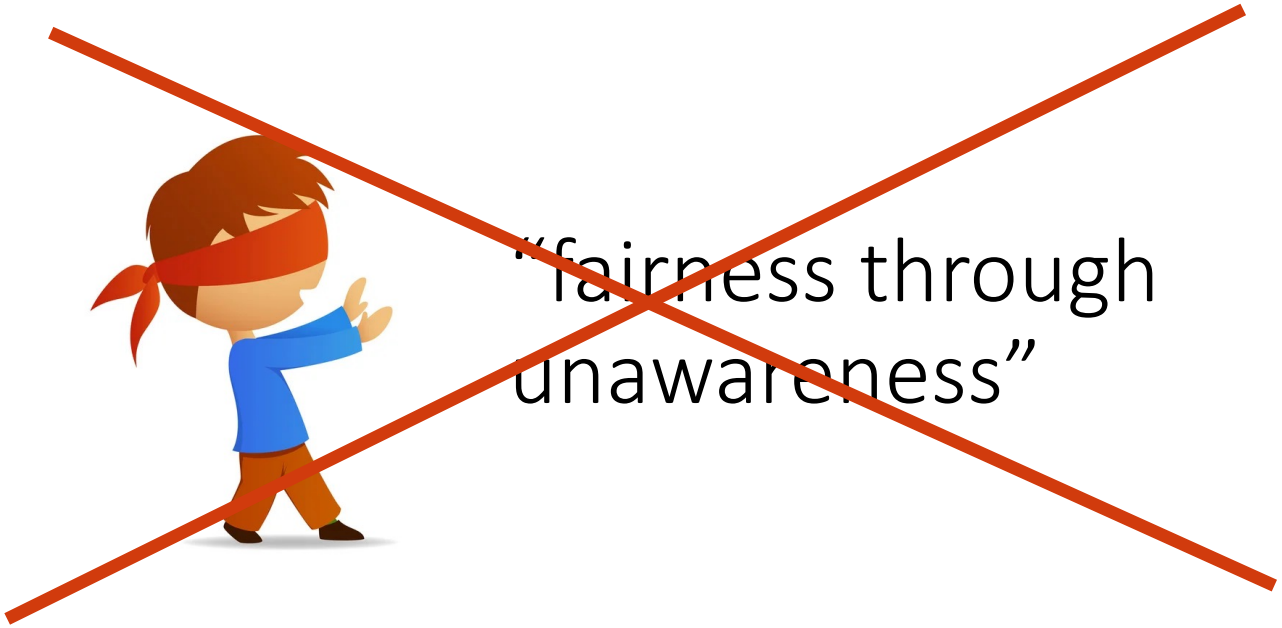
Formal Measures of Fairness: Beyond the Impossibility Theorem

