2)$$

- Prediction does not provide information about *A* beyond what Y already does
- Can conflict with other fairness metrics like demographic parity

Example from COMPAS

The False Positive and False Negative rates vary significantly across groups and Black defendants experience consistently unfair treatment as compared to other defendants.



https://afraenkel.github.io/fairness-book/content/08-compas-2.html

COMPAS fails to achieve fairness (by this definition)

Equality of Opportunity

Ensures that the recall (true positive rate) is equal across groups.

$$P(\hat{Y}=1|Y=1,A=a_1)=P(\hat{Y}=1|Y=1,A=a_2)$$

- i.e., conditioned on outcomes, predictions should be the same across groups.
- In a binary case, we think of Y=1 as the "advantaged" outcome. Require non-discrimination only within this outcome.
- E.g., people who pay back their loan ought to have an equal opportunity of getting the loan in the first place.

Calibration

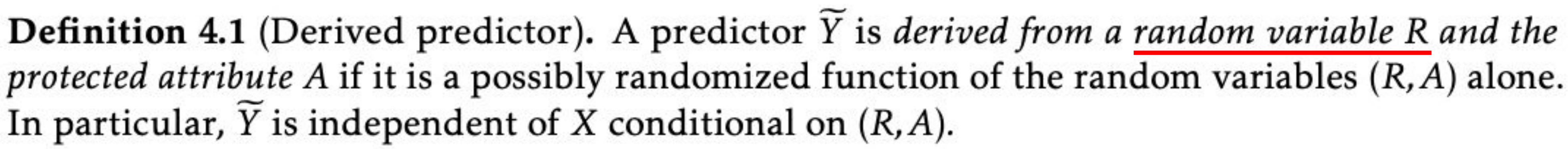
Ensures that predicted probabilities are equally reliable across groups.

$$P(Y=1|\hat{P}, A=a_1) = P(Y=1|\hat{P}, A=a_2)$$

- In a recidivism prediction model, if the predicted probability of reoffending is 70%, this should correspond to a 70% actual rate for all racial groups.

Achieving Eq. Odds and Eq. of Opportunity

- Goal: find an Eq. odds or Eq. Opportunity predictor $\tilde{Y}$
  - Derived from a (*possibly discriminatory*) predictor, $\hat{Y}$



Definition 4.1 (Derived predictor). A predictor $\widetilde{Y}$ is *derived from a random variable R and the protected attribute A* if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\widetilde{Y}$ is independent of X conditional on (R, A).

- The joint distribution is required at training time.
- At prediction time, we only have R, A.

- *A* – set of protected attributes
- *X* – all other observable attributes
- *U* – set of latent attributes not observed
- *Y* – outcome to be predicted
- $\hat{Y}$ – predictor, dependent on A, X, U.

Individual Fairness

With many slides from Moritz Hardt - Fall 2017

Disparate Impact

Measures whether outcomes disproportionately affect one group

$$\text{Disparate Impact Ratio} = \frac{P(\hat{Y} = 1 | A = a_1)}{P(\hat{Y} = 1 | A = a_2)}$$

- In algorithmic hiring, if the hiring rate for one gender is less than 80% of the rate for another gender, it may be considered discriminatory.

Credit Application

User visits `capitalone.com`

Capital One uses tracking information provided by the tracking network to personalize offers

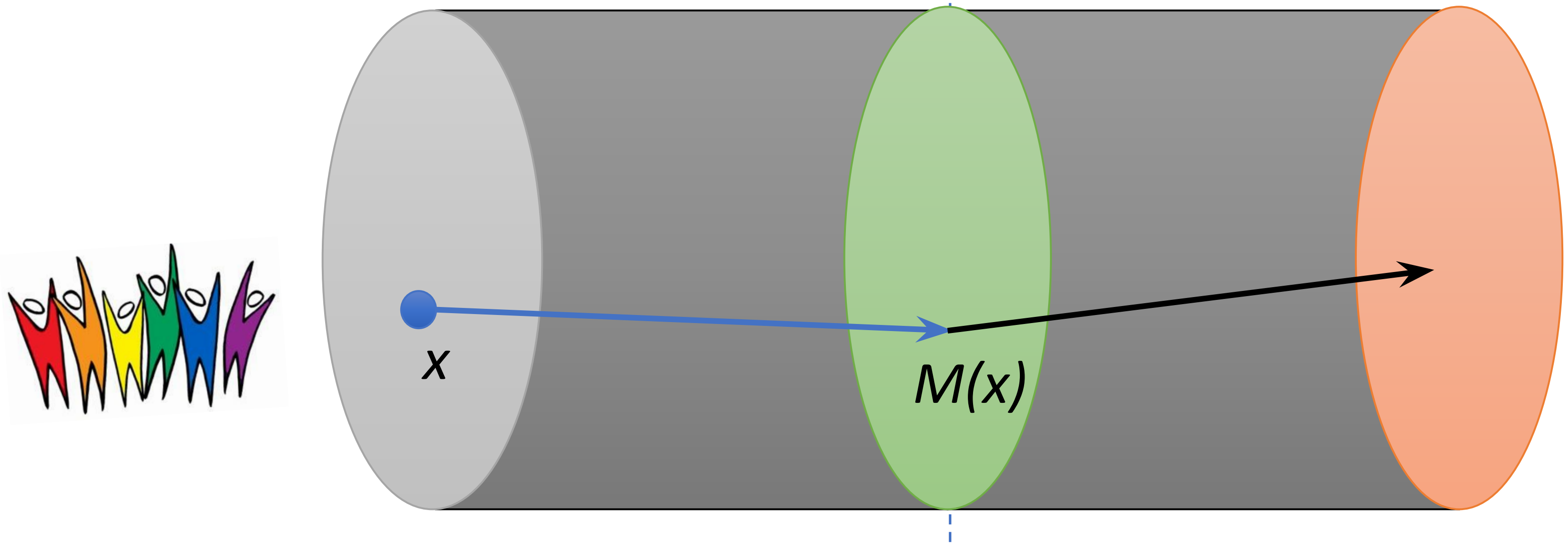
Concern: *Steering* minorities into higher rates (illegal)

Classifier
(eg. ad network)

$M : V \rightarrow O$

Vendor
(eg. capital one)

