Fairness and Machine Learning: Limitations and Opportunities

Moritz Hardt
UC Berkeley

Where to start?

We live in a world of pervasive inequality, oppression, and discrimination.

As we use machine learning to formalize, scale, accelerate processes in this world, we run danger of perpetuating existing patterns of injustice.

But there's also a (somewhat fragile) opportunity to revisit decision making in various domains and reform existing processes for the better.

Important work to start with

Ruha Benjamin. Race After Technology: Abolitionist Tools for the New Jim Code

Meredith Broussard. Artificial Unintelligence: How Computers Misunderstand the World

Virginia Eubanks. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor

Safiya Noble. Algorithms of Oppression: How Search Engines Reinforce Racism

Cathy O'Neil. Weapons of math destruction

Joy Boulamwini, Kate Crawford, Timnit Gebru, Latanya Sweeney, Meredith Whitaker and many others.

“the New Jim Code”: *the employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era.*

- Ruha Benjamin

Focus for this tutorial

Discrimination in consequential decision making settings

- This excludes many other forms of injustice (and even unfairness)

US centric perspective (insofar as the examples and legal backdrop go)

Formal models and frameworks

- This is not meant to decenter the scholarship just mentioned
- This decidedly leaves room for non-technical interventions

Is discrimination not the point of machine learning?

Our concern is with *unjustified basis for differentiation*

- *Practical irrelevance*
  - *Sexual orientation in employment decisions*
- *Moral irrelevance*
  - *Disability status in hiring decisions*

Discrimination is *not* a general concept

Domain specific:

Concerned with important opportunities that affect people's lives

Group specific:

Concerned with socially salient categories that have served as the basis for unjustified and systematically adverse treatment in the past

Regulated domains (based on US law)

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **‘Public Accommodation’** (Civil Rights Act of 1964)

Extends to *marketing and advertising*; not limited to final decision

This list sets aside complex web of laws that regulates the government

Legally recognized ‘protected classes’ in the US

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Courts & Law

Supreme Court says gay, transgender workers protected by federal law forbidding discrimination

Source: Washington Post
June 15, 2020

Two legal doctrines in the US

Disparate treatment

Purposeful consideration of group membership

Intentional discrimination without consideration of group membership

Goal: Procedural fairness

Disparate impact

Avoidable or unjustified harm, possibly indirect

Goal: Distributive justice, minimize differences in outcomes

Some well-recognized tension between the two.

Some caveats about the law

Anti-discrimination law does not reflect one moral theory

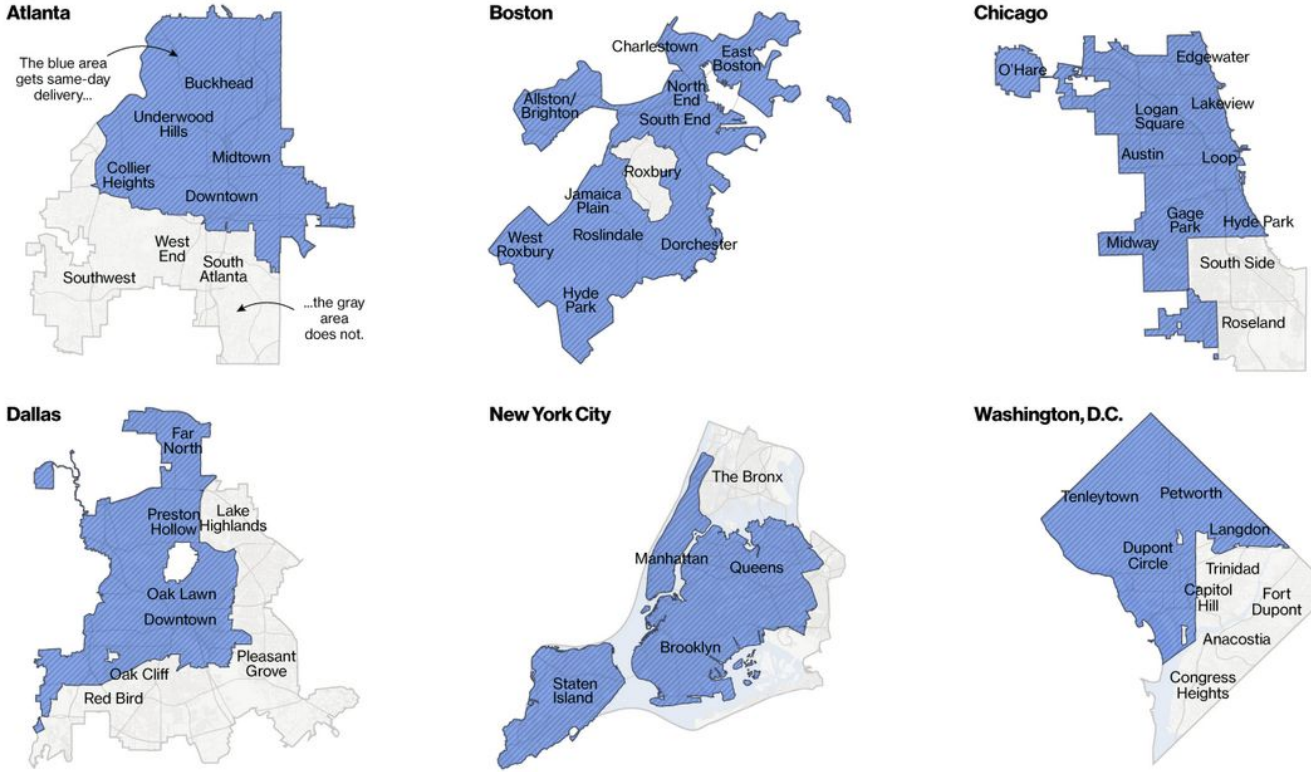
Legislations often were responses to civil rights movements, each hard fought through decades of activism