Zina Ward (Dept. of Philosophy)
ML Group 12/8/23

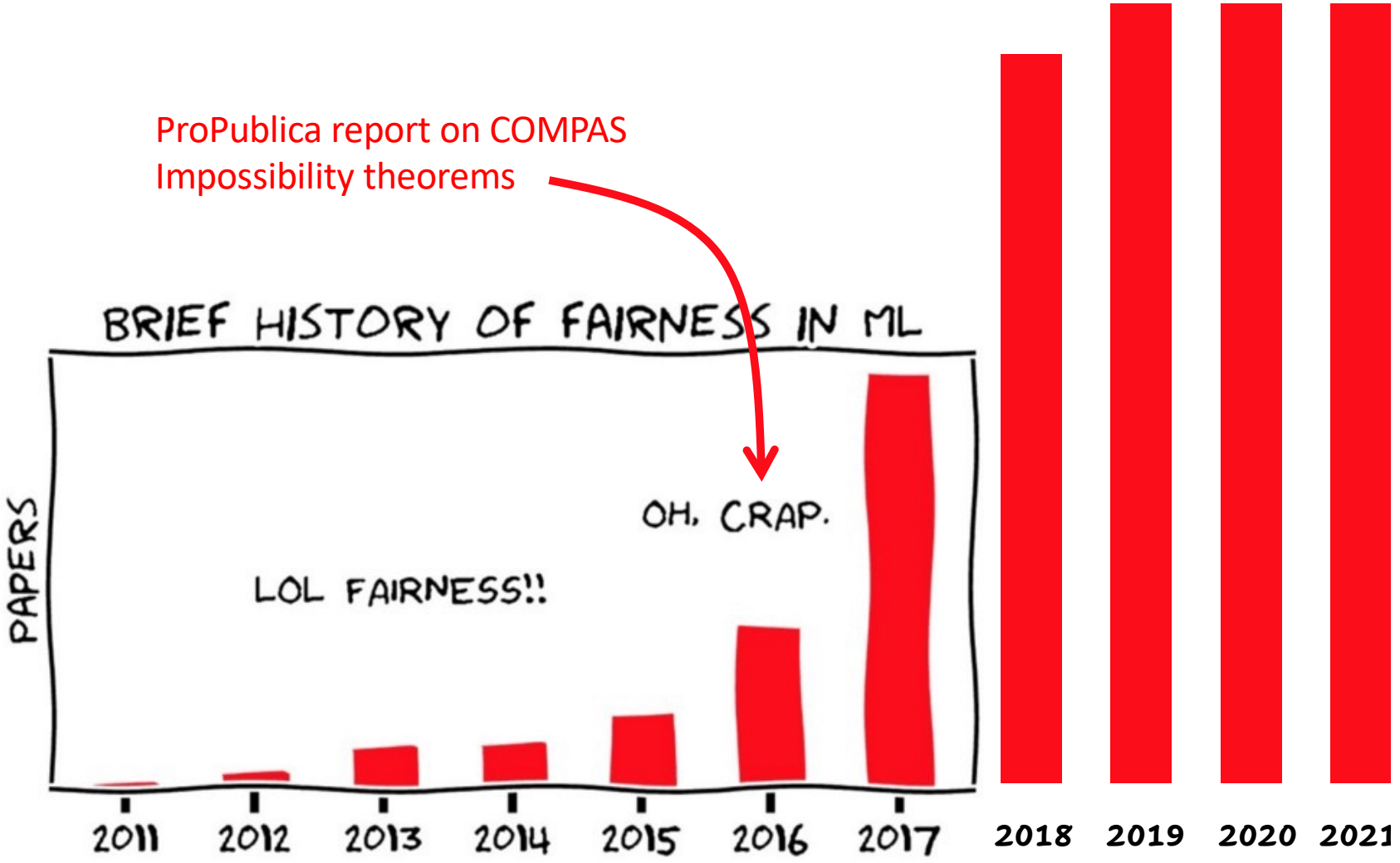
How do I know if my model is *fair* to members of sensitive classes?

e.g., sex, race, age, disability, national origin, religion



How do I know if my model is *fair* to members of sensitive classes?

e.g., sex, race, age, disability, national origin, religion

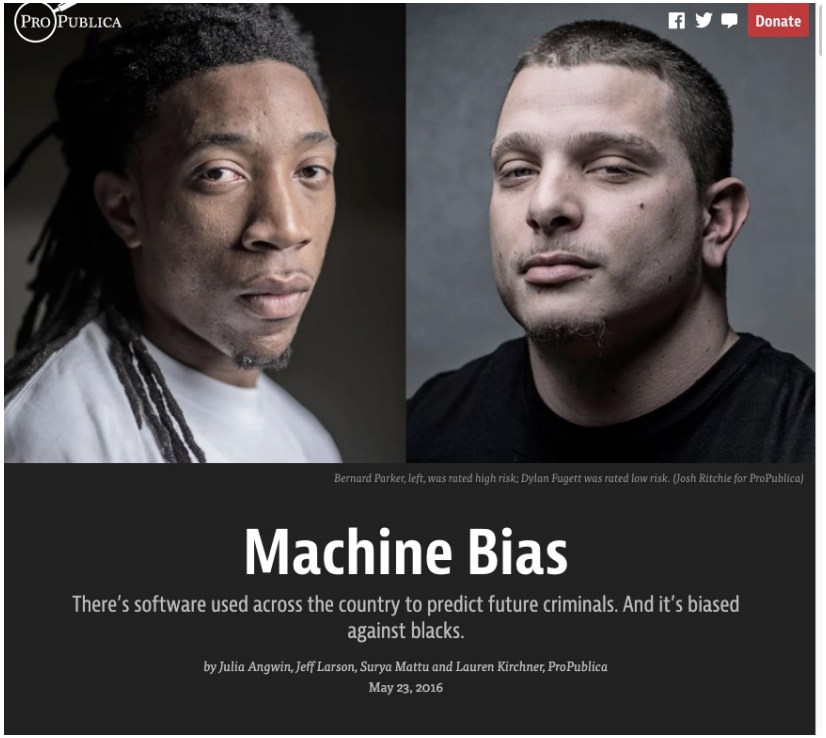

ProPublica report on COMPAS
Impossibility theorems
BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Plan for the talk