$f : O \rightarrow A$



V: Individuals

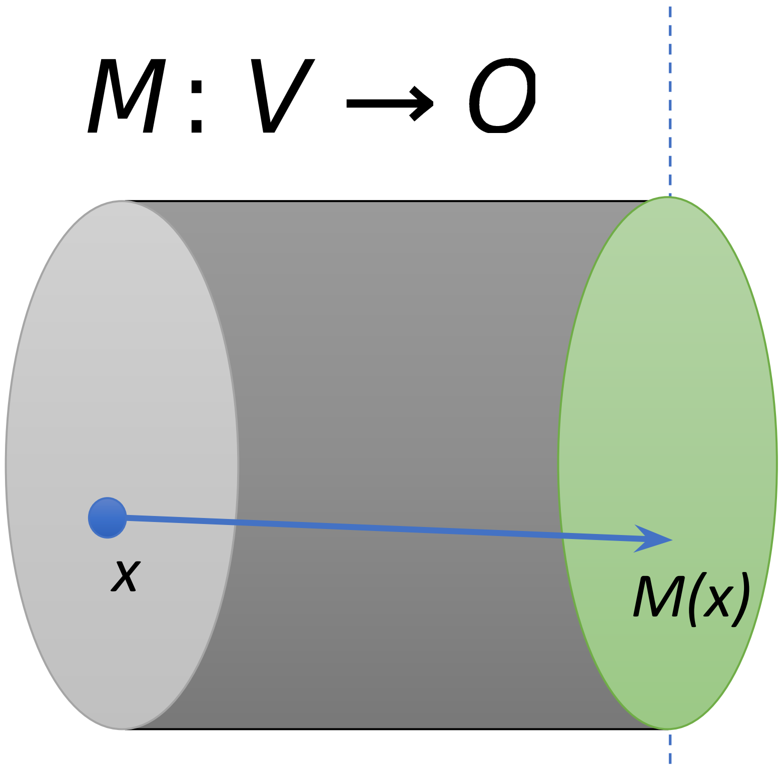
O: outcomes

A: actions

Goal:

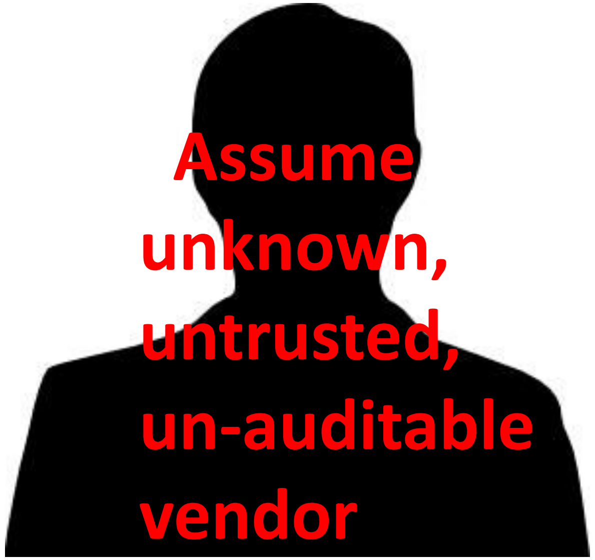
Achieve Fairness in the classification step

$$M : V \rightarrow O$$



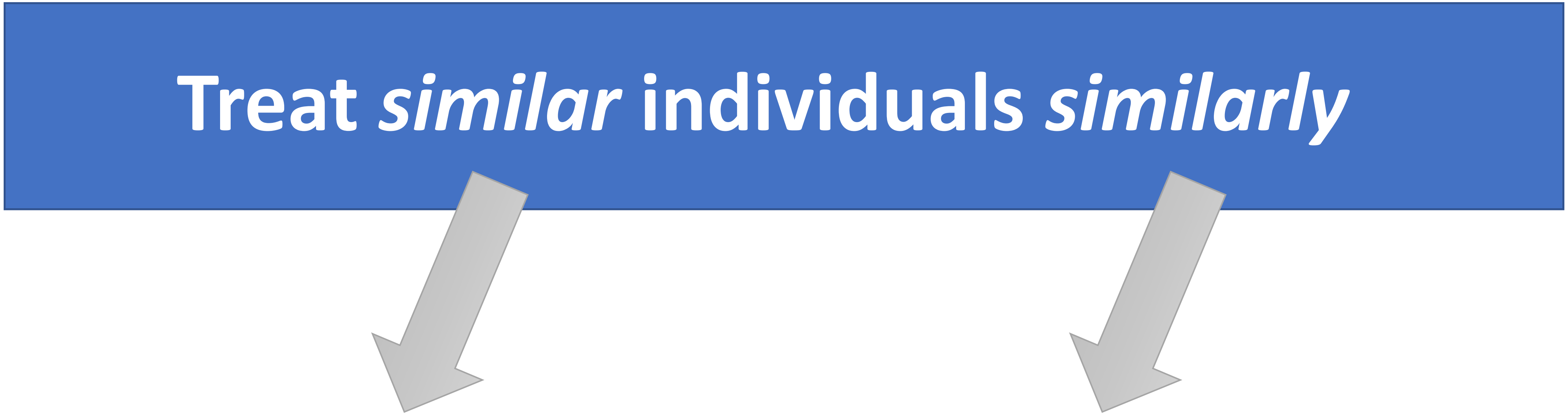
V: Individuals

O: outcomes



Assume unknown, untrusted, un-auditable vendor

Individual Fairness



Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

If x_1 and x_2 are similar, then $\hat{Y}(x_1) \approx \hat{Y}(x_2)$.

Metric

- Assume ***task-specific similarity metric***
  - Extent to which two individuals are similar w.r.t. the classification task at hand
- Ideally captures *ground truth*
  - Or, society's best approximation
- Open to public discussion, refinement
  - In the spirit of Rawls
- Typically, does not suggest classification!