The law does not give us a “fairness definition” that we could readily formalize and operationalize

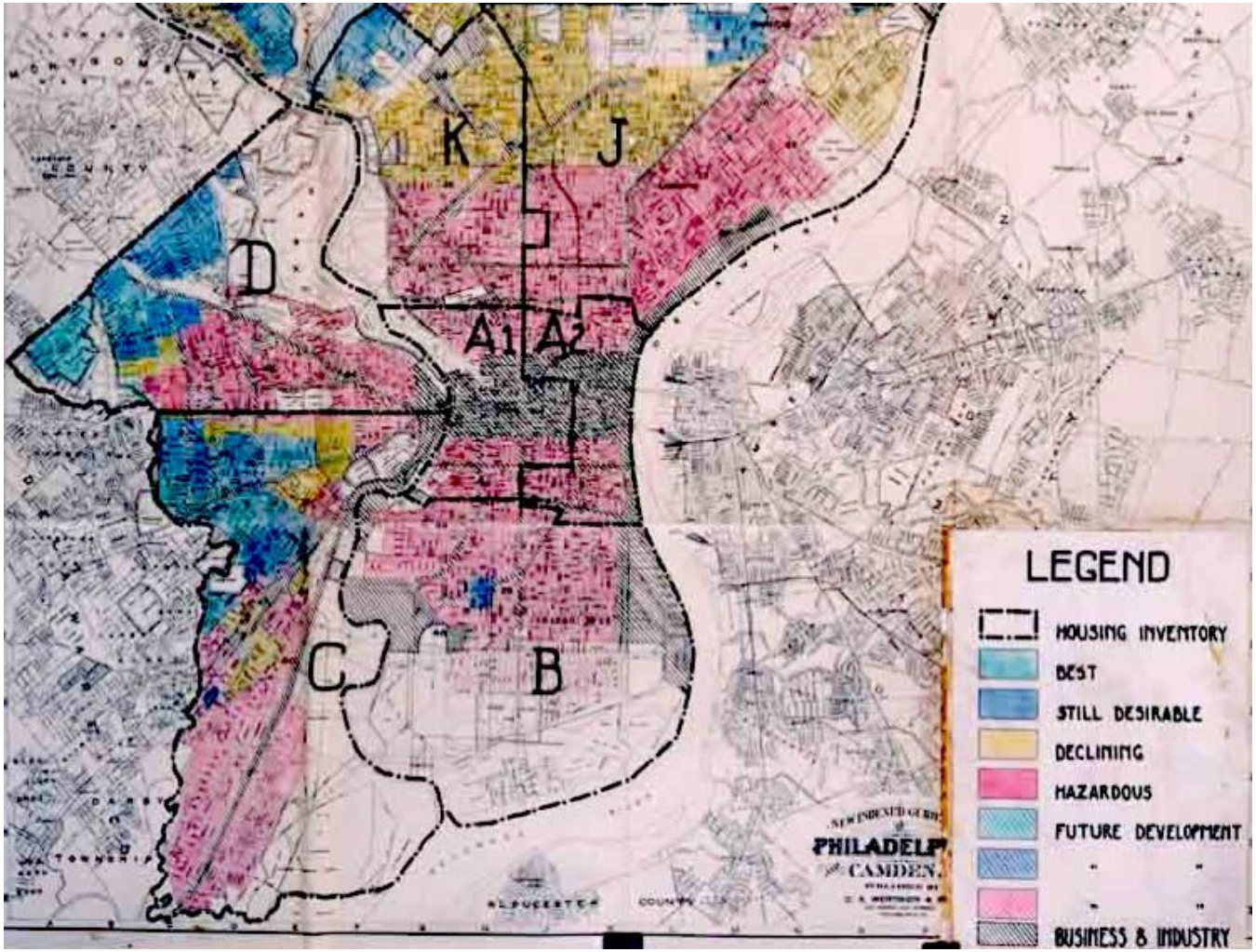
The failure of *fairness through unawareness*

