DUKE
COMPUTER SCIENCE

CompSci 590.01

Causal Inference in Data Analysis with Applications to Fairness and Explanations

Lecture 6:
Fairness
&
Causal Fairness

Sudeepa Roy

Reading

(these papers can be covered more)

- FairML book, chapter 4: https://fairmlbook.org/
- Loftus et al., Arxiv (2018): Causal Reasoning for Algorithmic Fairness (https://arxiv.org/pdf/1805.05859.pdf)
- Kusner et al., NeurIPS (2017): Counterfactual Fairness (https://papers.nips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf)
- Salimi et al., SIGMOD (2019): Interventional Fairness: Causal Database Repair for Algorithmic Fairness (https://dl.acm.org/doi/10.1145/3299869.3319901)

Acknowledgement (big thanks!):

Most of the slides are by Prof. Babak Salimi (UCSD) with small modifications from a joint short course taught at the Reasoning Web Summer School in 2022.

Announcements: Feb 2

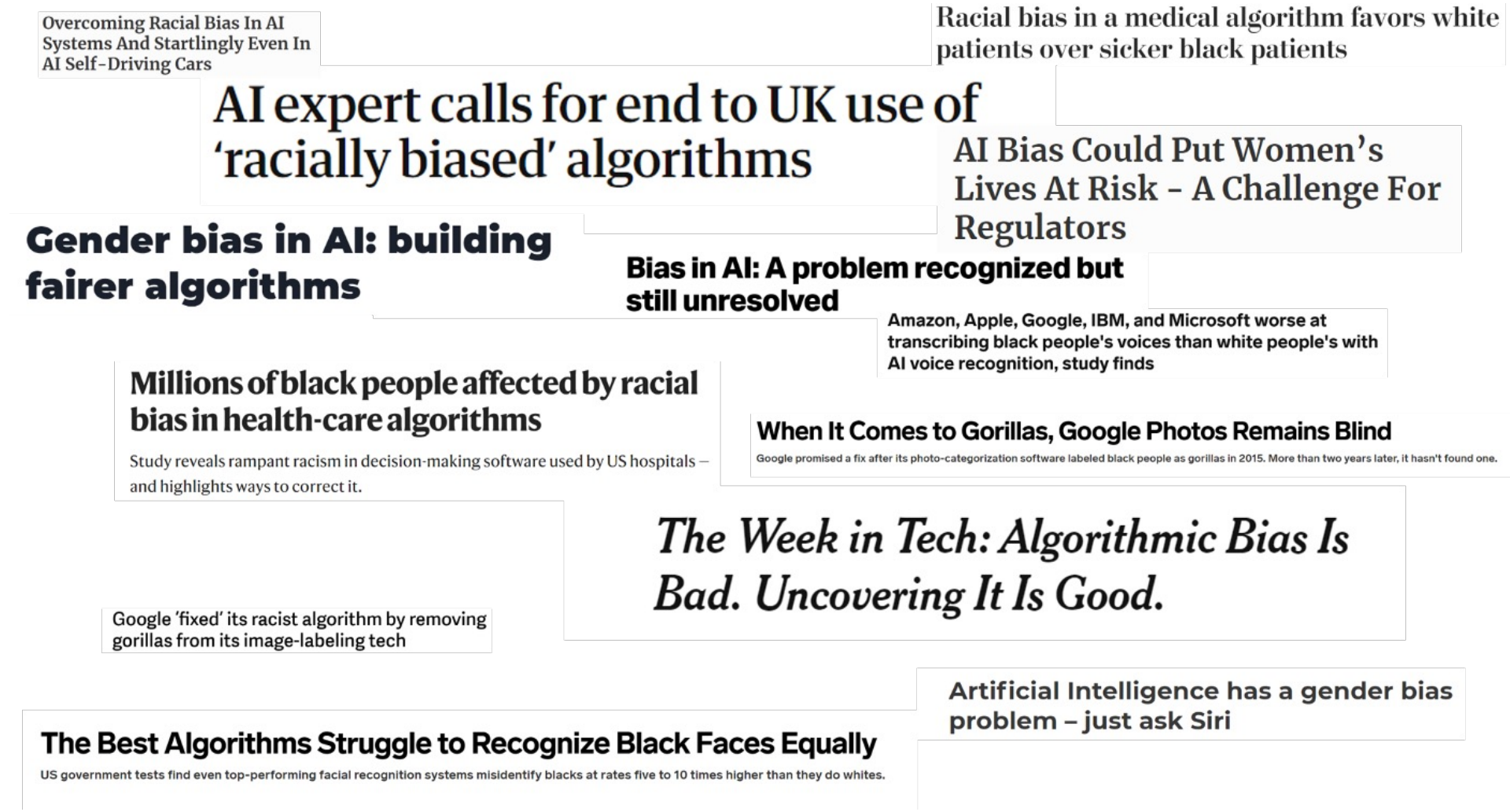
Timeline:

- Presentation topics & presnters' names due Tuesday 2/2
  - Then Sudeepa checks all and makes a balance of topics covering important papers & marks with green if a topic is final
  - First presentation starts on 2/14 – topic is final
  - Check out "paper review" instructions on Ed
- Initial project ideas & teammates' names due: Tuesday 2/7
  - Please share on Google doc or Overleaf (latex)
  - You may think about a project idea along the papers' topic you are presenting, and post on Ed for another teammate if you need one. Project in groups of 2 is also fine.
- Project proposal due: Tuesday 2/14

Causal Inference and Fairness

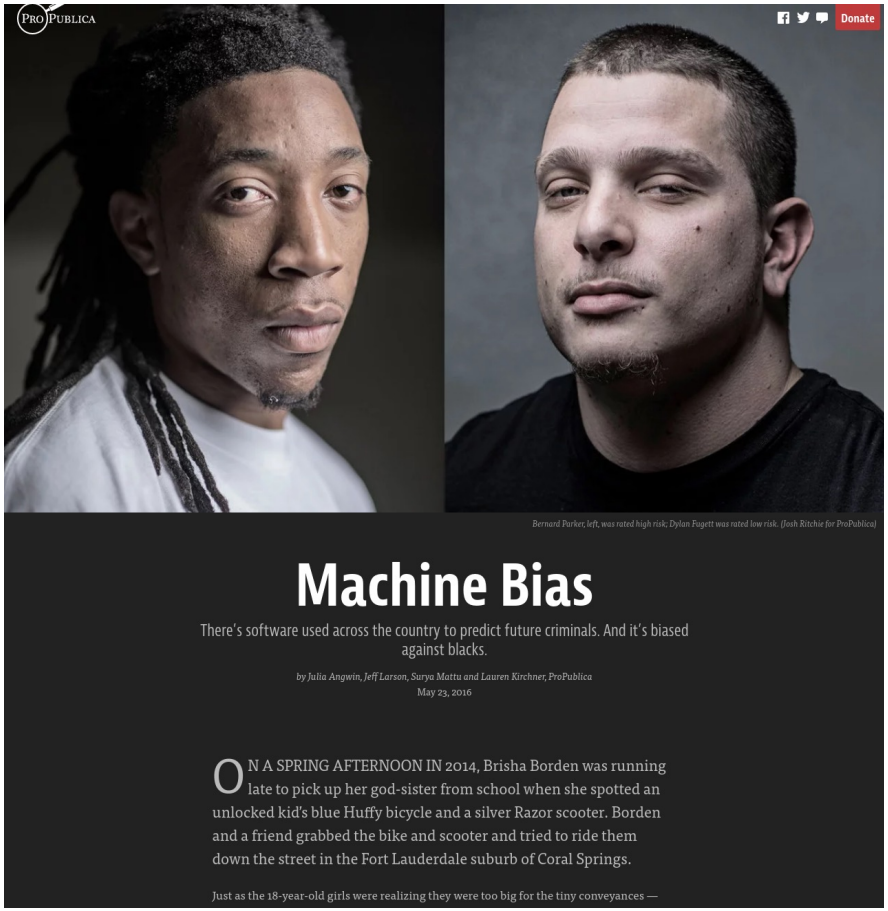
Fairness – very important in Responsible Data Science

Algorithmic Fairness



source: https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

Algorithmic Fairness



In 2016, a team of journalists from ProPublica constructed a dataset of more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014 in order to analyze the efficacy of COMPAS.