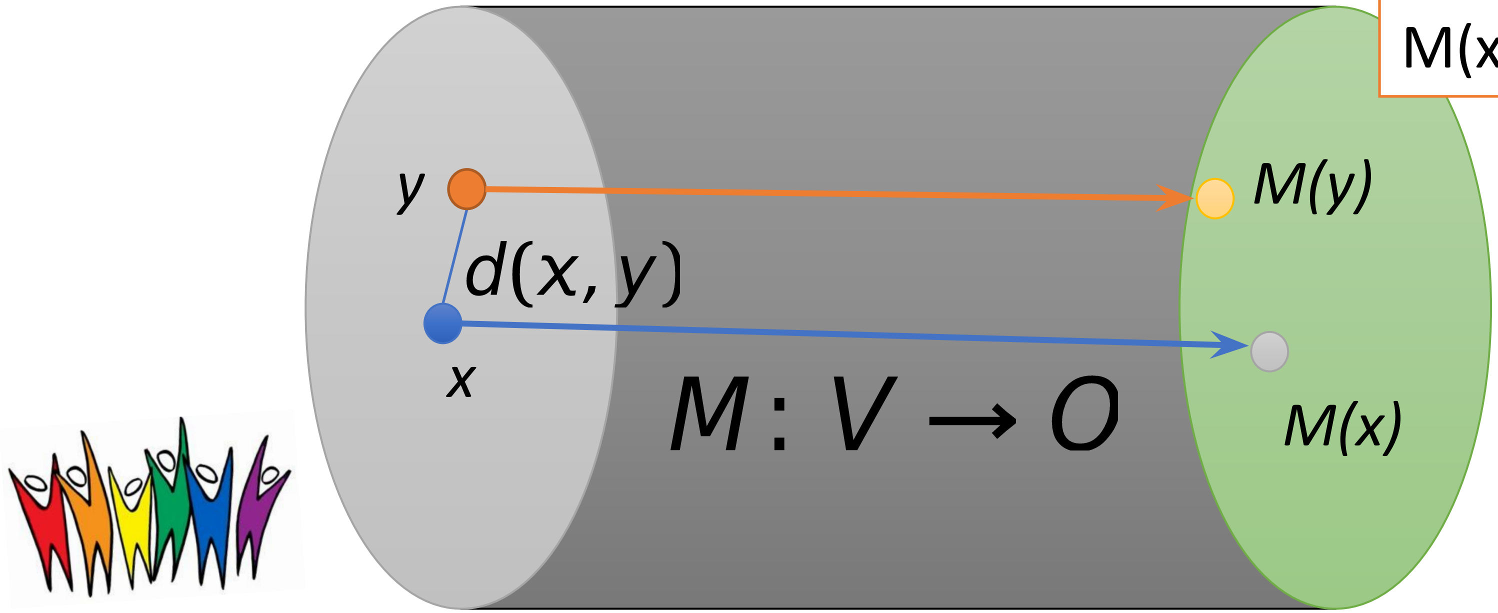
Examples

- Financial/insurance risk metrics
  - Already widely used (though secret)
- In a job recommendation system, two applicants with similar qualifications should receive similar recommendation.
- AALIM health care metric
  - health metric for treating similar patients similarly
- Roemer’s relative effort metric
  - Well-known approach in Economics/Political theory

How to formalize this?

Think of V as space
with metric d(x,y)
similar = small d(x,y)

How can we compare M(x) with M(y)?

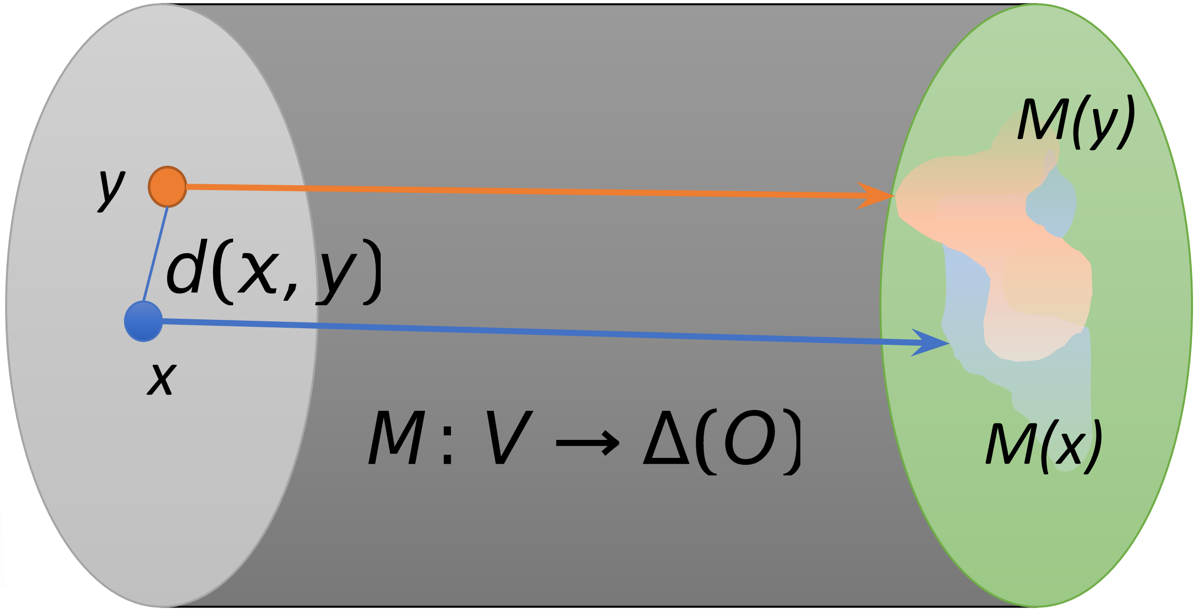


V: Individuals

O: outcomes

Distributional outcomes

How can we compare M(x) with M(y)?



Statistical distance!

V: Individuals

O: outcomes

Individual vs Group fairness

- Statistical metrics focus on group-level outcomes
- Statistical metrics can hide significant variations within groups
- Two individuals who are identical except for their protected attributes might receive different treatments
- Example: A system could achieve statistical parity by discriminating against half of each protected group while favoring the other half
  - The average looks fair, but individual treatment is arbitrary

When does Individual Fairness imply Group Fairness?

Suppose we enforce a metric *d.*

Question: Which *groups of individuals* receive (approximately) equal outcomes?

Theorem:
Answer is given by **Earthmover distance** (w.r.t. *d*) between the two groups.

Connection to differential privacy

- Close connection between individual fairness and **differential privacy** [Dwork-McSherry-Nissim-Smith’06]

  DP: Fairness condition on set of databases

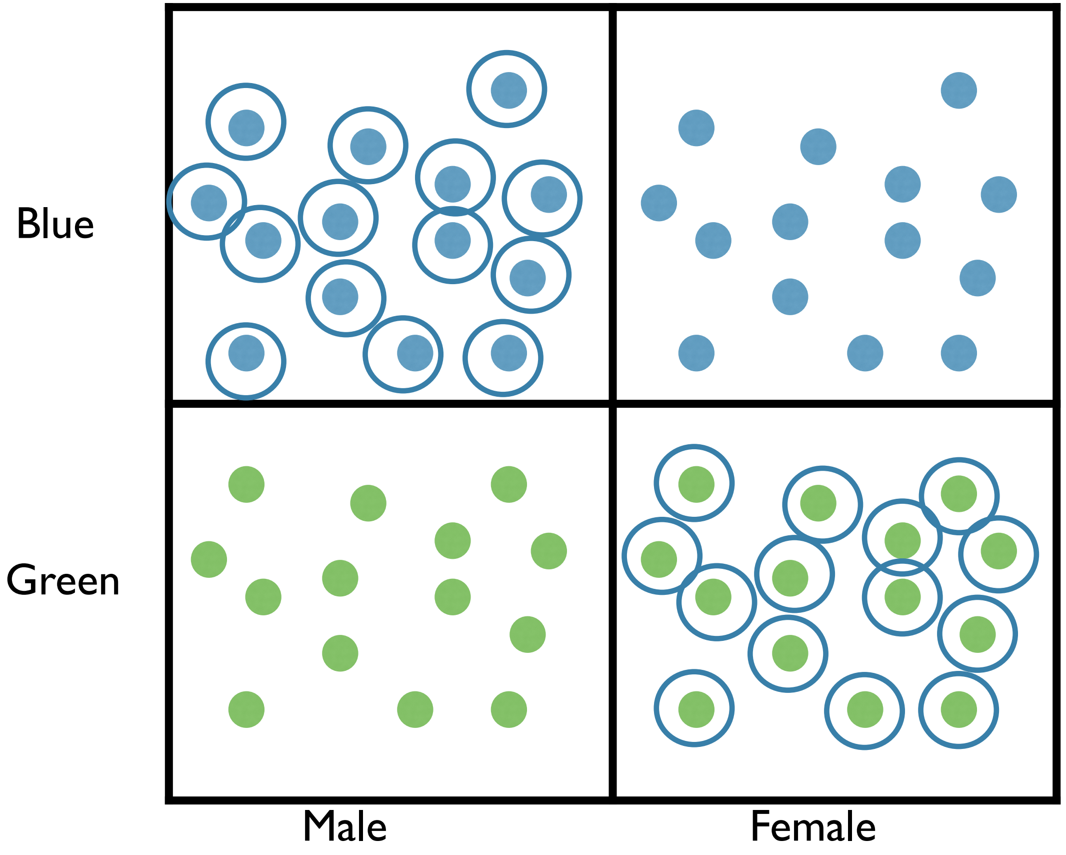
  IF: Fairness condition on set of individuals