First part: Background on the COMPAS affair, brief(!) overview of impossibility theorems

Second part: What do we do now? Four strands in recent philosophical and legal work

The COMPAS Affair



Introducing COMPAS

- risk assessment algorithm
- “Correctional Offender Management Profiling for Alternative Sanctions”
- predicts a defendant’s likelihood of committing a crime in the future (output: risk score from 1-10)
- proprietary: Northpointe, Constellation Software → Equivant
- based on (6 of) 137 questions answered by defendants, or pulled from criminal record (NOT including race)

Use of COMPAS

Many risk assessment tools: designed to help judges decide on treatment (drug treatment? mental health counseling? probation?).

In practice, COMPAS (and other risk assessment tools) are used for…

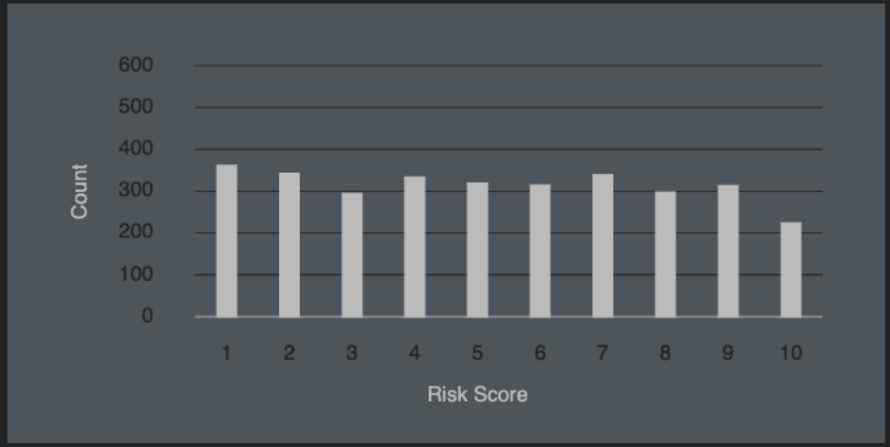
- treatment decisions
- sentencing decisions
- decisions about pre-trial release

ProPublica: Accusations of Unfairness

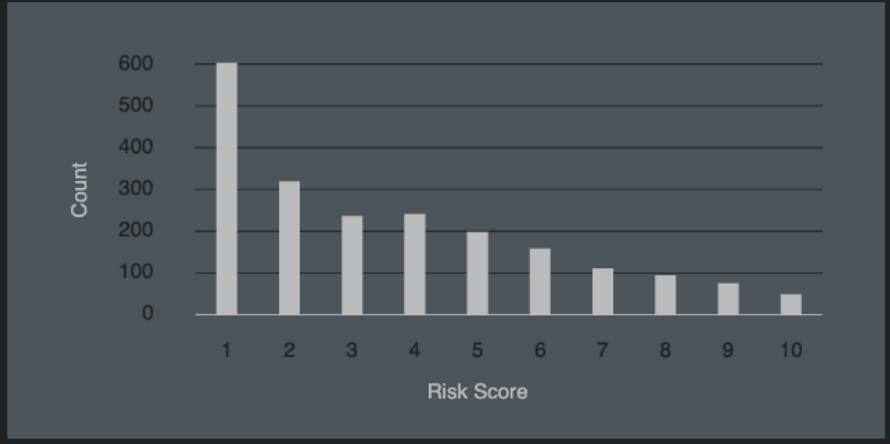
- obtained risk scores of 7,000 people arrested in Broward County in 2013 and 2014
- followed them for 2 years: how many were charged with new crimes?
- same accuracy: ~62% for both white and black defendants
- different false positive/negative rates for black white defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Black Defendants' Risk Scores



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Northpointe's Response

"Northpointe does not agree that the results of [ProPublica's] analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model."