In addition, they collected data on future arrests for these defendants through the end of March 2016.

"<…> 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."

Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS

Images and excerpts from

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

Algorithmic Fairness

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning

""Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry."

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Announcements: Feb 7

Timeline:

- Presentation topics almost final!
  - First presentation starts on 2/14
  - Check out “paper review” instructions on Ed
- Initial project ideas & teammates’ names due TODAY: Tuesday 2/7
  - See Ed post
  - Template on overleaf shared
  - You may think about a project idea along the papers’ topic you are presenting, and post on Spreadsheet/ Ed for more teammates. Project in groups of 2 is also fine.
- Project proposal due: Tuesday 2/14

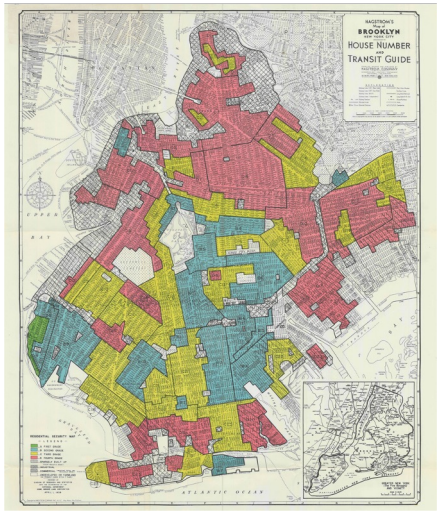
What is algorithmic bias?

- Algorithm bias is the lack of fairness that emerges from the output of a computer system

- Fairness is typically defined in terms of invariance of algorithmic decisions to variables that considered as sensitive

- Examples of sensitive variables: gender, ethnicity, sexual orientation, disability, etc.

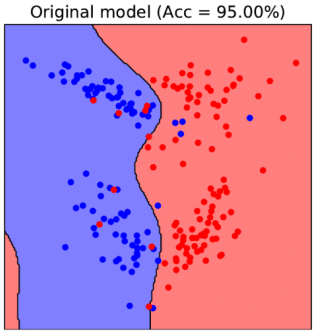
What are the sources of bias ?



src: NYTimes

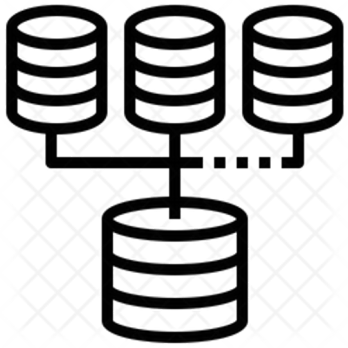
Historical bias in training data

src: openai.com



src: https://labs.f-secure.com

Adversarial data attacks

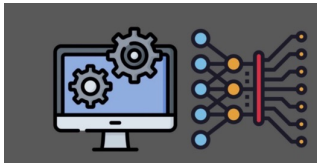


Data integration

src: nagwa.com

src: datacubed.com

Selection bias



Model design choices

Hooker, Sara. "Moving beyond “algorithmic bias is a data problem”." *Patterns* 2.4 (2021): 100241.

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

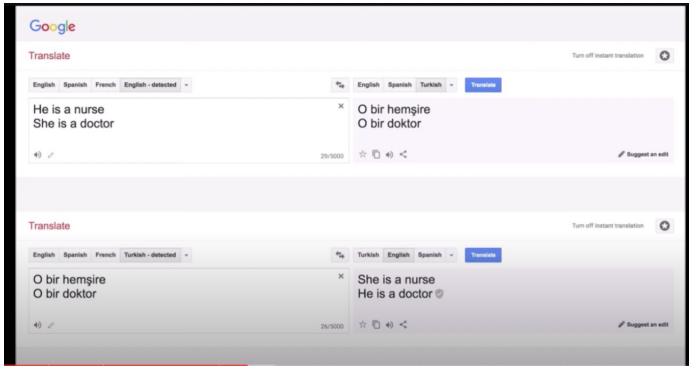
Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she−he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.



Screenshot from NeuRIPS'17 keynote by Kate Crawford

	denigration	stereotype	recognition	under-representation	ex-nomination
Image search for 'CEO' yields all white men on first page of results.			x	x	x
Google Photo mislabels black people as 'gorillas'	x				
YouTube speech-to-text does not recognize women's voices			x		x
HP Cameras' facial recognition unable to recognize Asian people's faces			x	x	x
Amazon labels LGBTQ literature as 'adult content' and removes sales rankings		x	x		x
Word embeddings contain implicit biases [Bolukbasi et al.]	x	x	x	x	x
Searches for African American-sounding names yield ads for criminal background checks [Sweeney]	x	x		x	

Screenshot from NeuRIPS'17 keynote by Kate Crawford

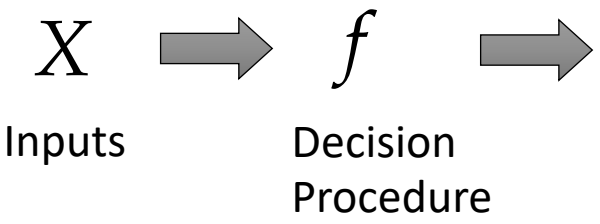
NeurIPS'16

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

How to formalize and measure “Fairness”?

Fair Classification: Think about some intuitive definitions of "fairness"



O = 1 Positive
O = 0 Negative

Sensitive attribute S :
S = 1 protected
S = 0 privileged

X: Features and qualifications: age, hobbies, test scores, grades, etc.

S: Gender

O: Admission Decisions

Other factors:
D = department they applied to
Y = Whether they successfully graduate if they are admitted

Fair Classification

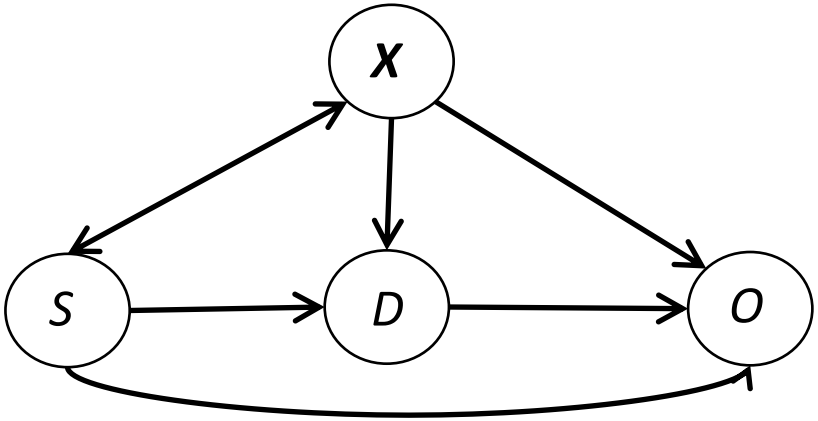
X → f → O

Inputs — Decision Procedure — Outcome

O = 1 Positive
O = 0 Negative

Sensitive attribute S :
S = 1 protected
S = 0 privileged

X: Features and qualifications: age, hobbies, test scores, grades, etc.