	Differential Privacy	Individual Fairness
Objects	Databases	Individuals
Outcomes	Output of statistical analysis	Classification outcome
Similarity	General purpose metric	Task-specific metric

Subgroup Fairness

Weaknesses of statistical fairness notions

- Statistical fairness metrics typically look at one protected attribute at a time
- Toy example: Protected subgroups are "Men", "Women", "Blue", "Green". Labels are independent of protected attributes.
- The following allocation achieves statistical parity and equalized odds. Satisfying fairness for "Men vs Women" and "Blue vs Green" separately doesn't guarantee fairness for intersectional groups (e.g., Blue Women)
- The same allocation could be deeply unfair to specific intersectional subgroups while still maintaining overall statistical parity

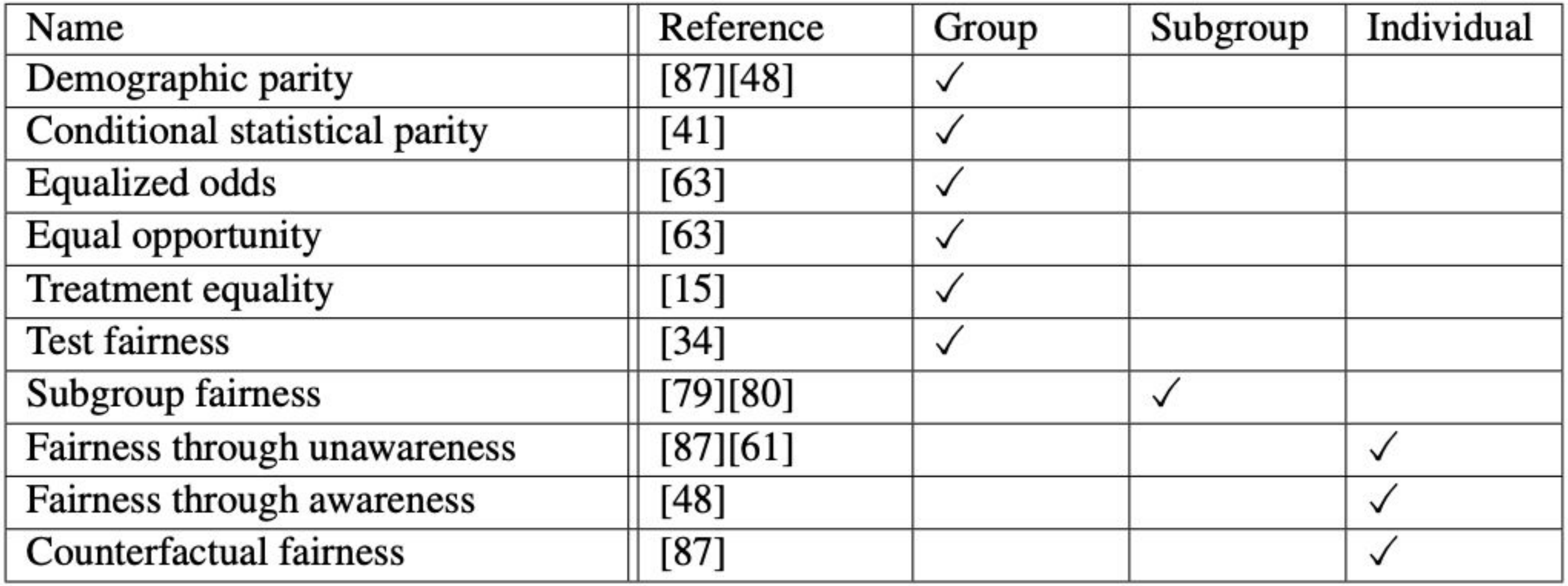


Statistical Fairness Notions

- The problem: Statistical constraints averaged over coarse subgroups can be "gerrymandered" to pack unfairness into structured subgroups. 不公正的划分
  - No reason to expect it won't happen with standard optimization techniques: we will see it does.
- Just add "green men", "blue women", etc. as protected subgroups?
  - What about other groups?