Removing (or not including) “sensitive attributes” is no cure for fairness concerns and can exacerbate them.

Amazon same-day delivery coverage



Source: https://www.bloomberg.com/graphics/2016-amazon-same-day/


K
J
D
A1
A2
C
B
LEGEND
HOUSING INVENTORY
BEST
STILL DESIRABLE
DECLINING
HAZARDOUS
FUTURE DEVELOPMENT
BUSINESS & INDUSTRY

The failure of fairness through *unawareness*

Perhaps Amazon was just predicting *number of purchases*, which correlates with affluence, which correlates with race in the United States. *Amazon almost certainly did not look at their customers' race when they built this product.*

“We don’t consider that in our data” is *never* a valid argument.

So what should we do instead?

Overview

Part I (today): From a narrower perspective

Fairness criteria in classification

Part II (Thu): Toward a broader perspective

Causal models of decision-making settings

Dynamic models of socio-technical systems

Part I

Formal work on fairness in classification and decision-making

Pioneering work in educational testing (Cleary 1968) and economics (Becker 1957, Phelps 1972, Arrow 1973) on the heels of civil rights movement.

Computer science: Mostly post 2010, explosive increase in work since 2016

Why today? Urgency, scale, reach, and impact of algorithmic decisions

Machine learning fuels adoption and motivates some new technical problems, but also forces us to revisit fundamental normative questions

Formal prediction and decision making setting

Data described by covariates $\boldsymbol{X}$

Outcome variable $\boldsymbol{Y}$ (often binary, sometimes called *target variable*)

Our goal is to *predict* $\boldsymbol{Y}$ from $\boldsymbol{X}$

Use supervised machine learning to produce a score function $\boldsymbol{R = r(X)}$

Make binary decisions according to threshold rule $\boldsymbol{D} = \mathbf{1}\{\boldsymbol{R} > t\}$

Note: Think of these as random variables in the same probability space.

Where do score functions come from

Score ***R*** could be:

- Based on parametric model of the data **(*X, Y*)**, e.g. *likelihood ratio test*
- Non-parametric score, such as, Bayes optimal score **$R = E[\, Y \mid X \,]$**
- Most commonly, learned from labeled data using supervised learning

Decision theory 101

Outcome Y	Decision *D* = 0	Decision *D* = 1
0	*True negative*	*False positive*
1	*False negative*	*True positive*

True positive rate = Pr[***D*** = 1 | ***Y*** = 1]

False positive rate = Pr[***D*** = 1 | ***Y*** = 0]

True negative rate = Pr[***D*** = 0 | ***Y*** = 0]

False negative rate = Pr[***D*** = 0 | ***Y*** = 1]

Statistical fairness criteria

Introduce additional random variable ***A*** encoding membership status in a protected class

Equalize different statistical quantities involving group membership ***A***

Idea dates back at least to the 1960s with work of Anne Cleary about group differences in educational testing*

We'll review *three* common criteria

* See Hutchinson, Mitchell (2018).

