On Formalizing Fairness in Prediction with Machine Learning

Pratik Gajane, Mykola Pechenizkiy

Jiawei Zhang (jiaweiz7)
2021/11/18

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Question: Why we need to care about fairness…?

Question: Why we need to care about fairness…?

Since it is highly related to our own benefit…

Job application, Addmision to college, Right to vote…

Question: Why we need to care about fairness…?

Since it is highly related to our own benefit…

Job application, Addmision to college, Right to vote…

"I try to be objective. I do not claim to be detached."

——— C. Wright Mills

Question: Why we need to care about fairness…?

Since it is highly related to our own benefit…

Job application, Addmision to college, Right to vote…

"I try to be objective. I do not claim to be detached."

——— C. Wright Mills

But NOT everyone can be so objective and unbiased…

Life is cruel…

Question: Algorithm is OBJECTIVE,
Can we use machine learning to encourage the fairness?

Question: Algorithm is OBJECTIVE,
Can we use machine learning to encourage the fairness?

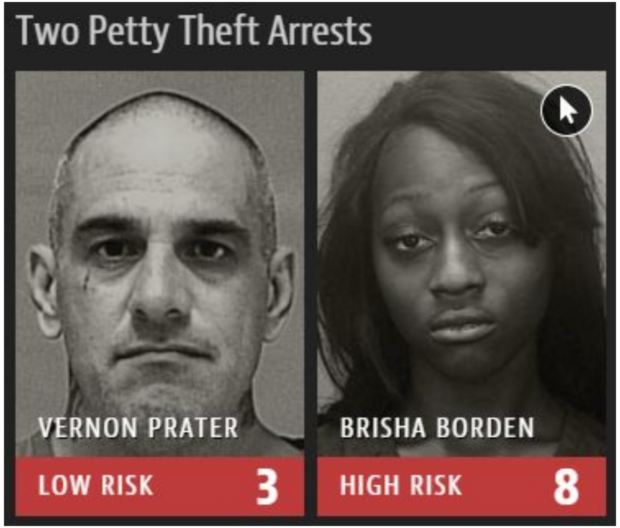
- Courts in United States use COMPAS algorithm for recidivism prediction...
- Amazon uses recommender system to decide the order of items appearing o
  a page...
- Linkedin uses ML to rank job candidates queried...
- etc...

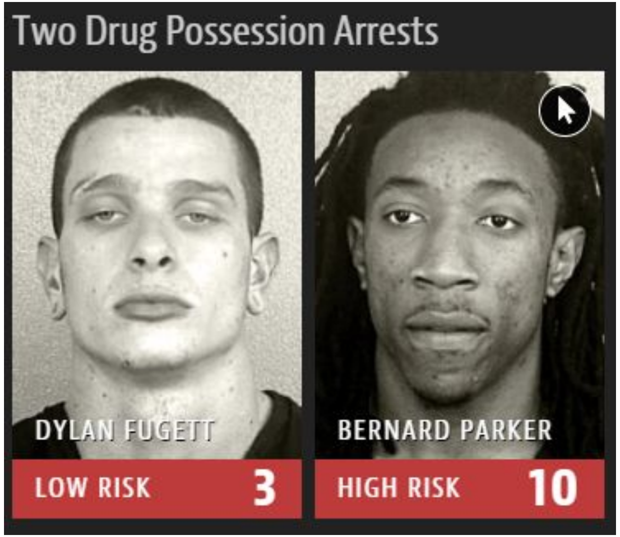
Question: Algorithm is OBJECTIVE,
Can we use machine learning to encourage the fairness?

- Courts in United States use COMPAS algorithm for recidivism prediction...
- Amazon uses recommender system to decide the order of items appearing on a page...
- Linkedin uses ML to rank job candidates queried...
- etc...

Discrimination still EXISTS..