S: Gender

O: Admission Decisions

D: choice of dept.

We will revisit these Causal DAGs later

Associational Fairness

Demographic Parity

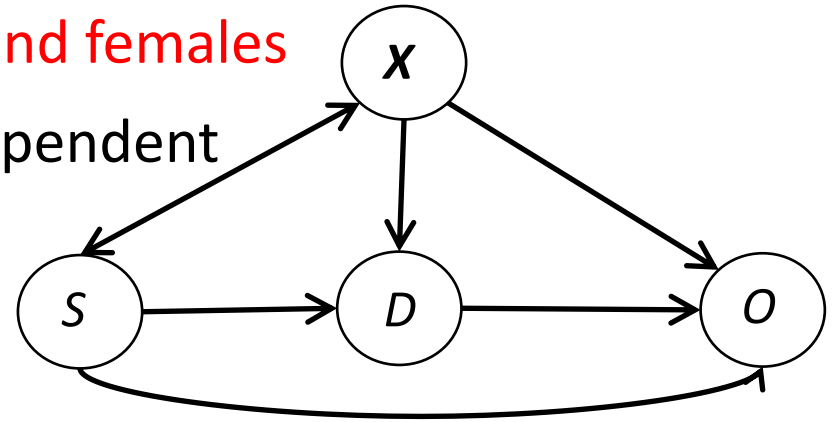
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Associational Fairness

Demographic Parity

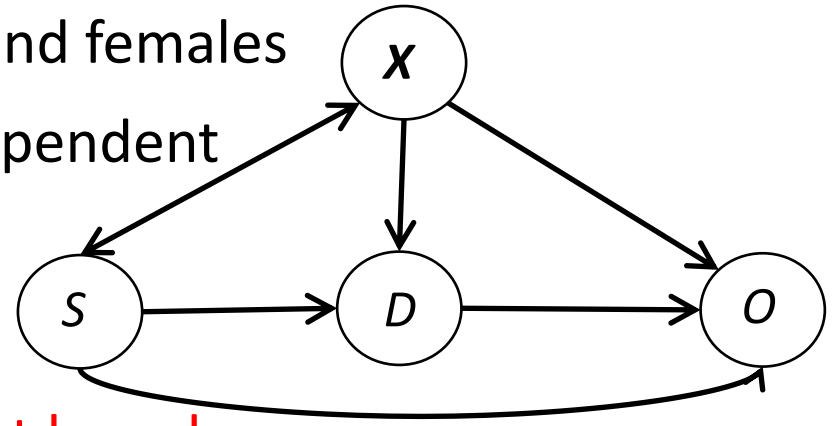
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$O \perp\!\!\!\perp S$



Can it be ensured if decision are not based on S? (Fairness through Blindness/unawreness)

Associational Fairness

Demographic Parity

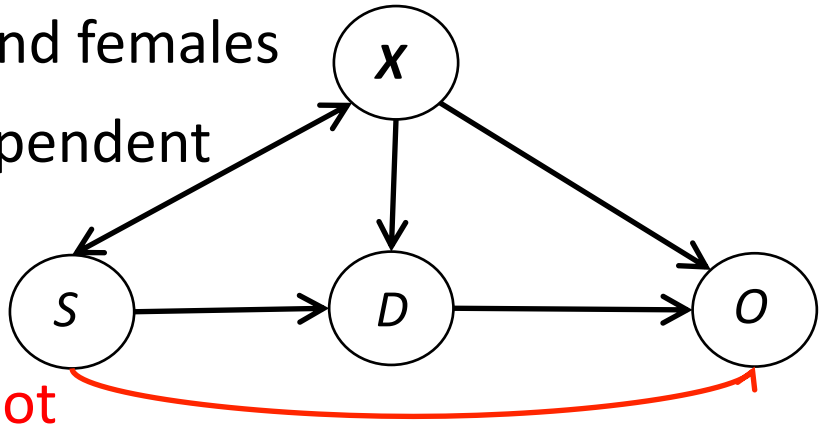
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Can it be ensured if decision are not based on S? (Fairness through Blindness)

Other issues?

Associational Fairness

Demographic Parity

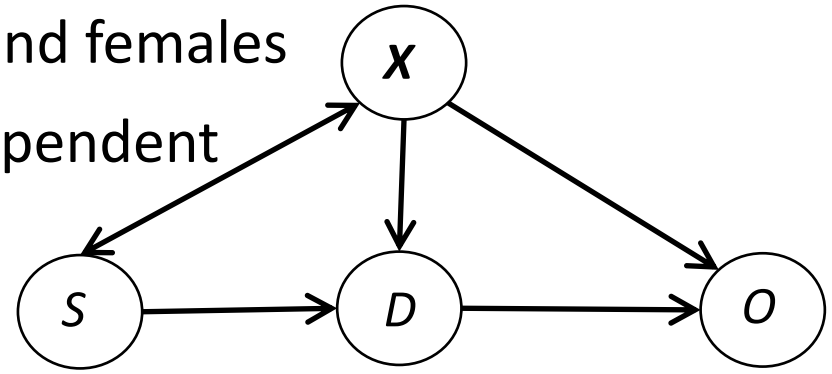
a.k.a. Statistical Parity or Benchmarking

$$\mathbb{P}(O=1|S=1)=\mathbb{P}(O=1|S=0)$$

Same fraction of admitted males and females

S and O should be marginally independent

$O \perp\!\!\!\perp S$



Suppose it happens that one of the S has very high quality applications than the other, or applied to a highly competitive dept

Associational Fairness

Conditional Statistical Parity

Admissible attributes

For any **A=a**

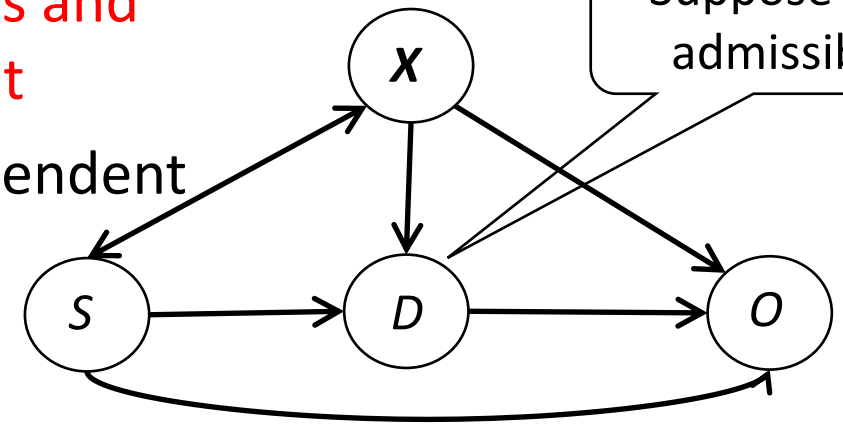
$$\mathbb{P}\{O=1|S=1, \mathbf{A=a}\}=\mathbb{P}\{O=1|S=0, \mathbf{A=a}\}$$