Equalizing acceptance rate

Equal positive rate: For any two groups *a, b*, require
Pr[***D*** = 1 | ***A*** = *a*] = Pr[***D*** = 1 | ***A*** = *b*]

“Acceptance rate” equal in all groups

Generalization: Require ***D*** to be independent of ***A*** (Independence)

All sorts of variants, relaxations, equivalent formulations

Why this does not rule out *unfair* practices

One unfair situation: Make good/informed decisions in one group, poor/arbitrary decisions in other groups. Equalize positive rate.

This can happen on its own if we have less data or poor data in one group.

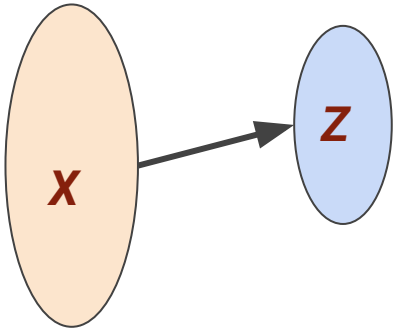
Example: Old *Framingham risk score* for coronary heart disease was created on cohort of white men, then used for other patients.

A *positive* call could be a false positive or a true positive. Moral intuition: You shouldn't get to match true positives in one group with false positives in another.

Achieving independence through representation learning

Lots of work out there on “fair representation” starting with work by Zemel et al. (2015).

General idea: Use deep learning tricks, such as adversarial learning, to train a representation of the data that is independent of group membership ***A***, while representing original data as well as possible.



$Z \perp A$

Equalizing error rates

For any two groups *a*, *b*, require

$\Pr[\boldsymbol{D} = 1 \mid \boldsymbol{Y} = 0, \boldsymbol{A} = a] = \Pr[\boldsymbol{D} = 1 \mid \boldsymbol{Y} = 0, \boldsymbol{A} = b]$ (equal false positive rate)

$\Pr[\boldsymbol{D} = 0 \mid \boldsymbol{Y} = 1, \boldsymbol{A} = a] = \Pr[\boldsymbol{D} = 0 \mid \boldsymbol{Y} = 1, \boldsymbol{A} = b]$ (equal false negative rate)

Generalization: Require ***D*** to be independent of ***A*** *given* ***Y***

Also makes sense for score: Require ***R*** to be independent of ***A*** *given* ***Y***

Error rate parity is a *post-hoc* criterion

At decision time, the decision maker doesn't know who is a positive/negative instance

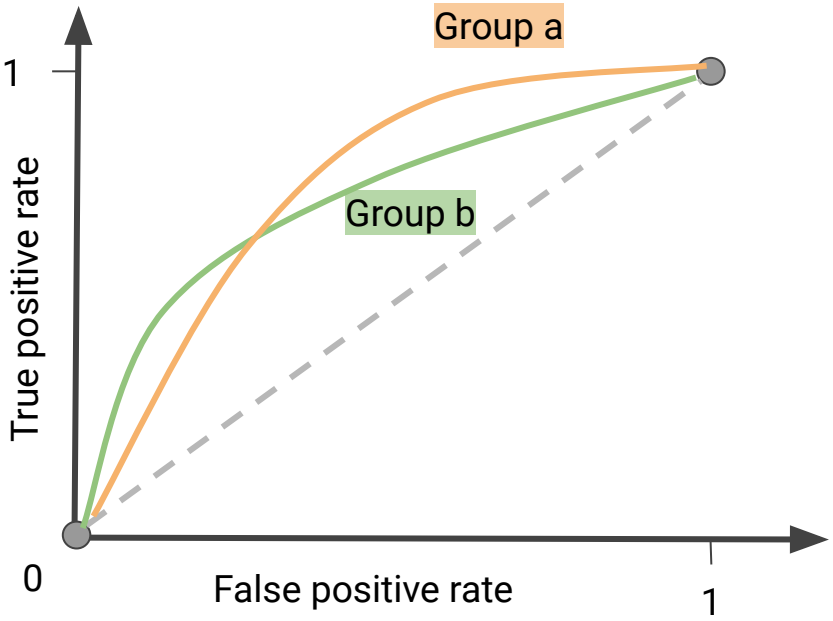
In hindsight, somebody can collect a group of positive instances and a group of negative instances and check how they were classified.

Group differences in this kind of post-hoc "audit" often strike people as unfair.