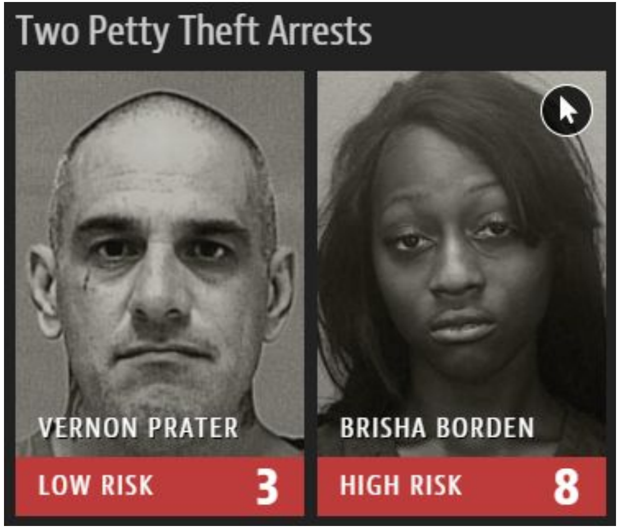
COMPAS for recidivism prediction

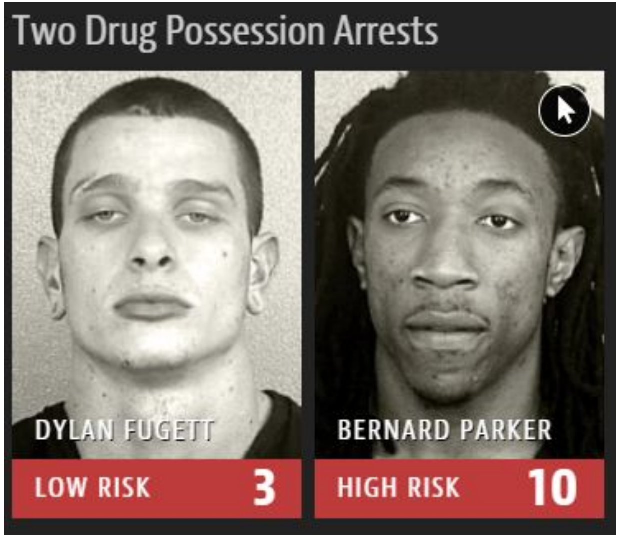




Source: [Larson et al. ProPublica, 2016]

COMPAS for recidivism prediction



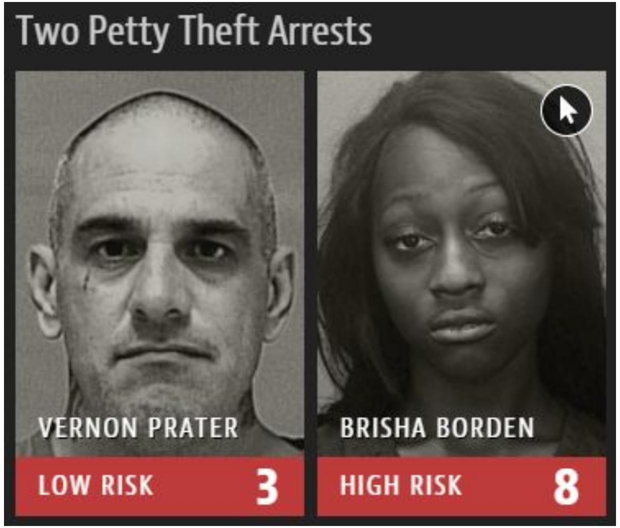


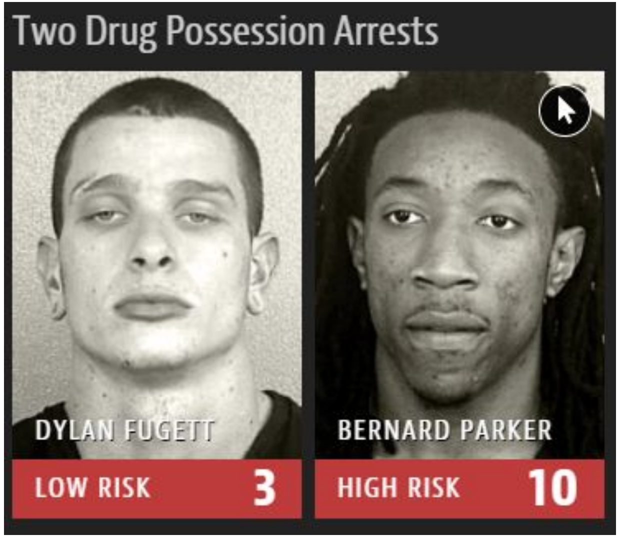
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Source: [Larson et al. ProPublica, 2016]