Recap: Fairness Definitions

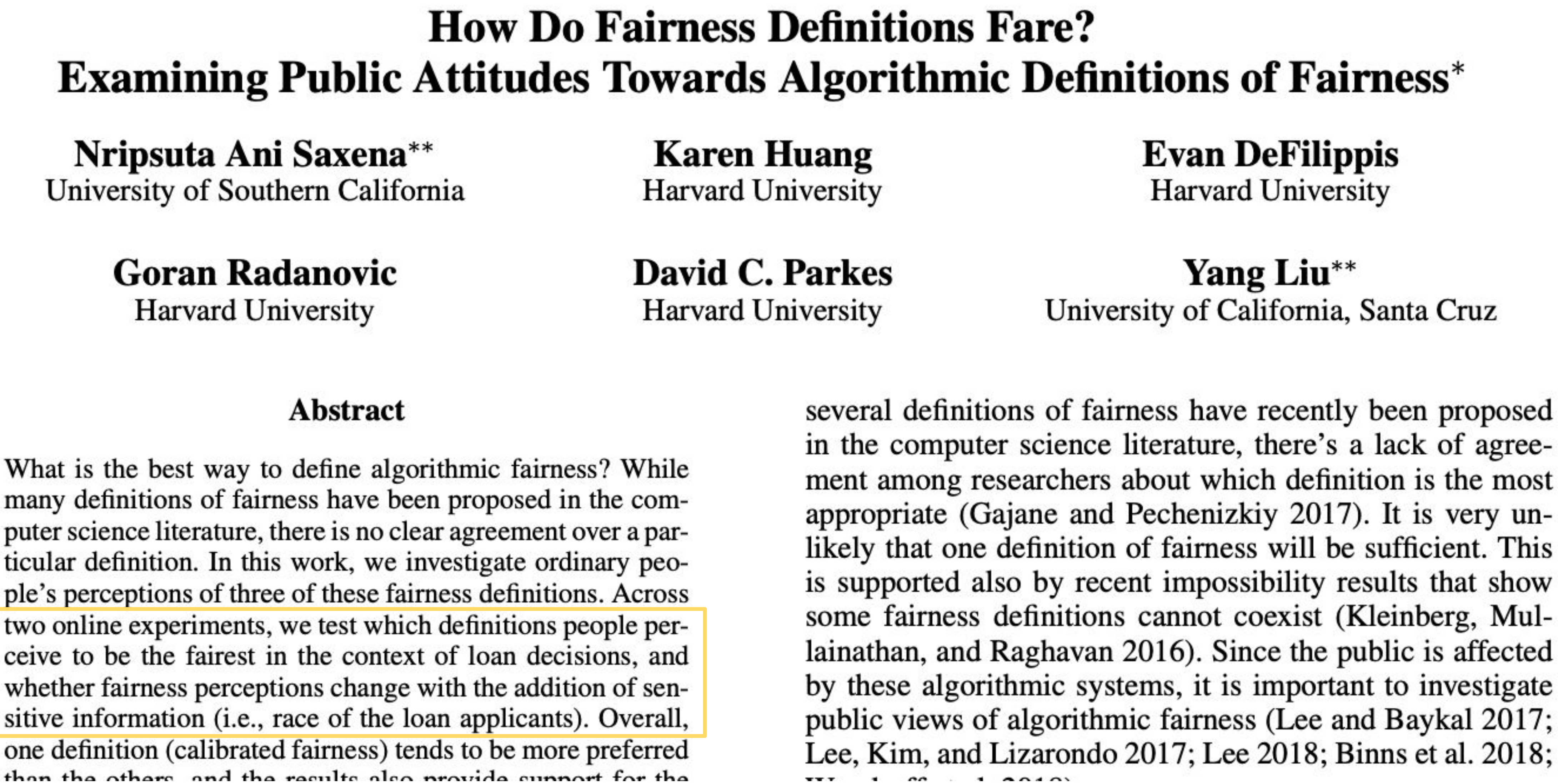
https://arxiv.org/pdf/1908.09635.pdf
Read Section 4.1



Name	Reference	Group	Subgroup	Individual
Demographic parity	[87][48]	✓		
Conditional statistical parity	[41]	✓		
Equalized odds	[63]	✓		
Equal opportunity	[63]	✓		
Treatment equality	[15]	✓		
Test fairness	[34]	✓		
Subgroup fairness	[79][80]		✓	
Fairness through unawareness	[87][61]			✓
Fairness through awareness	[48]			✓
Counterfactual fairness	[87]			✓

Table 1. Categorizing different fairness notions into group, subgroup, and individual types.

Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54.6 (2021): 1-35.



How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness[*]

Nripsuta Ani Saxena[**]
University of Southern California

Karen Huang
Harvard University

Evan DeFilippis
Harvard University

Goran Radanovic
Harvard University

David C. Parkes
Harvard University

Yang Liu[**]
University of California, Santa Cruz

Abstract

What is the best way to define algorithmic fairness? While many definitions of fairness have been proposed in the computer science literature, there is no clear agreement over a particular definition. In this work, we investigate ordinary people's perceptions of three of these fairness definitions. Across two online experiments, we test which definitions people perceive to be the fairest in the context of loan decisions, and whether fairness perceptions change with the addition of sensitive information (i.e., race of the loan applicants). Overall, one definition (calibrated fairness) tends to be more preferred than the others, and the results also provide support for the

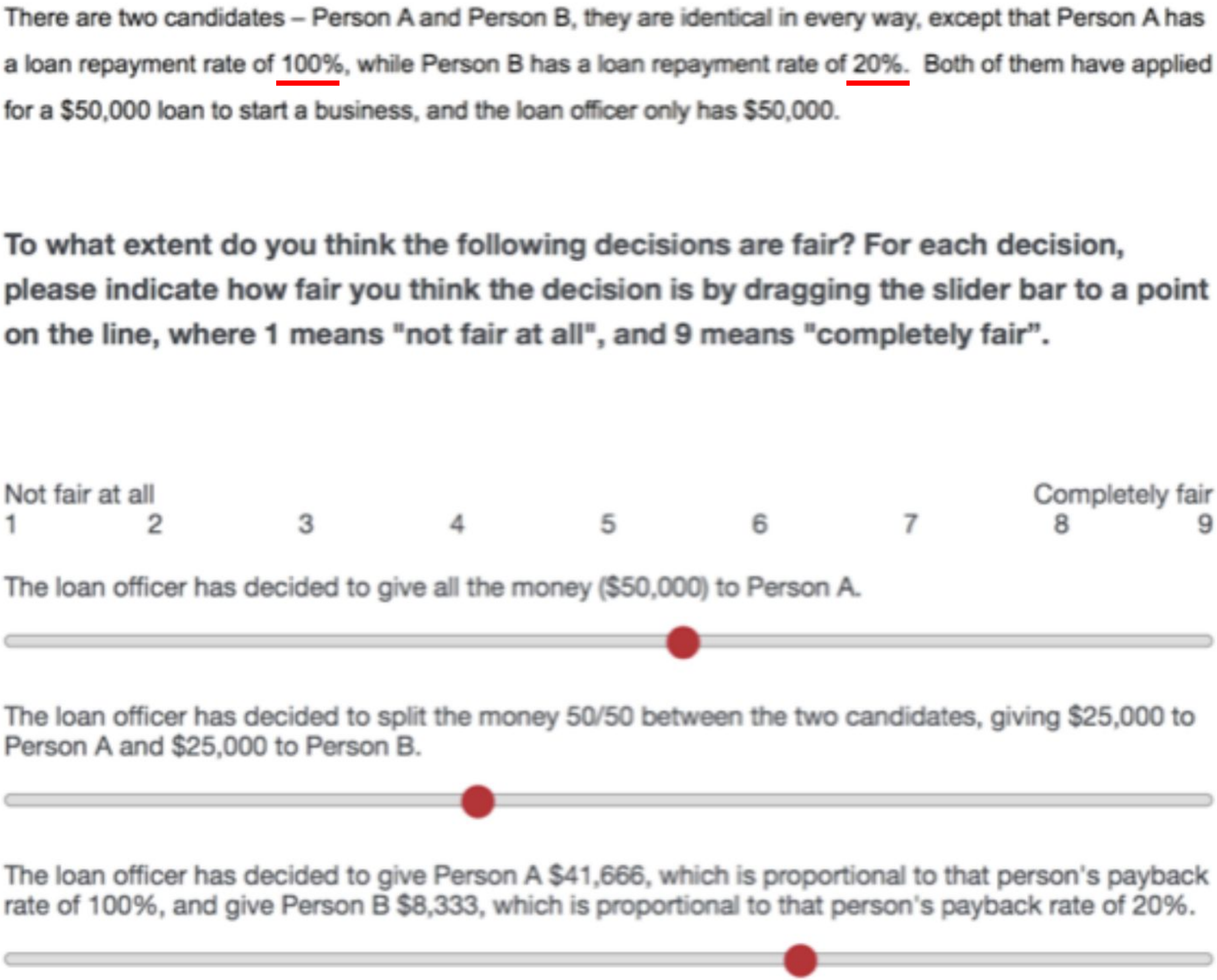
several definitions of fairness have recently been proposed in the computer science literature, there's a lack of agreement among researchers about which definition is the most appropriate (Gajane and Pechenizkiy 2017). It is very unlikely that one definition of fairness will be sufficient. This is supported also by recent impossibility results that show some fairness definitions cannot coexist (Kleinberg, Mullainathan, and Raghavan 2016). Since the public is affected by these algorithmic systems, it is important to investigate public views of algorithmic fairness (Lee and Baykal 2017; Lee, Kim, and Lizarondo 2017; Lee 2018; Binns et al. 2018; Woodruff et al. 2018).