Interpretation in terms of ROC curve

Suppose ***D*** is threshold of a score ***R***

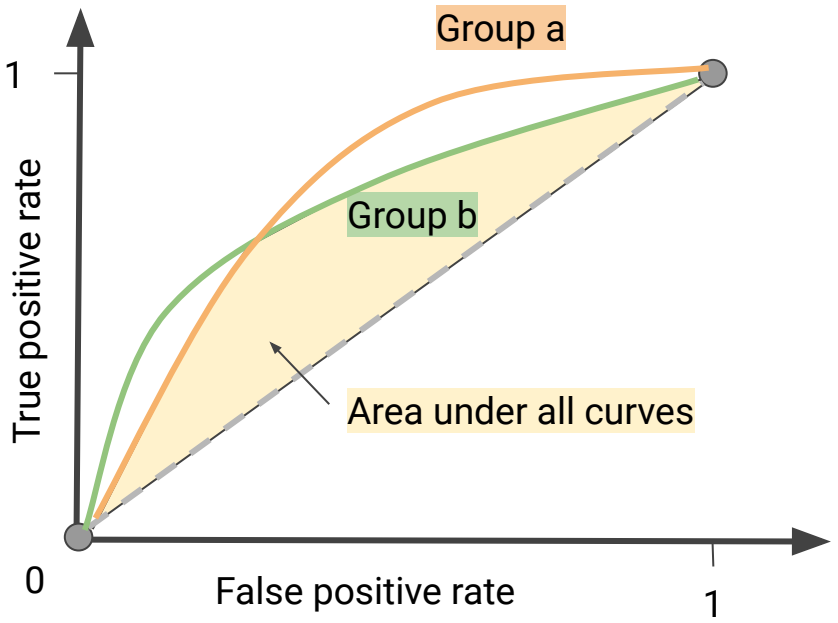
Error rate parity implies that ROC curve of score conditional on group must be under all curves.



Interpretation in terms of ROC curve

Suppose ***D*** is threshold of a score ***R***

Error rate parity implies that ROC curve of score conditional on group must be under all curves.

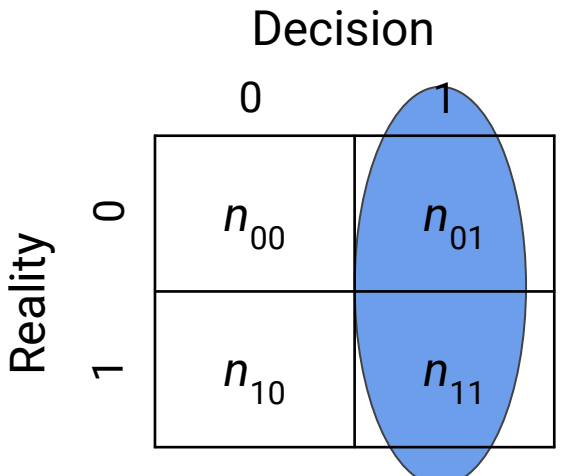


Column-wise criteria?

We could equalize expressions of the form
Pr[***Y*** = y | ***D*** = d, ***A*** = *a*]

These are called “column-wise” rates, i.e., false omission and false discovery rate.

Nothing wrong with this, but something closely related but different is more common.



Calibration

A score ***R*** is calibrated if: $\Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{R} = r] = r$

“You can pretend score is a probability” - although it may not actually be one!

Score value *r* corresponds to positive outcome rate *r*

Calibration by group: $\Pr[\boldsymbol{Y} = y \mid \boldsymbol{R} = r, \boldsymbol{A} = a] = r$

Follows from: ***Y*** independent of ***A*** conditional on ***R***

Calibration is an *a priori* guarantee

The decision maker sees the score value *r* and knows based on this what the frequency of positive outcomes is.

E.g., score 0.8 means 80% rate of heart failure on average over people who receive score 0.8.