COMPAS for recidivism prediction

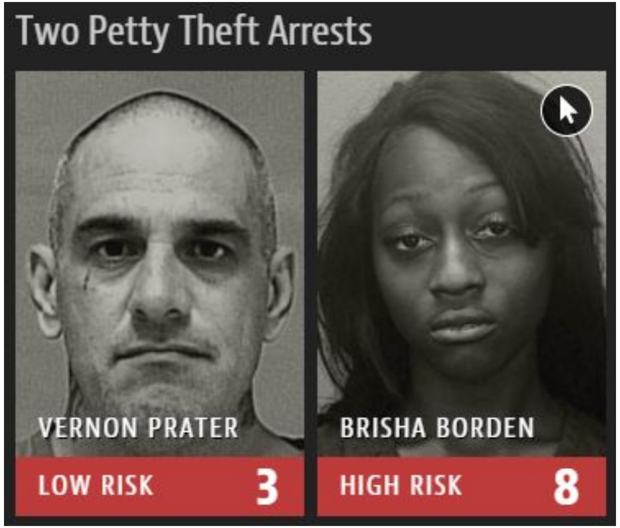


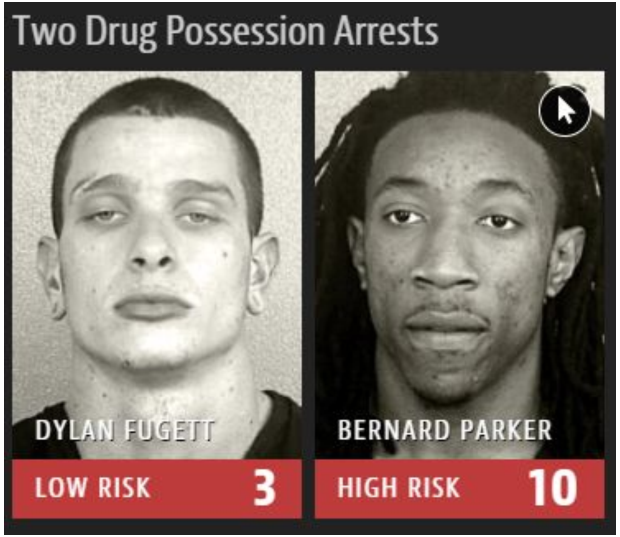


Skewed/Tainted/Limited samples, Sample size disparity...

Source: [Larson et al. ProPublica, 2016]

COMPAS for recidivism prediction





Skewed/Tainted/Limited samples, Sample size disparity...

The training DATA itself is biased,

which reflects the prejudices inherent in our human society.

Source: [Larson et al. ProPublica, 2016]

We need to first formalizing the fairness…

We need to first formalizing the fairness…

- What’s the PROTECTED attributes we care about in fariness?

Legally recognized PROTECTED attributes.

- **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);

…

Source: NIPS 2017 Workshop

We need to first formalizing the fairness...

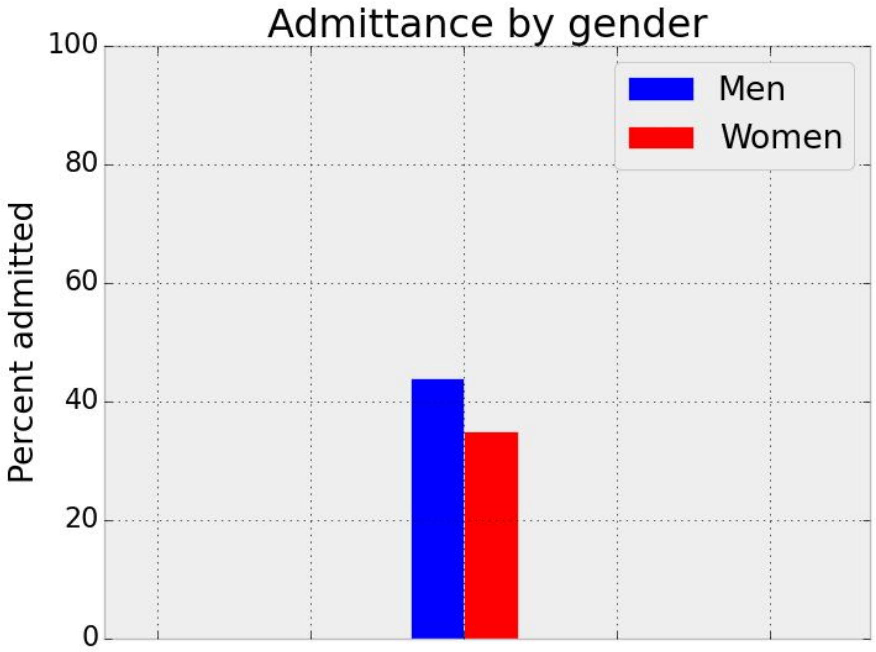
- What's the PROTECTED attributes we care about in fariness?
- How we define FAIRNESS, for group or individual?

Formal setup

- X a set of individuals, i.e., a *population*.
- A the protected attributes, e.g., race, gender.
- Z the remaining attributes, e.g., GPA, GRE score.
- Y the outcome for each individual, e.g., admitted or not.
- $\mathcal{H} : X \to Y$ predictor.
- $\mathcal{H}_S : X \to Y$ group-conditional predictor.

  $S \subset X$ e.g., S represents different race.