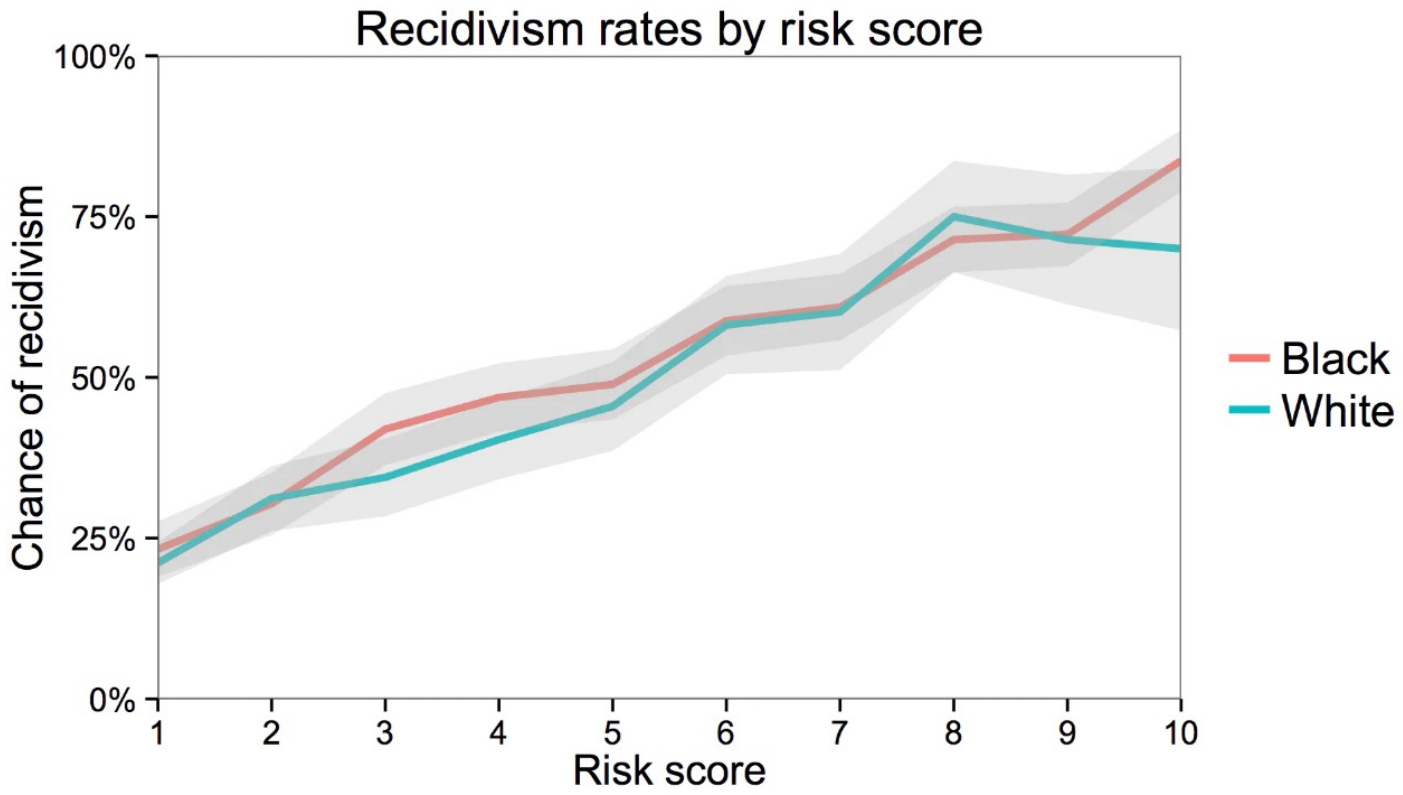
NORTHPOINTE

CONSTELLATION SOFTWARE INC.

COMPAS is fair because its scores mean the same thing regardless of race:

ex. 60% of white defendants who scored a "7" reoffended.

61% of black defendants who scored a "7" reoffended.



Recidivism rate by risk score and race. White and black defendants with the same risk score are roughly equally likely to reoffend. The gray bands show 95 percent confidence intervals.

Corbett-Davies et al. 2016

Subsequent Academic Work

There are multiple ways of formalizing fairness^, and under certain conditions*, it's impossible for a risk score to satisfy all fairness criteria at once.