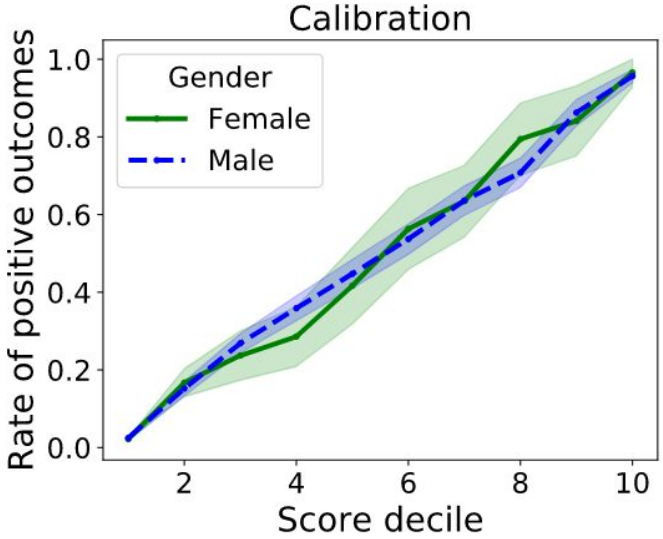
This guarantee (usually) does not hold at the individual level, e.g., “Mary’s individual risk of heart failure is 80%.”

Group calibration often follows from unconstrained learning

Informal theorem: Under reasonable conditions, the deviation from satisfying group calibration is upper bounded by the excess risk of the learned score relative to the Bayes optimal score function.

See Liu, Simchowitz, H (2019)

In other words, you shouldn't be surprised to see calibration follow approximately from unconstrained supervised learning.



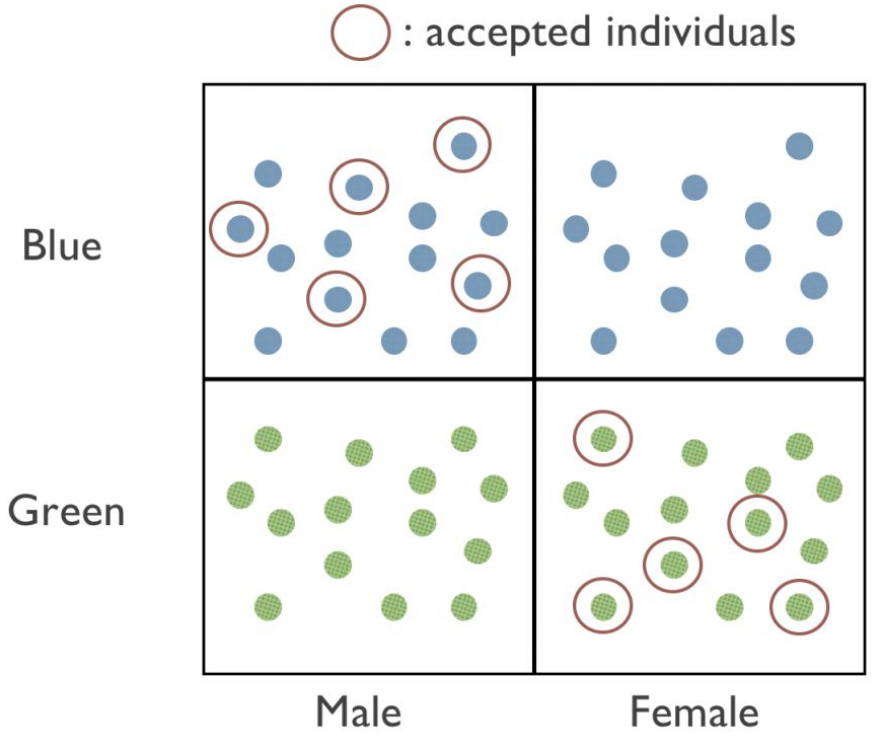
Example calibration of unconstrained learning on UCI adult data set.

Subgroup fairness

Ensuring fairness criteria between two groups can lead to violations within groups.

This motivated work on ensuring **subgroup fairness**.

See Kearns, Neel, Roth, Wu (2018), and Hébert-Johnson, Kim, Reingold, Rothblum (2018).



Illustrative example
from Kearns et al. (2018)

Recap

We saw three criteria:

- ***R*** independent of ***A*** (implies equal acceptance rate)
- ***R*** independent of ***A*** conditional on ***Y*** (implies equal error rates)
- ***Y*** independent of ***A*** conditional on ***R*** (implies calibration by group)

Can we have them all?

Incompatibility results

Informal theorem: Any two of these criteria are mutually exclusive in general.

Error rate parity vs calibration

Theorem:

1. Assume unequal base rates: $\Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{A} = a] \neq \Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{A} = b]$
2. Assume imperfect decision rule: $\boldsymbol{D}$ has nonzero error rates

Then, calibration by group implies that error rate parity fails.