Simpson's Paradox

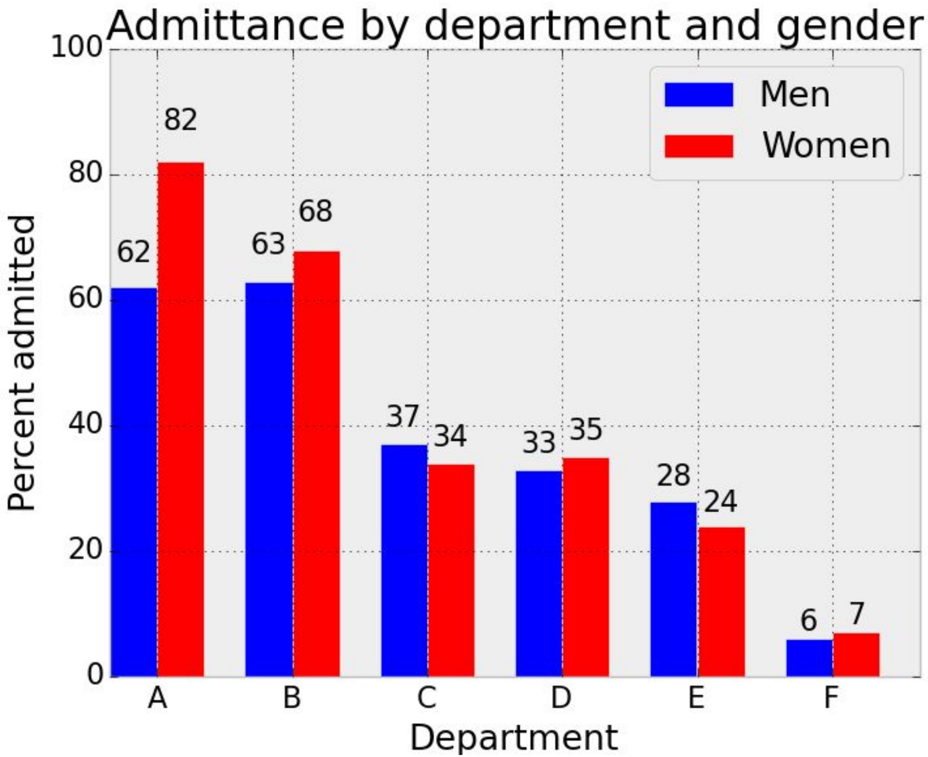


P(Admitted | Men) = 44%

P(Admitted | Woman) = 35%

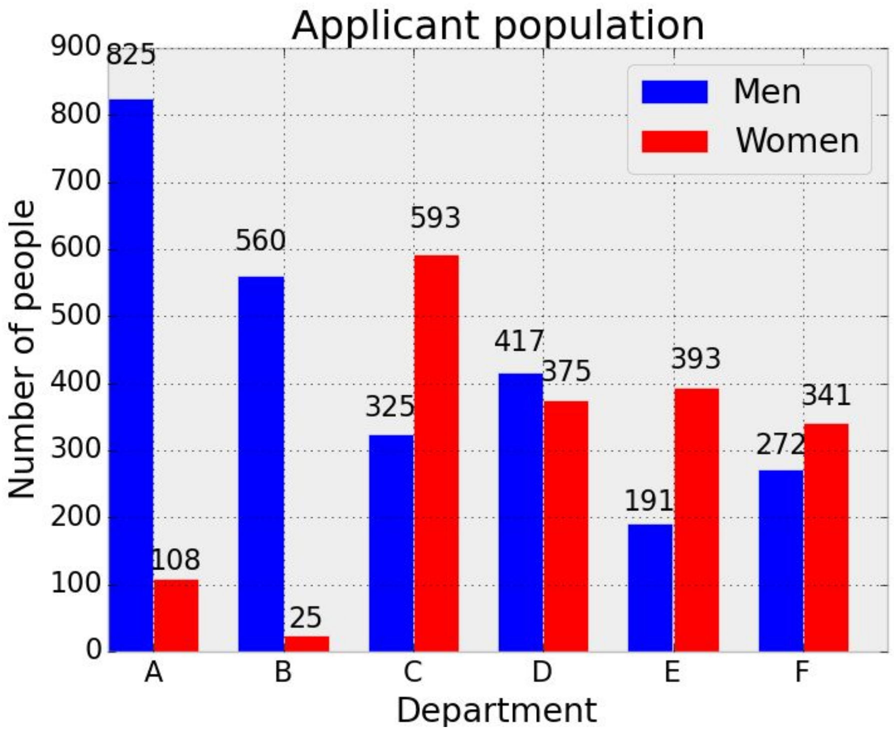
In 1973, UC Berkeley was sued for discrimination against women in graduate school admissions…

Simpson's Paradox



In 4/6 of the departments, a female applicant is more likely to be accepted than a male applicant– the opposite conclusion of Berkeley being biased against females!