Same fraction of admitted males and females in each department

S and O should be marginally independent conditioned on D

$$O \perp\!\!\!\perp S \mid D$$

Suppose D is admissible



Associational Fairness

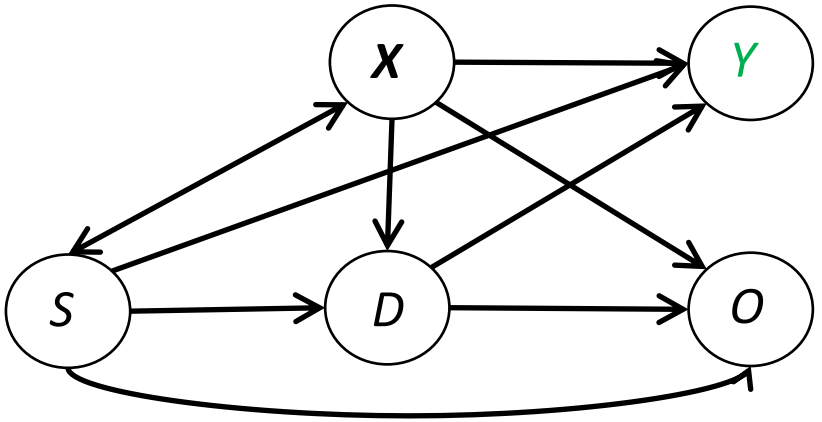
Equalized odds, conditional procedure accuracy equality and disparate mistreatment,

FNR = $\mathbb{P}\{O=0|S=1,Y=1\}=\mathbb{P}\{O=0|S=0,Y=1\}$

FPR = $\mathbb{P}\{O=1|S=1,Y=0\}=\mathbb{P}\{O=1|S=0,Y=0\}$

$O \perp\!\!\!\perp S \mid Y$

Among those applicant who (do not) graduate the rate of admitted students should be independent of applicants' gender.



Y be a binary variable that indicates degree attainment

Associational Fairness

Predictive Parity, Outcome Test or Test-fairness or Calibration

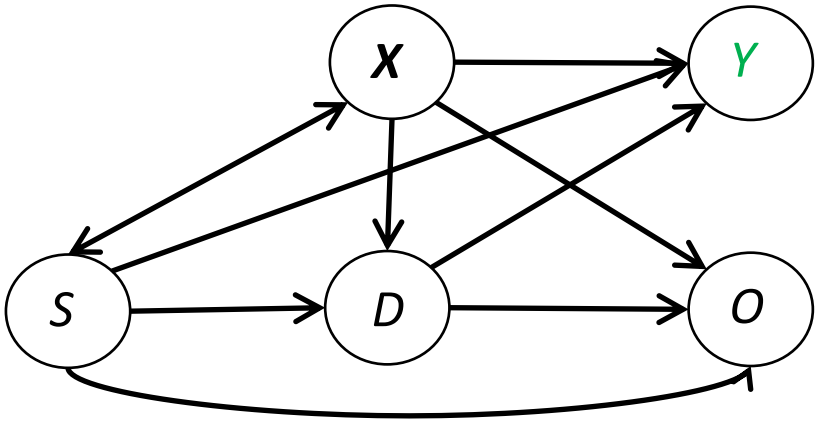
the same predicted positive value (PPV)

$$\mathbb{P}\{Y=1|S=1,O=1\}=\mathbb{P}\{Y=1|S=0,O=1\}$$

$$\mathbb{P}\{Y=1|S=1,O=0\}=\mathbb{P}\{Y=1|S=0,O=0\}$$

$$Y \perp\!\!\!\perp S \mid O$$

Among those applicant that are admitted, the rate of those who attain colleague degree should be the same for males and females



Y be a binary variable that indicates degree attainment

An Associational Debate

FP rate for African-Americans (44.9%)
FP rate for white people (23.5%)
FN rate for whites (47.7%)
FN rate for African-Americans (28.0%)

The likelihood of recidivism among high-risk offenders is the same regardless of race

The COMPAS risk tool is *unfair it violates equalized odds*

The COMPAS risk tool is *fair*. It satisfies predictive parity.

An Associational Debate

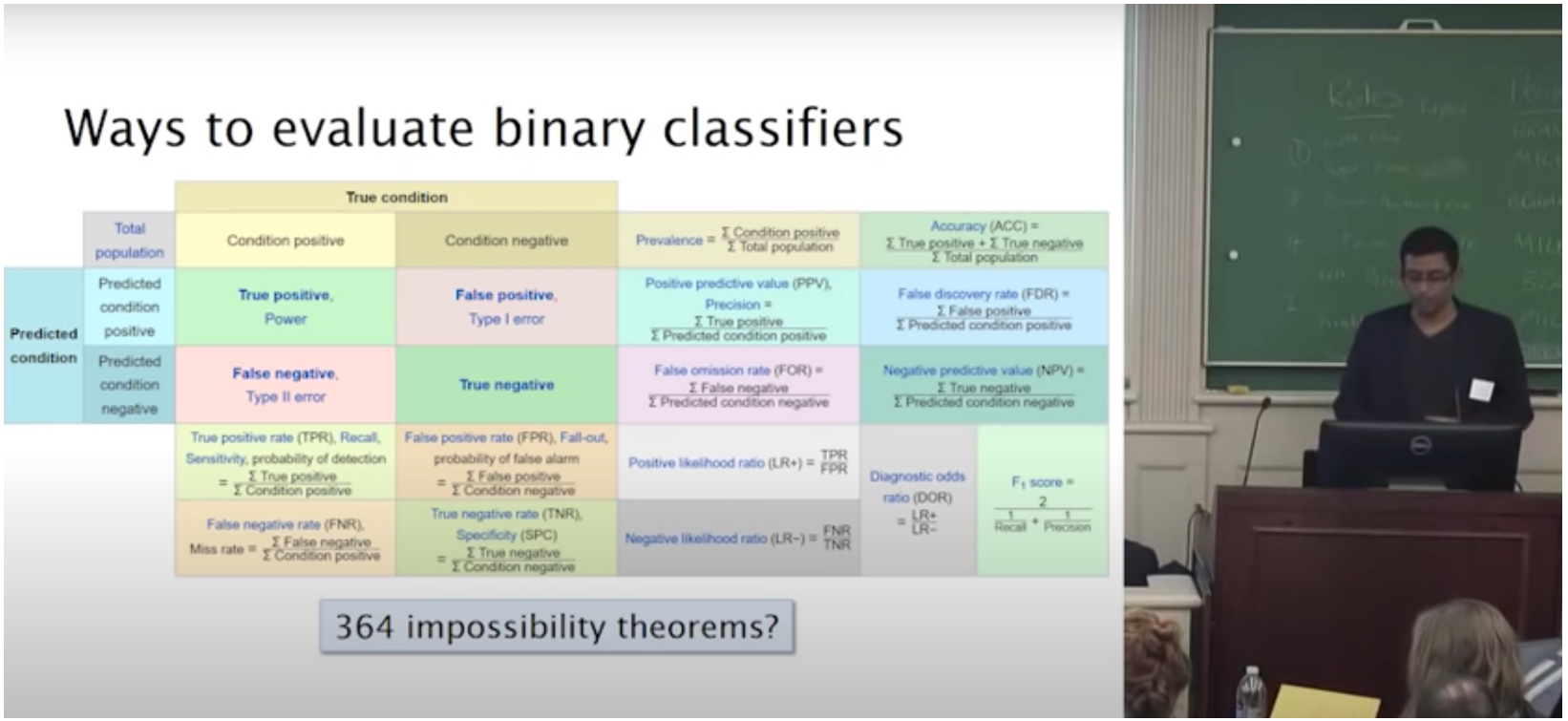
[Chouldechova 16], [Kleinberg, Mullainathan, Raghavan 16]:

"If the base rates differ between two populations,

$$P(Y = 1 \mid S = 0) <> P(Y = 1 \mid S = 1)$$

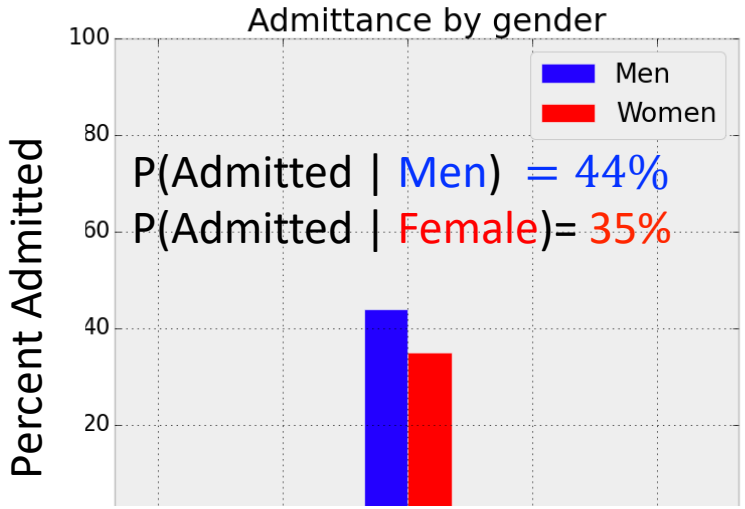
then no non-trivial classifier can simultaneously satisfy equalized odds and predictive parity unless it is perfect (i.e., FPR = FNR = 0)".

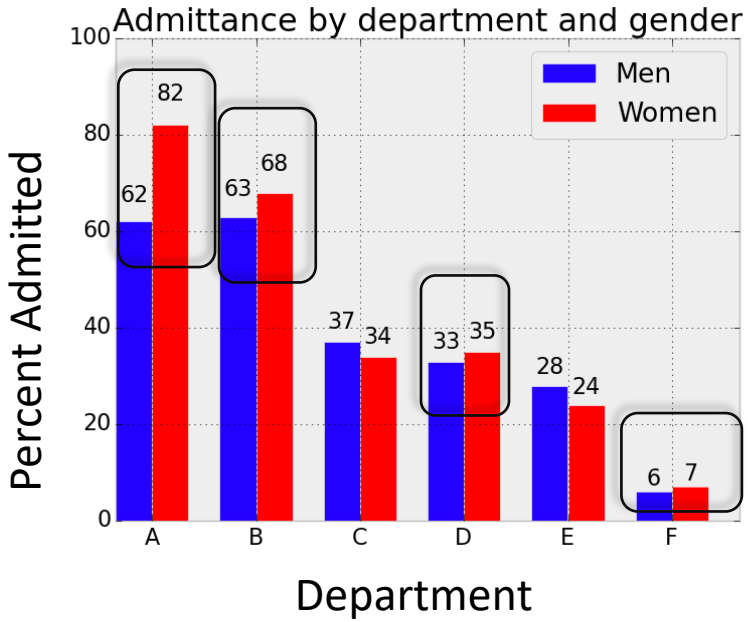
An Associational Debate



Tutorial: 21 fairness definitions and their politics
Arvind Narayanan
https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

An Associational Debate



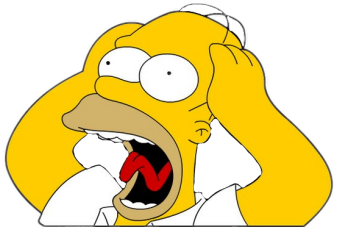


[UC Berkeley 1973 grad school admissions]

Gender is correlated with Admitted

Disparity against females!

Disparity against males!



Discrimination is a causal concept

- Associational notions of fairness are inconsistent and could be misleading
- To prove discrimination, one must show sensitive attribute causes the decisions
- This conception can be traced back to legal systems and literature (The but-for test)
- The but for test broadly asks: **"But for the actions of the defendant (X), would the harm (Y) have occurred?"**

Discrimination in legal system

JUSTIA US Supreme Court

private developers leeway to state and explain the valid interest their policies serve, an analysis that is analogous to Title VII's business necessity standard. It would be paradoxical to construe the FHA to impose onerous costs on actors who encourage revitalizing dilapidated housing in the Nation's cities merely because some other priority might seem preferable. A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity. A robust causality requirement is important in ensuring that defendants do not resort to the use of racial quotas. Courts must therefore examine with care whether a plaintiff has made out a prima facie showing of

source: https://supreme.justia.com/cases/federal/us/576/13-1371

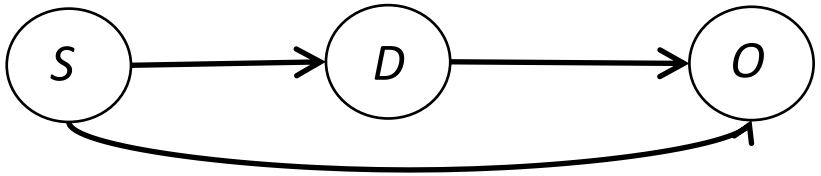
or renovate housing units. And as Judge Jones observed below, if the [plaintiff] cannot show a causal connection between the Department's policy and a disparate impact—for instance, because federal law substantially limits the Department's discretion—that should result in dismissal of this case." *Id.* at 20-21.

source: https://www.jdsupra.com/legalnews/supreme-court-allows-disparate-impact-47404/

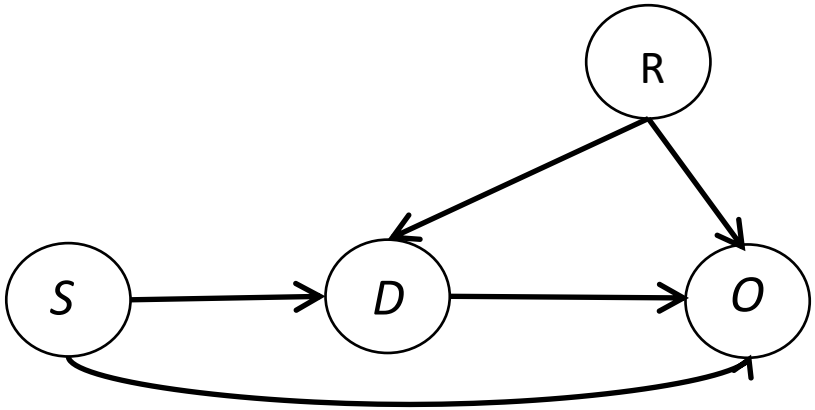
Causal Fairness

Causal DAG: Direct Effect

S = gender
D= department
O = admission decision



Idea 1: ”Disable” all paths between A and Y except for the direct link.
Hold D fixed?
What if another confounder exist? Like state of residence R.
D becomes collider