Related result due to Chouldechova (2016), Kleinberg, Mullainathan, Raghavan (2017)

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

Essence of COMPAS debate

There's a risk score used, called COMPAS, used by many jurisdictions in the United States to assess "risk of recidivism". Judges may detain defendant in part based on this score.

ProPublica: Black defendants face higher false positive rate, i.e., more Black defendants labeled "high risk" end up ***not*** committing a crime upon release than among Whites labeled "high risk"

COMPAS maker Northpointe: But our scores are calibrated by group and Black defendants have a higher recidivism rate! Hence, this is unavoidable.

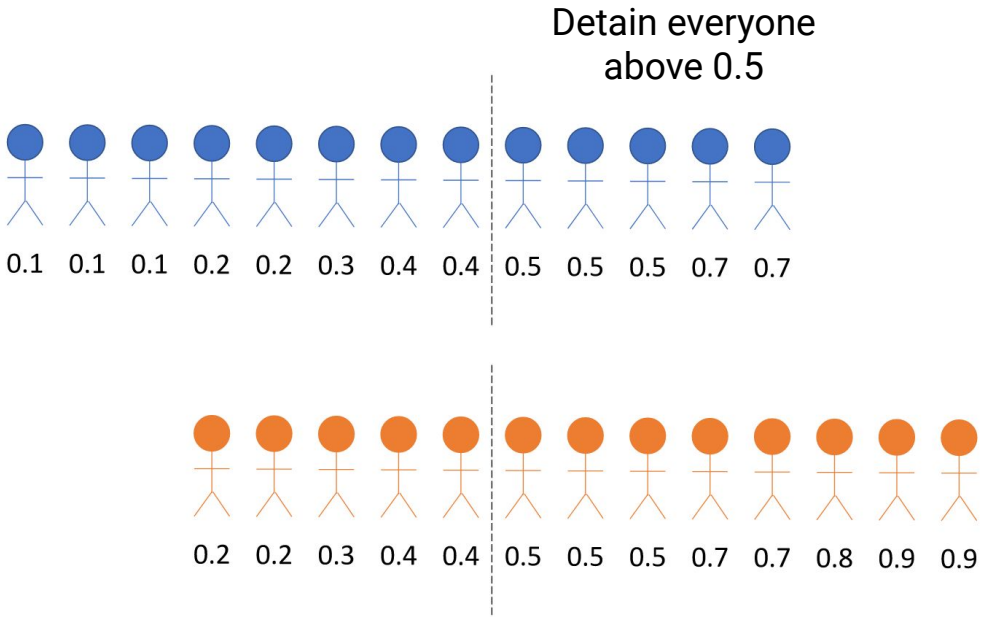
A first word of caution about COMPAS debate

Neither error rate parity nor calibration rule out blatantly unfair practices.

What's fair in criminal justice is not settled by appeal to one or the other criterion.

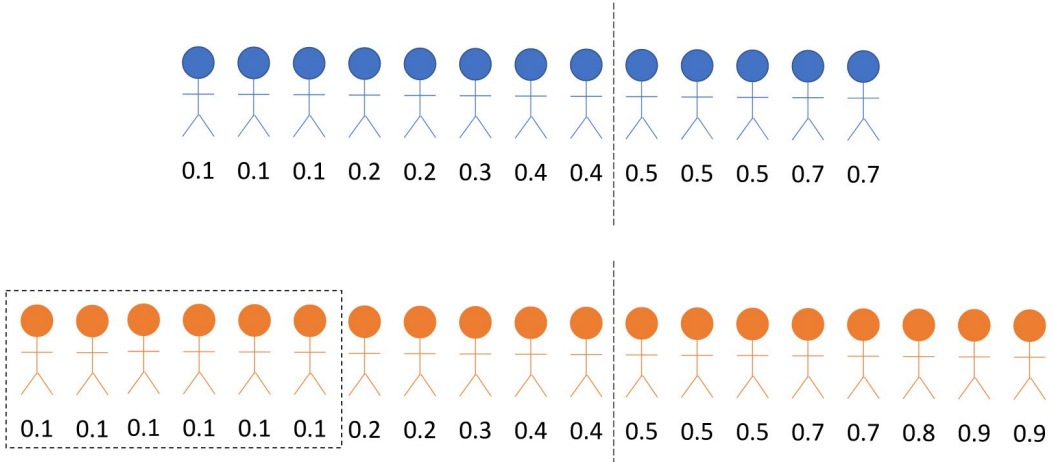
These properties are not meant to be "fairness certificates".