Simpson's Paradox



Females applied to more competitive departments than the males did.

=> As a whole, it was more likely that a male applicant would be accepted to Berkeley.

The Simpson's Paradox elucidates the need to be ***skeptical*** of reported statistics that may be extremely dependent upon ***how the data is aggregated.***

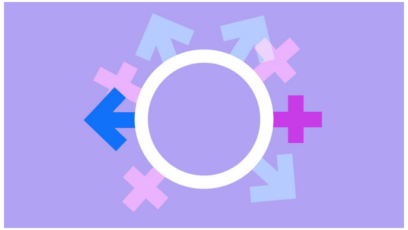
=> Critically think about the definition of fairness later!

P. J. Bickel, E. A. Hammel, J. W. O'Connell. Sex Bias in Graduate Admissions: Data from Berkeley. Science 187, (4175). 1975. pp. 398-404.

Fairness through unawareness

Consider the previous case about getting admitted to college:

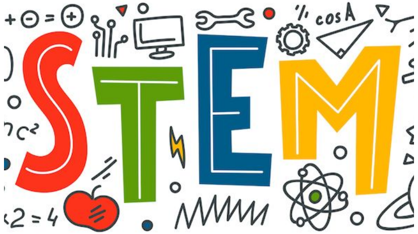
Race



Gender

GPA

Publication



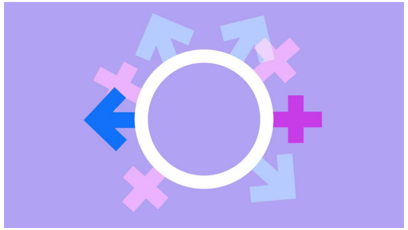
Department

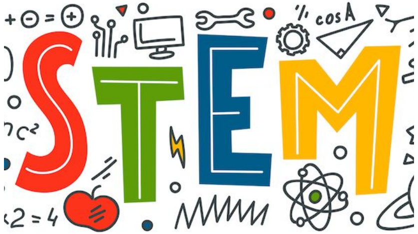
Predictor

Admitted?

Fairness through unawareness

Consider the previous case about getting admitted to college:

Race   Gender   GPA   Publication   Department

PROTECTED !

Predictor

Admitted?

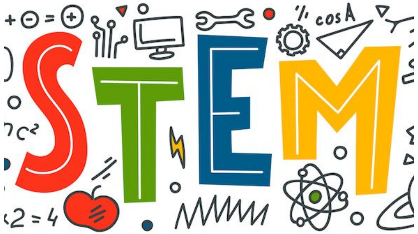
Fairness through unawareness

Consider the previous case about getting admitted to college:

[redacted]

GPA

Publication



Department

PROTECTED !

Predictor

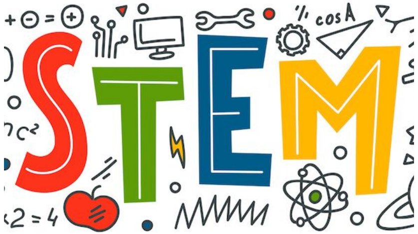
Admitted?

Fairness through unawareness

Consider the previous case about getting admitted to college:

GPA

Publication



Department

PRO

Definition 1: (fairness through unawareness) *A predictor is said to achieve fairness through* ***unawareness*** *if* ***<u>protected attributes are not explicitly used in the prediction process</u>***.

Definition 1: (fairness through unawareness) *A predictor is said to achieve fairness through* ***unawareness*** *if <u>protected attributes are not **explicitly** used in the prediction process</u>.*

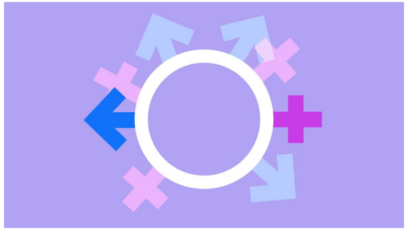
Weakness: the remaining attributes may be highly correlated with the protected attribute…

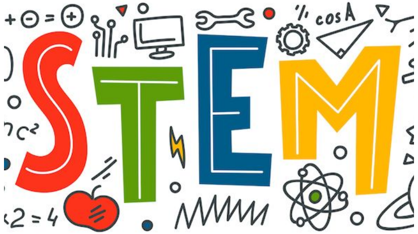
Definition 1: (fairness through unawareness) *A predictor is said to achieve fairness through* ***unawareness*** *if <u>protected attributes are not explicitly used in the prediction process</u>.*

Weakness: the remaining attributes may be highly correlated with the protected attribute...
e.g. the race may influence professors to grade...
the gender may influence the choice of the department for individual...