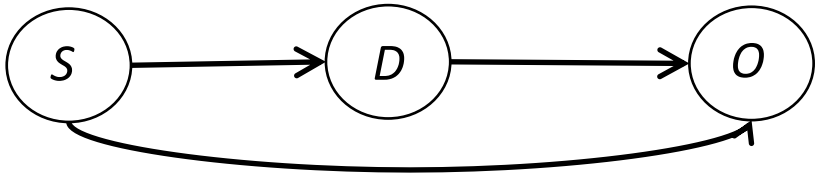


Direct effect = explicit use of S in the decision.
“Blind decision rule”, say does not ask for S.
But still may be discriminatory, in the presence of “proxy” vars in the application

Ref: FairML book, Chapter 4

Causal DAG: Indirect Effect

S = gender
D= department
O = admission decision



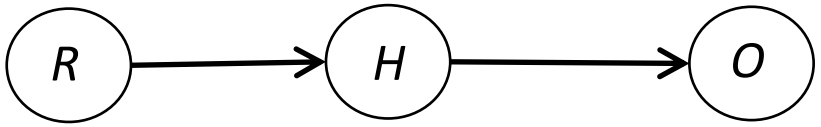
- Indirect effect of gender on admission that goes through department choice.
- Does the indirect path encode a pattern of discrimination?
  - D may be the applicant's inherent department preferences and the department is not responsible for the applicant's preferences. So no discrimination.
  - Not that clear – e.g., there may be Ad may discourage women from applying, a track record of hostile behavior against women, compensating women at a lower rate than equally qualified male students -- all correspond to an indirect effect mediated by department choice.

In general, indirect effects can be estimated only by counterfactuals, not intervention, as direct effects cannot be disabled

Ref: FairML book, Chapter 4

Causal DAG: Indirect Effect

R = Race
H= high-school diploma
O = employment in a high Paying job



To appreciate this point, contrast our Berkeley scenario with the important legal case *Griggs v. Duke Power Co.* that was argued before the U.S. Supreme Court in 1970. Duke Power Company had introduced the requirement of a high school diploma for certain higher paying jobs. We could draw a causal graph for this scenario not unlike the one for the Berkeley case. There's a mediating variable (here, level of education), a sensitive category (here, race) and an employment outcome (here, employment in a higher paying job). The company didn't directly make employment decisions based on race, but rather used the mediating variable. The court ruled that the requirement of a high school diploma was not justified by business necessity, but rather had adverse impact on ethnic minority groups where the prevalence of high school diplomas is lower. Put differently, the court decided that the use of this mediating variable was not an argument against, but rather for discrimination.

Ref: FairML book, Chapter 4 – example and screenshot

Causal Fairness

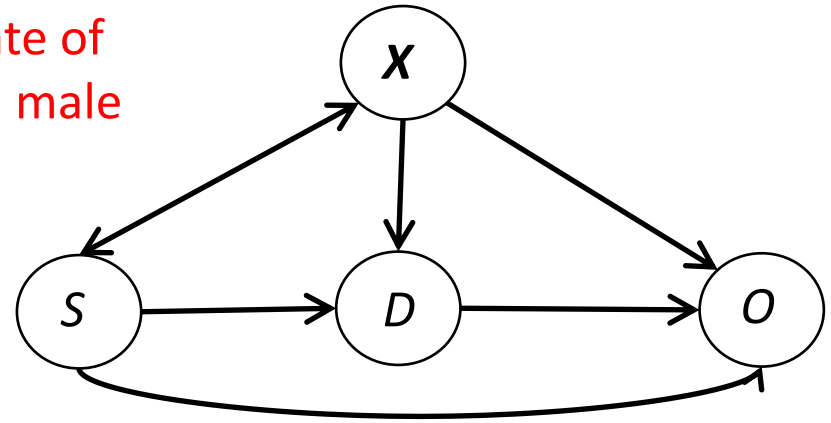
Total Causal Effect Fairness

$$\mathbb{P}(O=1 \mid Do(S=1)) = \mathbb{P}(O=1 \mid Do(S=0))$$

$$\mathbb{P}(O_{S \leftarrow 1} = 1) = \mathbb{P}(O_{S \leftarrow 0} = 1)$$

The rate of admitted students had all students been female should be equal to the rate of admitted student had all students been male

Sufficient Condition:
No causal path from S to O



Causal Fairness

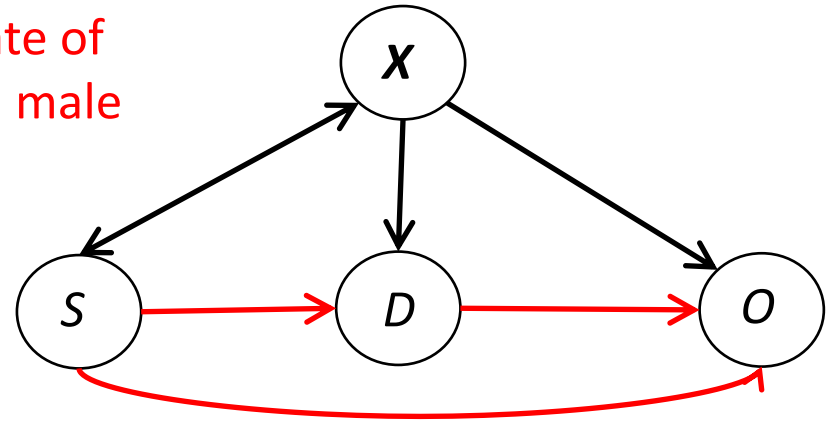
Total Causal Effect Fairness

$$\mathbb{P}(O=1|\ Do(S=1))=\mathbb{P}(O=1\mid Do(S=0))$$

$$\mathbb{P}(O_{S\leftarrow 1}=1)=\mathbb{P}(O_{S\leftarrow 0}=1)$$

The rate of admitted students had all students been female should be equal to the rate of admitted student had all students been male

Sufficient Condition:
No causal path from S to O



1. Direct Path
2. Indirect path (D is a “mediator”)

Causal Fairness

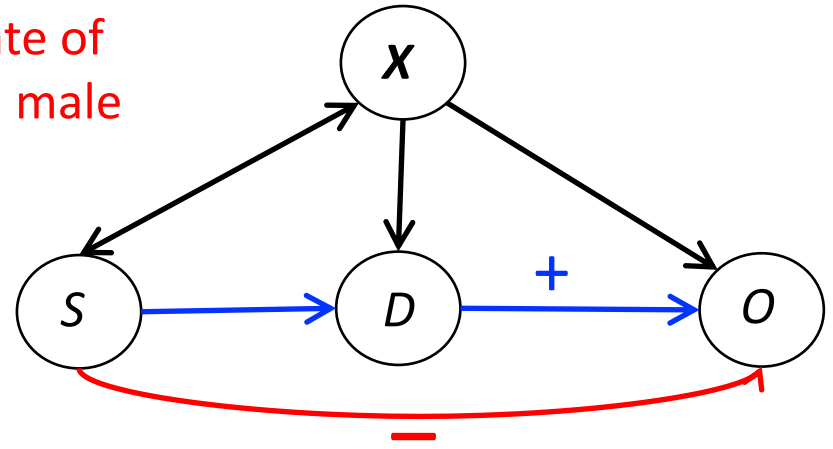
Total Causal Effect Fairness

$$\mathbb{P}(O=1|\ Do(S=1))=\mathbb{P}(O=1\mid Do(S=0))$$

$$\mathbb{P}(O_{S\leftarrow 1}=1)=\mathbb{P}(O_{S\leftarrow 0}=1)$$

The rate of admitted students had all students been female should be equal to the rate of admitted student had all students been male