^ more than 20, by recent counts

* different base rates, e.g., the overall recidivism rate for black defendants (52%) is higher than for white defendants (39%)

Three (Families of) Observational Fairness Criteria

Fairness as a property of joint distribution of…

sensitive attribute *A*

target variable *Y*

score/classifier *R*

Table 3: Non-discrimination criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

see Barocas et al. (2022), *Fairness and Machine Learning: Limitations and Opportunities*, Chapter 3, "Classification"

Impossibility Theorems

In most cases, you can’t satisfy any two of these criteria at once.

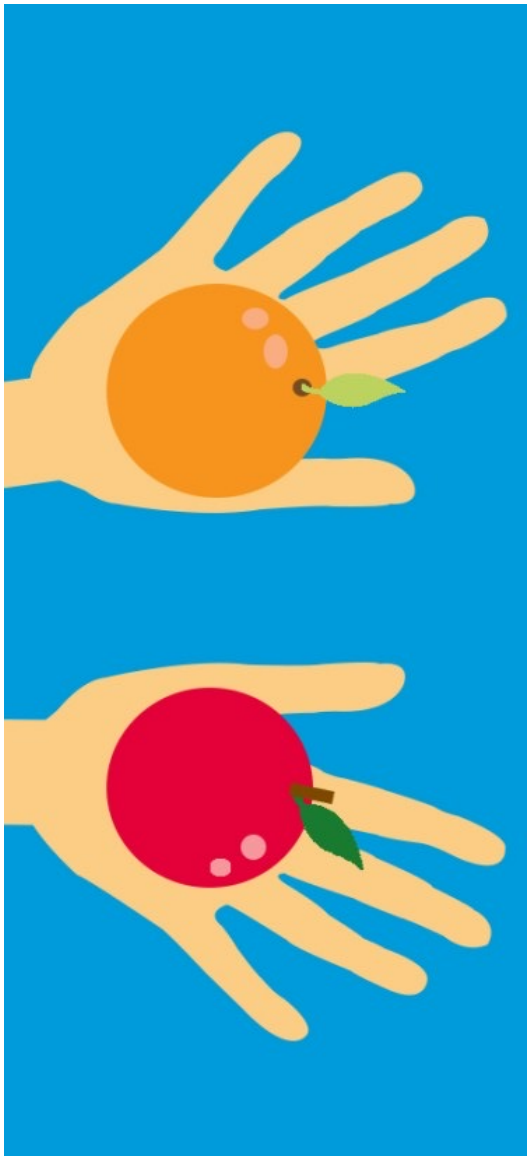
the sensitive attribute is correlated with the target variable

Impossibility Theorems

Trade-offs are necessary: In most cases, you can't satisfy any two of these three fairness criteria.

COMPAS: Separation vs. Sufficiency

- **Kleinberg et al. (2016),** "Inherent Trade-Offs in the Fair Determination of Risk Scores"
- **Chouldechova (2017),** "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments"
- **Corbett-Davies et al. (2017),** "Algorithmic decision making and the cost of fairness"



How should we *conceive of* fairness in light of these trade-offs?

How should we *promote* fairness in algorithmic decision-making?

Is it true that “total fairness cannot be achieved” (Berk et al. 2018)?

Plan for the talk

First part: Background on the COMPAS affair, brief(!) overview of impossibility theorems

Second part: What do we do now? Four strands in recent philosophical and legal work

Recent Philosophical and Legal Scholarship

1) The trade-offs between fairness criteria are not an artifact of mathematization: they reflect deep and genuine tensions within moral ideals.
2) Which fairness criterion we should prioritize depends on the context.
3) Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.
4) Satisfying formal fairness criteria alone doesn't necessarily make the deployment of an algorithm fair: we also need to look at the broader context.

1. The trade-offs between fairness criteria reflect deep and genuine tensions within moral ideals.

Egalitarianism: All people should be treated as equals.

equality of opportunity vs. **equality of outcome**

sufficiency: each person should have an equal chance to obtain goods

independence: each person should have an equal share of a good

Tension: The egalitarian often has to *choose* between equality of opportunity and equality of outcome.

Friedler et al. 2016, Binns 2017

1. The trade-offs between fairness criteria reflect deep and genuine tensions within moral ideals.

Discrimination Law: Two doctrines

disparate treatment vs. **disparate impact**

ensure that similar people are treated similarity

ensure that members of a protected class are not differently impacted

sufficiency

separation

aim: procedural fairness

aim: distributive justice

Tension: The legal system sometimes has to navigate trade-offs: to avoid disparate impact, one has to violate disparate treatment; to ensure disparate treatment, one has to put up with disparate impact.