Group fairness (Statistical/demographic parity)

Consider the previous case about getting admitted to college:

Race

Gender

GPA

Publication

Department

Predictor

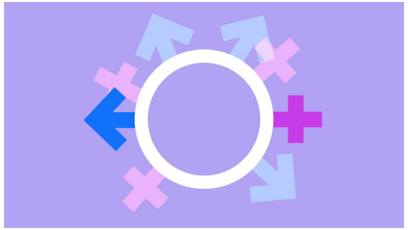
Gender = Male

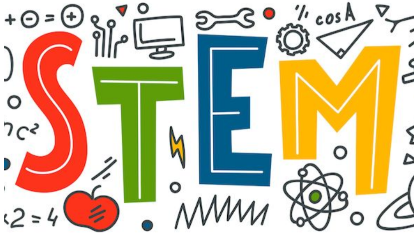
Gender = Female

Same admission rates

Group fairness (Statistical/demographic parity)

Consider the previous case about getting admitted to college:





Race   Gender   GPA   Publication   Department

Definition 2: (Group fairness) A *predictor $H : X \longrightarrow Y$ achieves group fairness with bias ε with respect to groups $S, T \subseteq X$ and $O \subseteq Y$ being any subset of outcomes iff*

$$\left|\mathbb{P}\left\{\mathcal{H}\left(x_i\right) \in O \mid x_i \in S\right\} - \mathbb{P}\left\{\mathcal{H}\left(x_j\right) \in O \mid x_j \in T\right\}\right| \leq \epsilon$$

Definition 2: (Group fairness) A *predictor* $H : X \longrightarrow Y$ *achieves group fairness with bias* ε *with respect to groups* $S, T \subseteq X$ *and* $O \subseteq Y$ *being any subset of outcomes iff*

$$\left|\mathbb{P}\left\{\mathcal{H}\left(x_i\right) \in O \mid x_i \in S\right\} - \mathbb{P}\left\{\mathcal{H}\left(x_j\right) \in O \mid x_j \in T\right\}\right| \leq \epsilon$$

Strength: Follows "four-fifth rule"

Definition 2: (Group fairness) A *predictor* $H : X \to Y$ *achieves group fairness with bias* ε *with respect to groups* $S, T \subseteq X$ *and* $O \subseteq Y$ *being any subset of outcomes iff*

$$|\mathbb{P}\{\mathcal{H}(x_i) \in O \mid x_i \in S\} - \mathbb{P}\{\mathcal{H}(x_j) \in O \mid x_j \in T\}| \leq \epsilon$$

Strength: Follows “four-fifth rule”

Weakness: 1、when P(Y = 1) is not the same for different gender then it rules out the best predictor H = Y

Definition 2: (Group fairness) A *predictor $H : X \to Y$ achieves group fairness with bias ε with respect to groups $S, T \subseteq X$ and $O \subseteq Y$ being any subset of outcomes iff*

$$\left|\mathbb{P}\left\{\mathcal{H}\left(x_i\right) \in O \mid x_i \in S\right\} - \mathbb{P}\left\{\mathcal{H}\left(x_j\right) \in O \mid x_j \in T\right\}\right| \leq \epsilon$$

Strength: Follows "four-fifth rule"

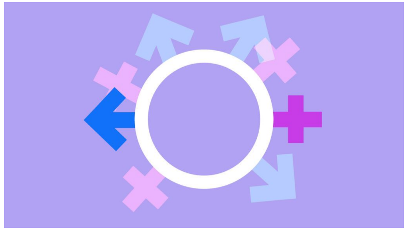
Weakness: 1、when P(Y = 1) is not the same for different gender then it ***rules out the best predictor H = Y***

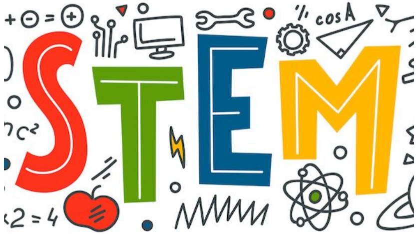
2、we only care about the ***proportion***:

Laziness: we can carefully admit quailified individuals from "female", but randomly admit from "male"…

Equalized odds (not included in paper)

Consider the previous case about getting admitted to college:





Race | Gender | GPA | Publication | Department

Predictor

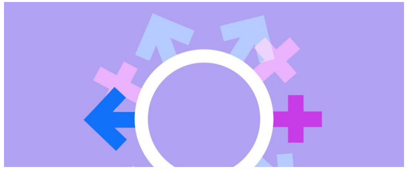
Admitted
(Y = 1)

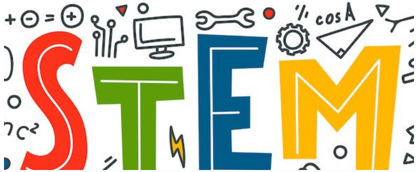
Same Probability for
male and female.