Sufficient Condition:
No causal path from S to O



Dependence between S and O
= Spurious correlation + Causal effect

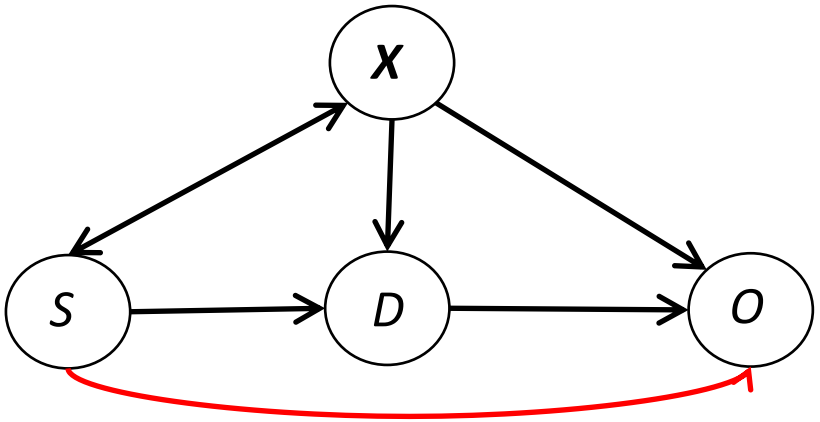
Causal Fairness

Direct Causal Effect Fairness

Total effect = Natural Direct Effect + Natural Indirect Effect

Forbids the natural direct causal effect of S on O

Dependence between S and O
= Spurious correlation + Direct causal effect + Indirect causal effect



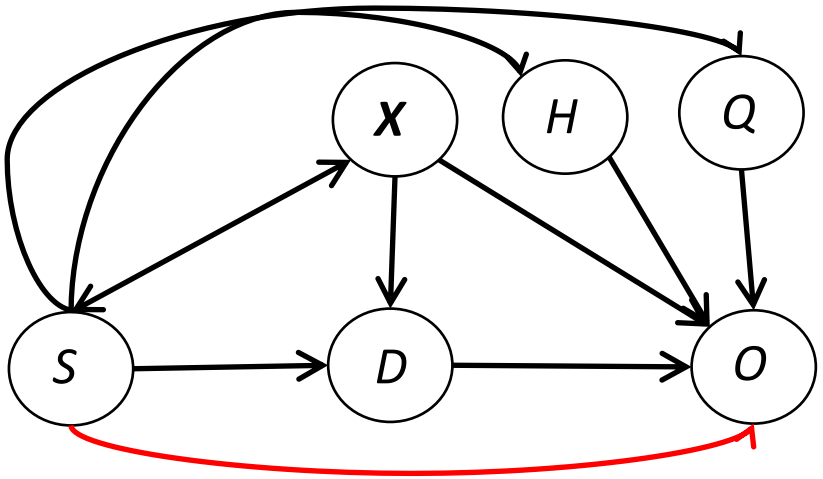
Causal Fairness

Direct Causal Effect Fairness

Total effect = Natural Direct Effect + Natural Indirect Effect

Forbids the natural direct causal effect of S on O

Dependence between S and O
= Spurious correlation + Direct causal effect + Indirect causal effect



All indirect influences of S on O are allowed!

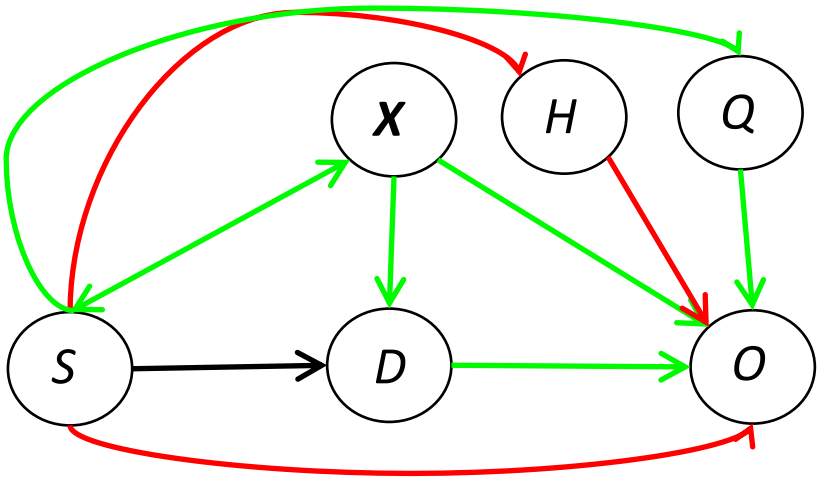
Causal Fairness

Path-Specific Fairness

Partition causal paths from S to O: fair/ discriminatory

S can influence O ONLY through fair causal paths

Red paths are discriminatory



Caveat: It is notoriously difficult to compute path specific effects

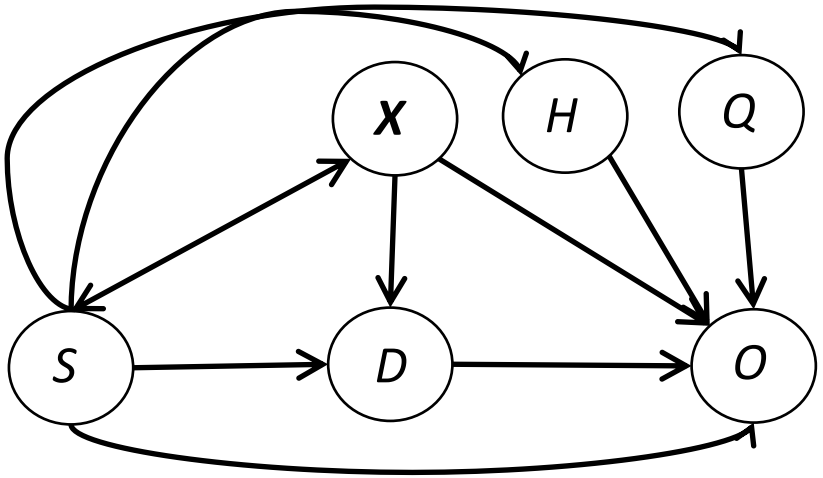
Causal Fairness

Proxy Fairness

[Kilbertus et al. NeurIPS'17]

P is a proxy for S (e.g., hobby?), and may include S

$$\mathbb{P}(O = 1 \mid DO(\mathbf{P}=\mathbf{p})) = \mathbb{P}(O = 1 \mid Do(\mathbf{P}=\mathbf{p}))$$