Consider these two groups



Detention rate	False pos. rate
38%	25%
61%	42%

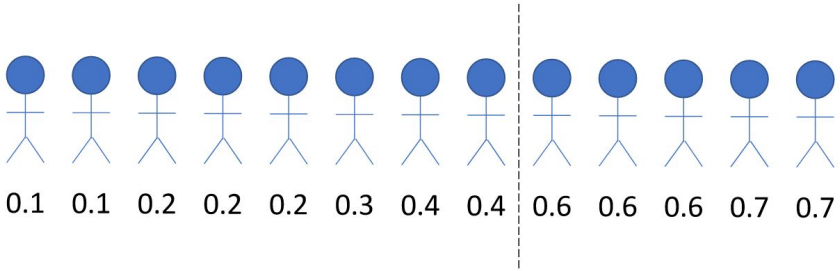
Equalizing rates may lead to undesired outcomes



Arrest more low risk individuals in orange group!

Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

An issue with calibration

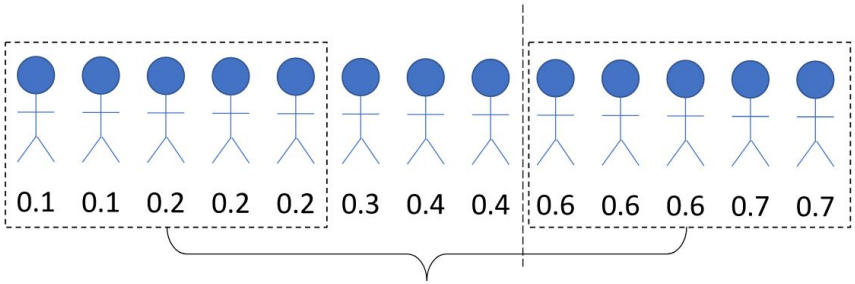


True probabilities of reoffending
(hypothetically, suppose we know them!)

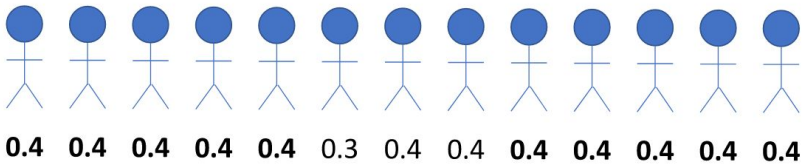
Detain everyone
above 0.5

Examples from: [Corbett-Davies, Pierson, Feller, Goel, Huq (2017)]

An issue with calibration



Average probability of re-offense is 0.4 in this subgroup



Calibrated new scores

No one is detained!

Is *prediction* too narrow a perspective?

Scholarly debate around COMPAS was largely about **tension between fairness criteria**.

Some rightfully point out **data and measurement problems**
(e.g., policing patterns influence variables such as criminal history and recidivism)

When is the issue not *how* we predict but *that* we predict?

Failure to appear in court

One approach: Predict failure to appear, jail if risk is high.

Alternative: Recognize that people fail to appear in court due to lack of child care and transportation, work schedules, or too many court appointments. Implement steps to mitigate these issues.

Alternative is part of the Harris County Lawsuit settlement: *"require Harris County to provide free child care at courthouses, develop a two-way communication system between courts and defendants, give cell phones to poor defendants and pay for public transit or ride share services for defendants without access to transportation to court." (Source: Houston Chronicle, April 2019)*

Toward a broader perspective

Statistical fairness criteria take data generating distribution as given and work with nothing but the joint statistics of ***(X, Y, R, A).***

If this statistical perspective is too narrow, how do we take salient social facts and context into account?

This will be the subject of Part II on Thursday.

Wrapping up

Fairness through unawareness *fails*!

Violations of fairness criteria trigger valid moral intuitions, and can surface normative questions about decision-making, as well as trade-offs and tensions between different interpretations of fairness.

However, statistical fairness criteria on their own cannot be a “proof of fairness”.

Nor are fairness criteria on their own a good objective function.

Background reading

A textbook in progress (mostly available online):

Barocas, H, Narayanan. Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org

Today's lecture roughly corresponds to Chapters 1 & 2.