Barocas & Hardt 2017

1. The trade-offs between fairness criteria reflect deep and genuine tensions within moral ideals.

genuine moral trade-offs → no technical fix!

"[R]ather than reveal that there are no right answers, Kleinberg et al. and Chouldechova show that there are no easy answers. The community has correctly recognized that fairness is a fundamentally hard problem, but misdiagnoses why.

Fair machine learning is hard not because of statistical or computational challenges, but because striving for fairness is ultimately a process of continual social negotiation and adjudication between competing needs and visions of the good."

(Green and Hu 2018, 3)

1. The trade-offs between fairness criteria reflect deep and genuine tensions within moral ideals.

Table 3 Key philosophical perspectives on inequality

Philosophical perspective	Acceptable inequalities	Unacceptable inequalities
Formal equality of opportunity/procedural fairness [26]	Any inequality as long as the opportunity was open to all	Treatment inequality
"Fair equality of opportunity" [49, 50]	Natural, talent, and preference inequalities	Socioeconomic, treatment inequalities
Rawlsian EOP + Difference principle [49]	Natural, talent, and preference inequalities, plus any inequality benefiting the most disadvantaged society members in long-term impact	Socioeconomic, treatment inequalities
Equality of outcome/condition/welfare [26]	None - all members should get the exact same outcome	All
Luck egalitarianism [15]	Effort-based inequalities (e.g. preference)	Circumstances (e.g. natural inequality)
Equality of freedom/autonomy [53]	Inequality resulting in "genuinely free" choices	Any inequality hindering freedom
Sufficiency/equality of capability [60]	Any inequality as long as everyone is above the level of sufficiency	Any resulting in people falling below sufficiency levels
Prioritarianism [47, 52]	Any inequality reduction should prioritise resource allocation to those who are worst off	None as long as the worst off are prioritised
Desert [31, 32]	Any inequality based on what he/she "deserves"	Any inequality that does not equate to the person's deservingness

Lee et al. 2021

Recent Philosophical and Legal Scholarship

1) The trade-offs between fairness criteria are not an artifact of mathematization: they reflect deep and genuine tensions within moral ideals.
2) Which fairness criterion we should prioritize depends on the context.
3) Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.
4) Satisfying formal fairness criteria alone doesn't necessarily make the deployment of an algorithm fair: we also need to look at the broader context.

2. Which fairness criterion we should prioritize depends on the context.

Philosophers and legal scholars have been developing arguments *for* or *against* prioritizing one fairness criterion or another:

Huq (2019): against separation

Hellman (2020): in defense of separation

Hedden (2021): in defense of sufficiency

Long (2022): in defense of sufficiency

Grant (2023): in defense of separation

Holm (2023): in defense of separation

2. Which fairness criterion we should prioritize depends on the context.

Others argue that the choice between criteria should be *context-sensitive* (e.g., Binns 2017).

Walzer (2008): How can goods and resources be fairly allocated?

No single universal principle of justice. Rather: different distributional principles apply in different spheres of life

- civil rights (e.g., voting): absolute equal distribution of the good
- economic goods, social positions (e.g., job): equality of opportunity

2. Which fairness criterion we should prioritize depends on the context.

"Equality of opportunity may be an appropriate metric to apply to model that will be deployed in the sphere of 'economic justice'; e.g**. selection of candidates for job interviews or calculation of insurance**. But in contexts which fall under the sphere of civil justice, we may want to impose more blunt metrics like equality of outcome. This might be the case for **airport security checks**, where it is important to a sense of social solidarity that no group is overexamined as a result of a predictive system, even if there really are differences in base rates.

We therefore can't assume that fairness metrics which are appropriate in one context will be appropriate in another."

(Binns 2017, 7)

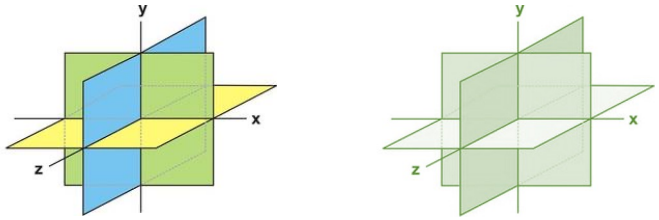
Recent Philosophical and Legal Scholarship

1) The trade-offs between fairness criteria are not an artifact of mathematization: they reflect deep and genuine tensions within moral ideals.
2) Which fairness criterion we should prioritize depends on the context.
3) Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.
4) Satisfying formal fairness criteria alone doesn't necessarily make the deployment of an algorithm fair: we also need to look at the broader context.

3. Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.

Individual Fairness

Dwork et al. (2012): ”Individuals who are similar with respect to a particular task should be classified similarly.”

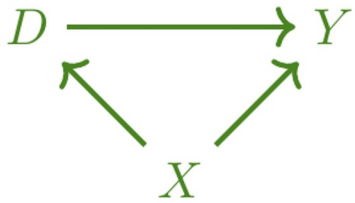


similarity space outcome space

fair = individuals close to one another in similarity space are close to one another in outcome space

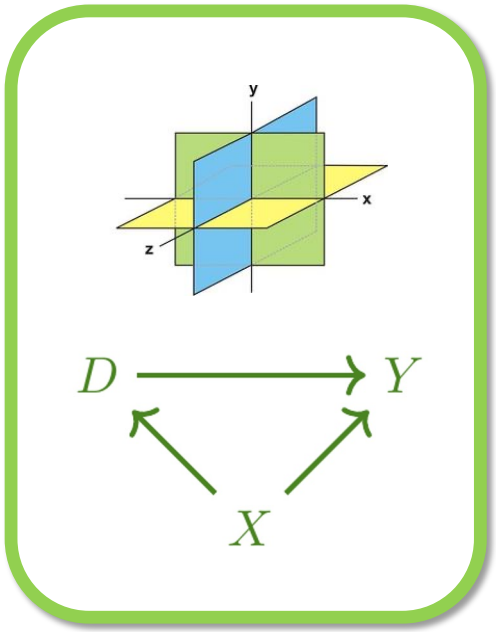
Causal/counterfactual fairness

Build a causal model of the data-generation mechanism. (e.g. Kusner et al. 2017).



fair = would give the same prediction had the individual been of another race, gender, etc.

3. Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.



Both individual fairness criteria and causal/counterfactual fairness criteria require *prior moral judgments about fairness*.

(Green & Hu 2018, Fazelpour & Lipton 2020, Binns 2020, Hu 2020, Fleisher 2021, Zimmerman & Lee-Stronach 2022)

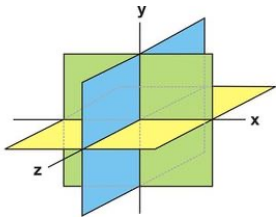
They might still be useful criteria, but we shouldn't take them to provide reductive, non-circular *definitions* of fairness.

3. Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.

Individual Fairness

What qualifies as *task-relevant similarity* depends on moral judgments about fairness.

e.g. hiring: include quality of high school?



Barocas et al. (2022): IF "possesses little normative substance"