Not Admitted
(Y = 0)

Equalized odds (not included in paper)

Consider the previous case about getting admitted to college:

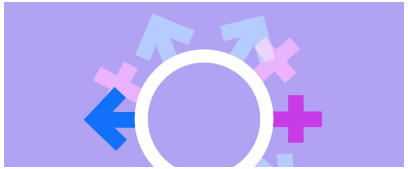
$$P(H=1|A=Male, Y=1) = P(H=1|A=Female, Y=1)$$

$$P(H=0|A=Male, Y=0) = P(H=0|A=Female, Y=0)$$

Definition 3: (Equalized odds) A *predictor* $H: X \rightarrow Y$ *satisfies this definition if the subjects in the protected and unprotected groups have equal true positive rate and equal false positive rate.*

Equalized odds (not included in paper)

Consider the previous case about getting admitted to college:



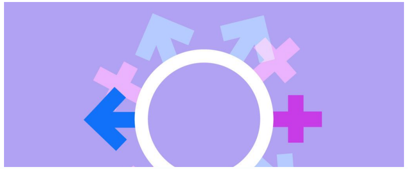
$$\boxed{P(H=1|A=Male,Y=1)=P(H=1|A=Female,Y=1)}$$

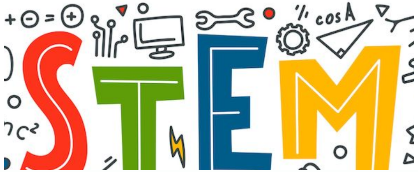
$$P(H=0|A=Male,Y=0)=P(H=0|A=Female,Y=0)$$

Definition 3: (Equalized odds) A *predictor $H : X \rightarrow Y$ satisfies this definition if the subjects in the protected and unprotected groups have* ***equal true positive rate*** *and equal false positive rate.*

Equalized odds (not included in paper)

Consider the previous case about getting admitted to college:





$$\boxed{P(H=1|A=Male,Y=1)=P(H=1|A=Female,Y=1)}$$

$$P(H=0|A=Male,Y=0)=P(H=0|A=Female,Y=0)$$

Definition 4: (Equal opportunity) A predictor is said to satisfy equal opportunity with respect to group S iff

$$\mathbb{P}\left\{\mathcal{H}\left(x_i\right)=1 \mid y_i=1, x_i \in S\right\}=\mathbb{P}\left\{\mathcal{H}\left(x_j\right)=1 \mid y_j=1, x_j \in X \backslash S\right\}$$

Definition 4: (Equal opportunity) A predictor is said to satisfy equal opportunity with respect to group S iff

$$\mathbb{P}\{\mathcal{H}(x_i) = 1 \mid y_i = 1, x_i \in S\} = \mathbb{P}\{\mathcal{H}(x_j) = 1 \mid y_j = 1, x_j \in X \backslash S\}$$

Strength: 1、Allows H = Y

2、Penalize laziness mentioned before

Definition 4: (Equal opportunity) A predictor is said to satisfy equal opportunity with respect to group S iff

$$\mathbb{P}\left\{\mathcal{H}\left(x_i\right)=1 \mid y_i=1, x_i \in S\right\}=\mathbb{P}\left\{\mathcal{H}\left(x_j\right)=1 \mid y_j=1, x_j \in X \backslash S\right\}$$

Strength: 1、Allows H = Y

2、Penalize laziness mentioned before

Weakness: Still may not help closing the gap between two groups...

Admission = 30

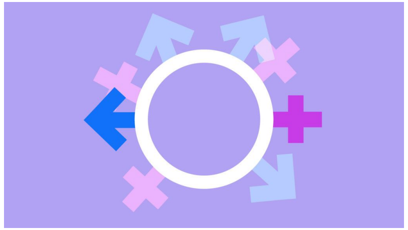
Black: White= 100 : 100

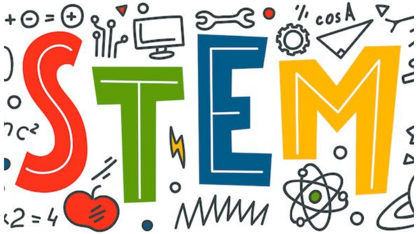
Qualified Black: Qualified White= 2 : 58

Admitted Black: Admitted White= 1 : 29

Individual fairness

Consider the previous case about getting admitted to college:





Race | Gender | GPA | Publication | Department

Definition 5: (Individual fairness) A predictor achieves individual fairness iff $H(x_i) \approx H(x_j) \mid d(x_i, x_i) \approx 0$ where $d : X \times X \rightarrow R$ is a distance metric for individuals.