- Three definitions of algorithmic fairness:
  - Treating similar individuals similarly (based on task-relevant metrics)
  - Never favoring worse individuals over better ones (meritocratic fairness)
  - Calibrated fairness (allocating resources in proportion to merit)

Two online experiments using loan allocation scenarios

- Asked participants to evaluate three possible loan allocation decisions:
  - "All A": Give all money to candidate with higher repayment rate
  - "Equal": Split money 50/50 between candidates
  - "Ratio": Split money proportionally based on repayment rates

Setup - FTU



There are two candidates – Person A and Person B, they are identical in every way, except that Person A has a loan repayment rate of 100%, while Person B has a loan repayment rate of 20%. Both of them have applied for a $50,000 loan to start a business, and the loan officer only has $50,000.

To what extent do you think the following decisions are fair? For each decision, please indicate how fair you think the decision is by dragging the slider bar to a point on the line, where 1 means "not fair at all", and 9 means "completely fair".

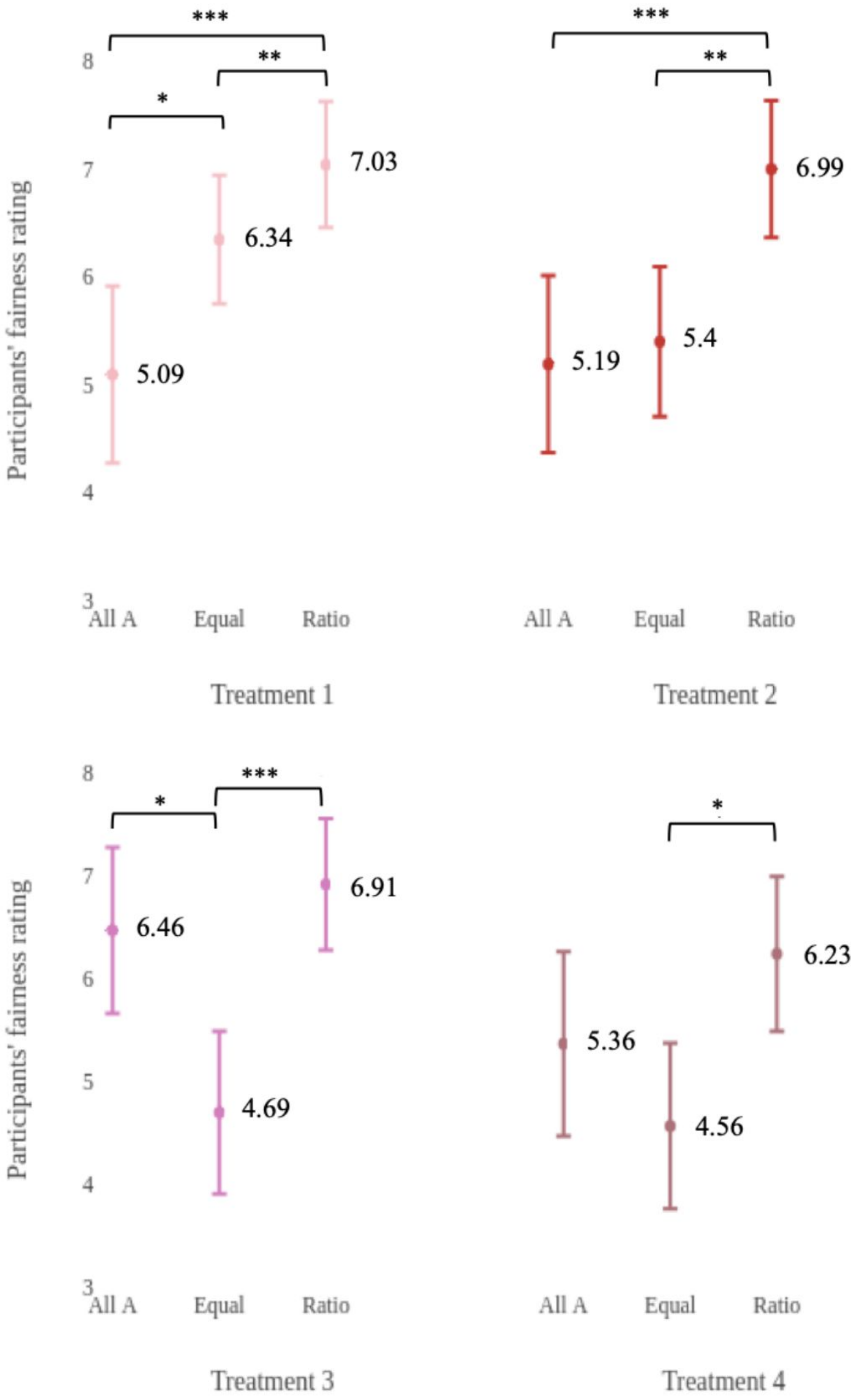
The loan officer has decided to give all the money ($50,000) to Person A.

The loan officer has decided to split the money 50/50 between the two candidates, giving $25,000 to Person A and $25,000 to Person B.

The loan officer has decided to give Person A $41,666, which is proportional to that person's payback rate of 100%, and give Person B $8,333, which is proportional to that person's payback rate of 20%.