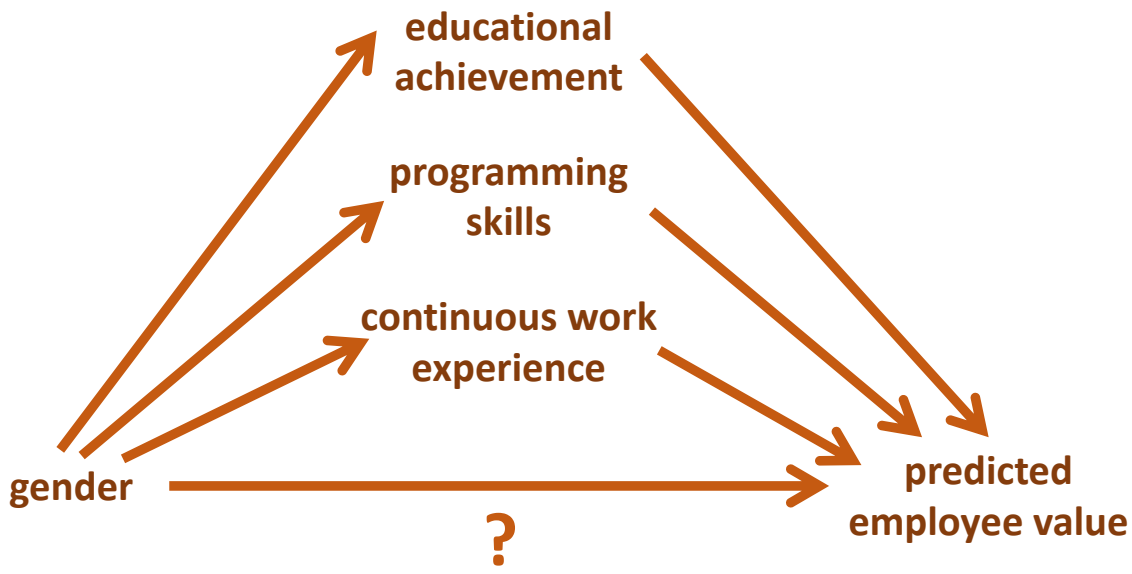
Causal/counterfactual fairness

Fairness is model-relative. And building causal models requires judgments about what variables to include.

e.g. hiring: should continuous work experience be included in model?

Is this hiring model fair to women?

Causal criterion: The model is unfair only if gender is a direct cause of predicted value to the company *while controlling for other, indirect causes.*

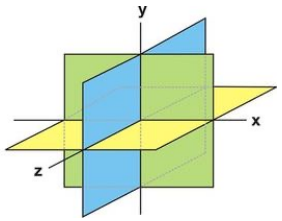


3. Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.

Individual Fairness

What depends as *task-relevant similarity* depends on moral judgments about fairness.

e.g. hiring:
include prestige of undergraduate institution?



Barocas et al. (2022): IF "possesses little normative substance"

Causal/counterfactual fairness

Fairness is model-relative. And building causal models requires judgments about what variables to include.

"When you draw a causal diagram that looks to illuminate the causal effects of any social category… you can't claim to only be using statistical methods to expose social scientific causes and effects. From the start, you are engaging in good old social and political analysis, with all its attendant normativity." (Hu 2020)

Recent Philosophical and Legal Scholarship

1) The trade-offs between fairness criteria are not an artifact of mathematization: they reflect deep and genuine tensions within moral ideals.
2) Which fairness criterion we should prioritize depends on the context.
3) Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.
4) Satisfying formal fairness criteria alone doesn't necessarily make the deployment of an algorithm fair: we also need to look at the broader context.

4. Formal fairness criteria do not guarantee true fairness: one must also look at the broader context.

There is a "risk that human decision-makers will mistakenly think that they have achieved procedural justice after having implemented such a [formal fairness] strategy" (Zimmerman & Lee-Stronach 2022, 15).

But there can be unfairness associated with a model even if it satisfies formal fairness criteria.

Green & Hu 2018
Selbst et al. 2019
Fazelpour & Lipton 2020
Hellman 2020
Castro 2022
Green 2022
Zimmerman & Lee-Stronach 2022
Creel & Hellman 2022
Grant 2023

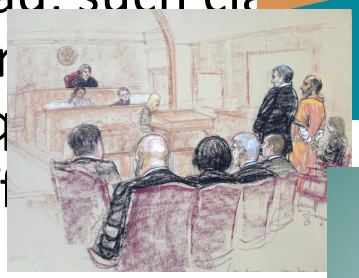
4. Formal fairness criteria do not guarantee true fairness: one must also look at the broader context.

Even if a model satisfies observational fairness criteria, its use may still be unfair if…

- it is unfair in a *non-comparative* sense
- it involves *procedural* unfairness
- it "*compounds injustice*" (Hellman 2020)

4. Formal fairness criteria do not guarantee true fairness: one must also look at the broader context.

> "[I]t is not sufficient to declare that an algorithm is fair
> the basis of mathematical tests alone. Instead, such cla
> must also account for relational and structur
> theories of change for remedying those ineq
> whether algorithms actually represent an eff
> advancing the desired reforms."
>
> (Green 2022, 90)



Recent Philosophical and Legal Scholarship

1) The trade-offs between fairness criteria are not an artifact of mathematization: they reflect deep and genuine tensions within moral ideals.
2) Which fairness criterion we should prioritize depends on the context.
3) Non-observational fairness criteria cannot be applied without taking on substantive moral commitments.
4) Satisfying formal fairness criteria alone doesn't necessarily make the deployment of an algorithm fair: we also need to look at the broader context.

thanks!

zward@fsu.edu