Causal Fairness

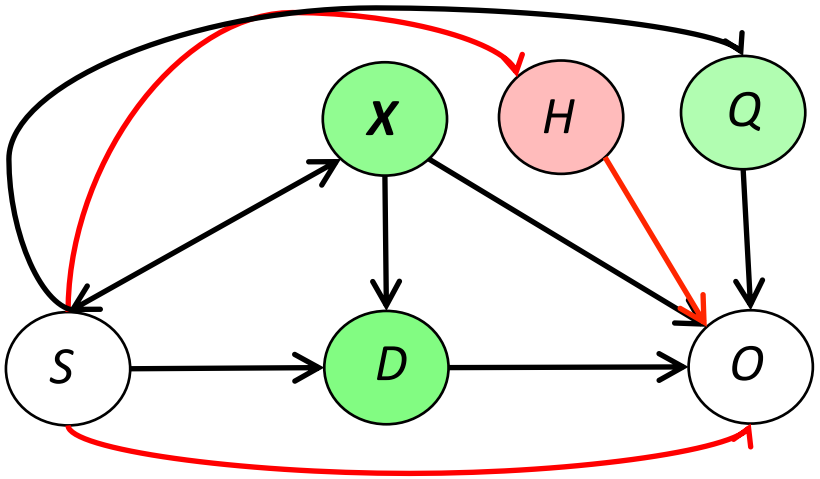
Interventional Fairness [Salimi et al. SIGMOD'19]

Partition variables into: Admissible/ Inadmissible

For any $\mathbf{k} \in \text{Dom}(\mathbf{K})$ and $\mathbf{K} \supseteq \{D, Q, X\}$

$$\mathbb{P}(O \mid DO(S=0), DO(\mathbf{K}=\mathbf{k})) = \mathbb{P}(O \mid Do(S=1), Do(\mathbf{K}=\mathbf{k}))$$

1. It is less expressive than path-specific fairness but easier to compute and enforce
2. captures group-level fairness



Causal Fairness

[Kusner et al. NeuRIPS'17]

Counterfactual Fairness

Total Causal Effect Fairness:

$$\mathbb{P}(O_{S\leftarrow 1}=1)=\mathbb{P}(O_{S\leftarrow 0}=1)$$

U = u: Exogenous variables
X = x: Any context

Gives individual effect: S should not be the cause for any individual instance

$$\mathbb{P}(O_{S\leftarrow 1}(u)=1 \mid X=x, S=1)=\mathbb{P}(O_{S\leftarrow 0}(u)=1 \mid X=x, S=1)$$

$$\mathbb{P}(O_{S\leftarrow 1}(u)=1 \mid X=x, S=0)=\mathbb{P}(O_{S\leftarrow 0}(u)=1 \mid X=x, S=0)$$

Can not be captured using the do-operator

~~$\mathbb{P}(O=1 \mid X=x, do(S=0), S=1)$~~

Causal Fairness

[Pfohl et al. ML for healthcare '19]

Equalized Counterfactual Odds

Equalized Odds:

$$\mathbb{P}\{O=1|S=1,Y=1\}=\mathbb{P}\{O=1|S=0,Y=1\}$$

$$\mathbb{P}\{O=1|S=1,Y=0\}=\mathbb{P}\{O=1|S=0,Y=0\}$$

Before intervention

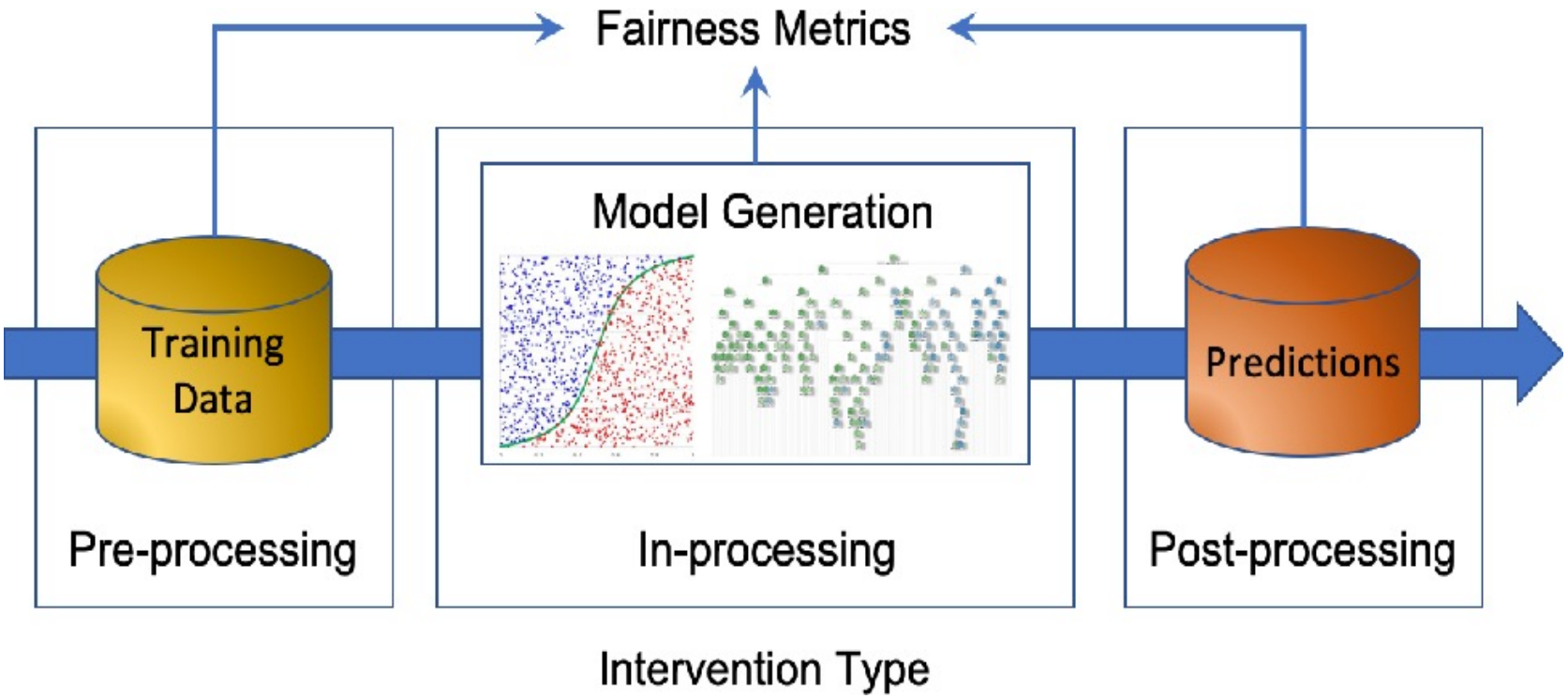
$$\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=1, Y=0)=\mathbb{P}(O_{S\leftarrow 1}=1 \mid S=0, Y=0)$$

After intervention

the predictor is counterfactually fair, conditioned on the factual outcome matching the counterfactual outcome

$$\mathbb{P}(O_{S\leftarrow 1}(u)=1 \mid S=1,X=x, Y_{S\leftarrow 1}=1) = \mathbb{P}(O_{S\leftarrow 0}(u)=1 \mid S=0,X=x, Y_{S\leftarrow 1}=1)$$
$$\mathbb{P}(O_{S\leftarrow 1}(u)=1 \mid S=1,X=x, Y_{S\leftarrow 1}=0) = \mathbb{P}(O_{S\leftarrow 0}(u)=1 \mid S=0,X=x, Y_{S\leftarrow 1}=0)$$

Building Fair Models



Source: Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." *arXiv preprint arXiv:2010.04053* (2020)

Take Aways

- Fairness is causal concept
- One can define a causal counterpart for any existing associational notions of fairness
- Causal reasoning enable disentangling the observed statistical dependence between sensitive attribute and outcome into fine grained causal quantities
- Proving discrimination is as difficult as establishing causation