Treatments:

1. **55/50**
2. **70/40**
3. **90/10**
4. **100/20**

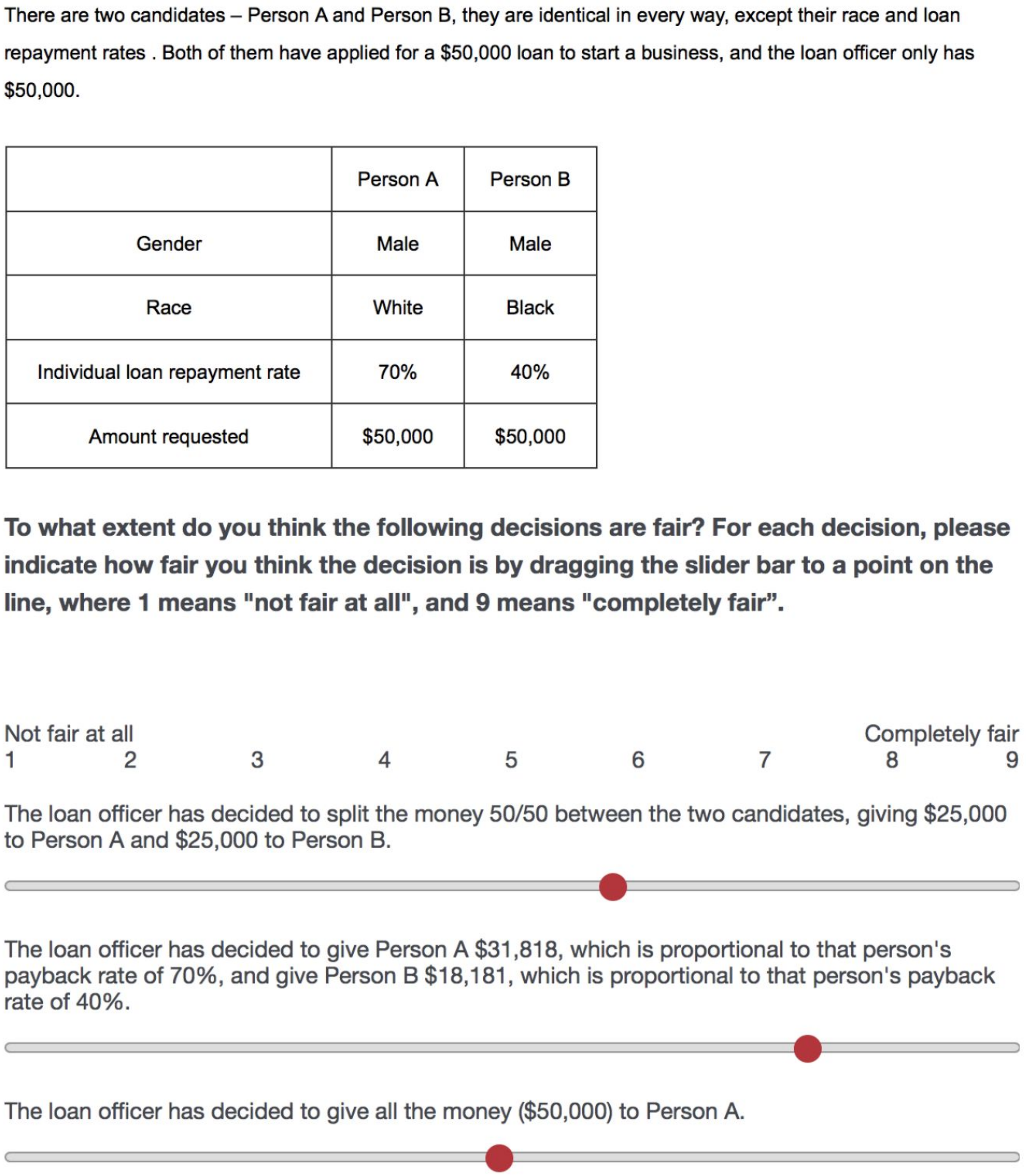

Participants' fairness rating
8
7
6
5
4
3
***
**
*
5.09
6.34
7.03
All A
Equal
Ratio
Treatment 1
***
**
5.19
5.4
6.99
All A
Equal
Ratio
Treatment 2
Participants' fairness rating
8
7
6
5
4
3
*
***
6.46
4.69
6.91
All A
Equal
Ratio
Treatment 3
*
5.36
4.56
6.23
All A
Equal
Ratio
Treatment 4

Task 2

Treatments:

1. **55/50**
2. **70/40**
3. **90/10**
4. **100/20**

Varied Race



There are two candidates – Person A and Person B, they are identical in every way, except their race and loan repayment rates . Both of them have applied for a $50,000 loan to start a business, and the loan officer only has $50,000.

	Person A	Person B
Gender	Male	Male
Race	White	Black
Individual loan repayment rate	70%	40%
Amount requested	$50,000	$50,000

To what extent do you think the following decisions are fair? For each decision, please indicate how fair you think the decision is by dragging the slider bar to a point on the line, where 1 means "not fair at all", and 9 means "completely fair”.

Not fair at all Completely fair
1 2 3 4 5 6 7 8 9

The loan officer has decided to split the money 50/50 between the two candidates, giving $25,000 to Person A and $25,000 to Person B.

The loan officer has decided to give Person A $31,818, which is proportional to that person's payback rate of 70%, and give Person B $18,181, which is proportional to that person's payback rate of 40%.

The loan officer has decided to give all the money ($50,000) to Person A.

Findings

Study 1 (No Race Information)

- People generally preferred the "Ratio" decision (calibrated fairness)
- When candidates had very similar repayment rates, people preferred equal division
- Support for giving everything to the better candidate increased as the difference in repayment rates grew larger

Study 2 (With Race Information)

- Still preferred the "Ratio" decision
- Race influenced decisions when differences in repayment rates were large
- People were more likely to support giving all money to the better candidate when that person was Black
- Suggests support for affirmative action principles

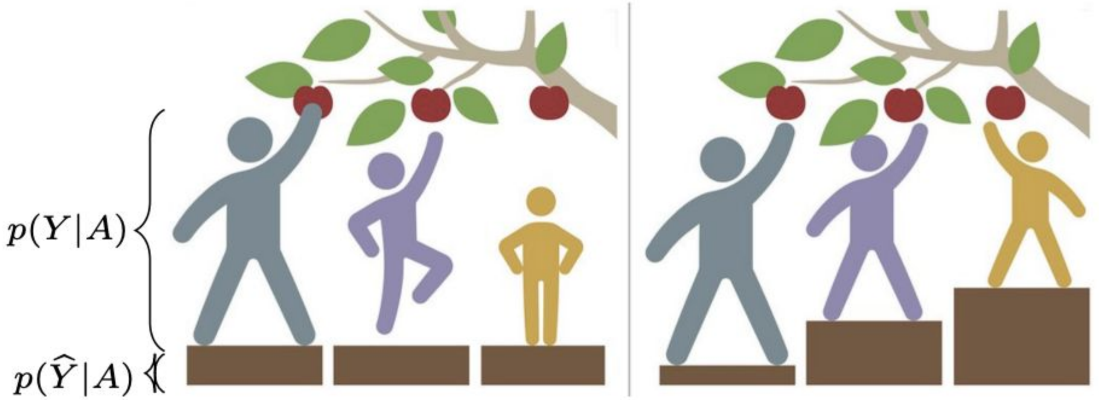
平权行动

Implications

- Public tends to favor calibrated fairness approaches
- People consider both merit-based metrics and sensitive attributes like race
- Fairness perceptions change based on context and the magnitude of differences between candidates
- Need to consider public attitudes when designing algorithmic decision systems

Other Definitions: Equity

Definition 1 (Statistical Equity) *A predictor is statically equitable among demographic groups, a and b, if it satisfies $p(\hat{Y}|A = a) + p(Y|A = a) = p(\hat{Y}|A = b) + p(Y|A = b)$.*

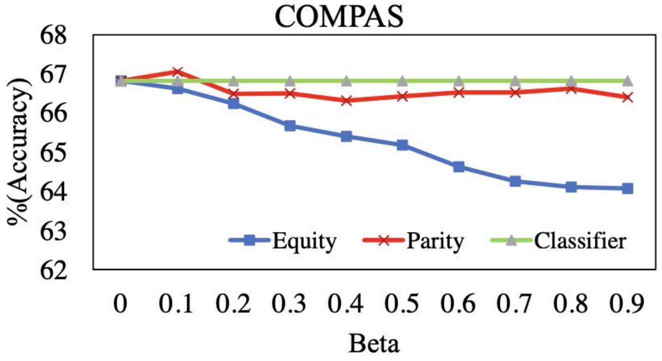


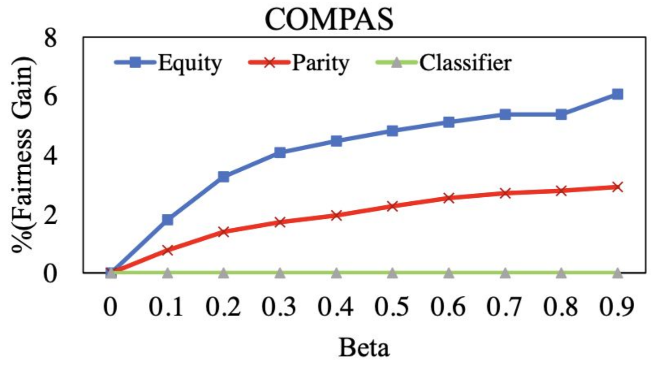
Equality

$$p(\widehat{Y}|A = \text{blue}) \\ = p(\widehat{Y}|A = \text{purple}) \\ = p(\widehat{Y}|A = \text{yellow})$$

Equity

$$p(\widehat{Y}|A = \text{blue}) + p(Y|A = \text{blue}) \\ = p(\widehat{Y}|A = \text{purple}) + p(Y|A = \text{purple}) \\ = p(\widehat{Y}|A = \text{yellow}) + p(Y|A = \text{yellow})$$

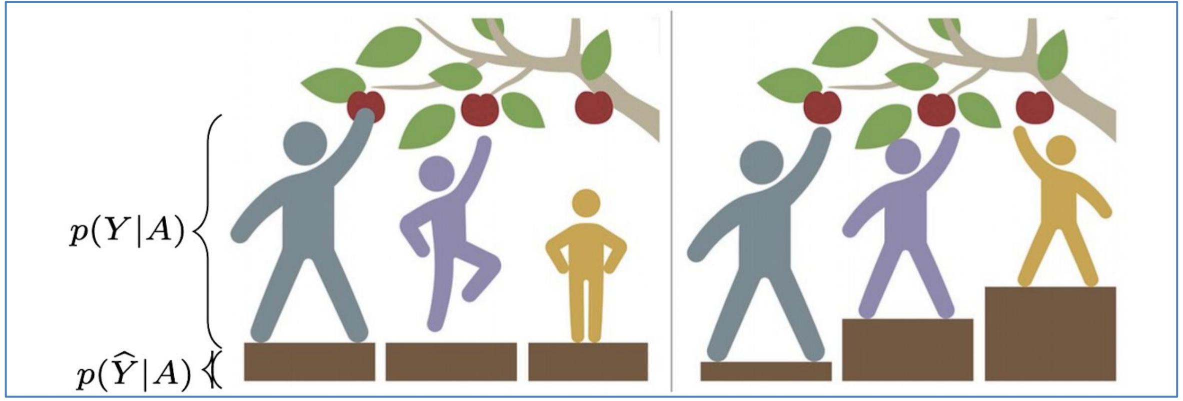




Fairness

Allocation Harms

- “A system withholds certain resources or opportunities from some groups.”*
- E.g. Loan allocation



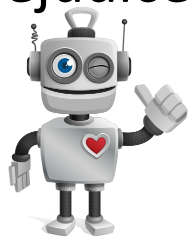
Mehrabi, N., Huang, Y., & Morstatter, F. (2020). Statistical Equity: A Fairness Classification Objective. arXiv preprint arXiv:2005.07293.

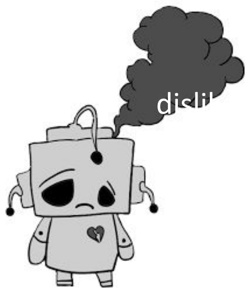
Representational Harms

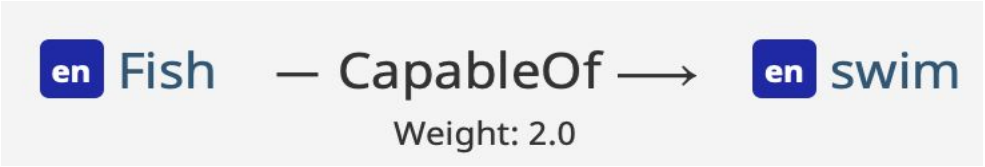
- Harm to the identity of certain groups.
- Regardless of allocating resources or opportunities.

Representational Harms

- Investigating representational harms in commonsense knowledge resources and tasks.
- What is Commonsense Knowledge?
  - “Facts about the world that all humans are expected to know.”**
  - Commonsense knowledge should be facts NOT stereotypes or biases in terms of favoritism or prejudice.

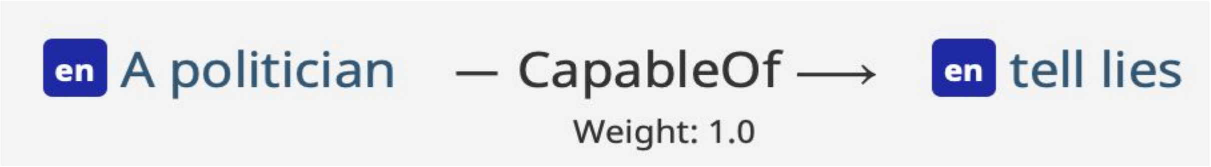






en Fish — CapableOf ⟶ en swim

Weight: 2.0



en A politician — CapableOf ⟶ en tell lies

Weight: 1.0

Conclusion

- Looked at different fairness definitions.
- Different levels: group, individual, subgroup.
- Saw reactions from the public.
- Not all fairness definitions can co-exist.