Definition 5: (Individual fairness) A predictor achieves individual fairness iff $H(x_i) \approxeq H(x_j) \mid d(x_i, x_i) \approxeq 0$ where $d : X \times X \rightarrow R$ is a distance metric for individuals.

the metric is hard to define…

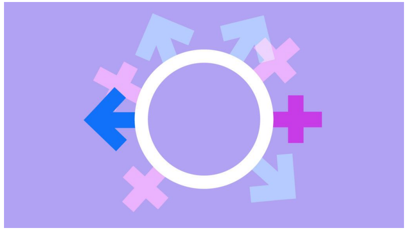
Race	Gender	GPA	Publication	Department
A: White	Male	3.98	None	CS
B: Black	Female	3.85	1 CVPR	CS
C: Black	Male	3.62	2 NIPS	Math

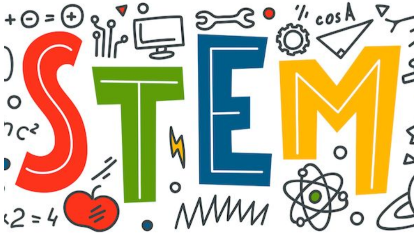
B and A are closer? Or B and C are closer?

How to do quantitative measure…?

Counterfactual fairness

Consider the previous case about getting admitted to college:

Race | Gender | GPA | Publication | Department

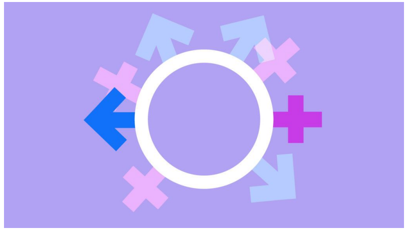
Predictor

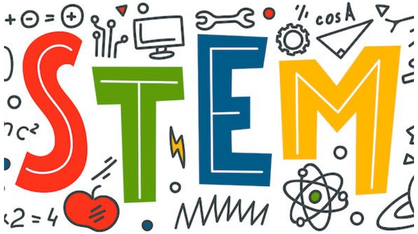
=> Same Probability

Actual world:				
White	Male	3.98	1 CVPR	CS
Counterfactual world:				
Black	Male	3.98	1 CVPR	CS

Counterfactual fairness

Consider the previous case about getting admitted to college:





Race Gender GPA Publication Department

Act

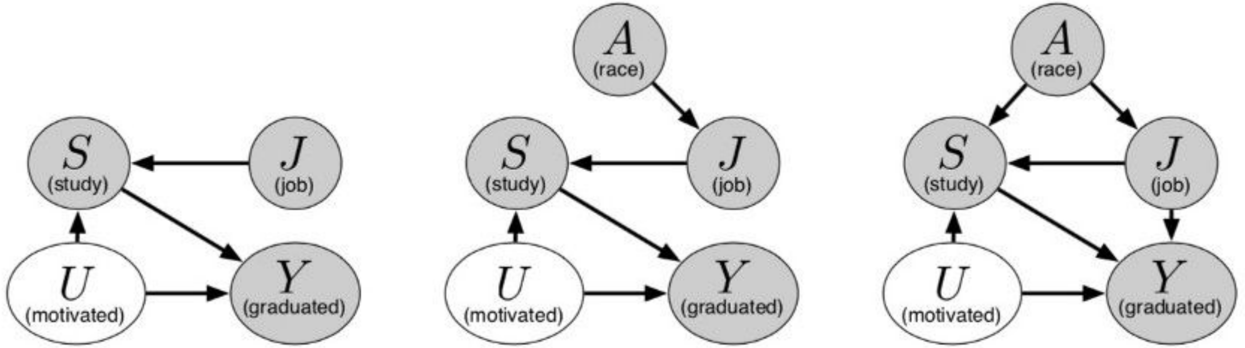
Cou

y

Definition 6: (Counterfactual fairness) A predictor H is counterfactually fair, given Z = z and A = a, for all y and a ≠ a′, iff

$$\mathbb{P}\{\mathcal{H}_{A=a} = y \mid Z = z, A = a\} = \mathbb{P}\{\mathcal{H}_{A=a'} = y \mid Z = z, A = a\}$$

Definition 6: (Counterfactual fairness) A predictor H is counterfactually fair, given Z = z and A = a, for all y and a ≠ a′, iff $\mathbb{P}\{\mathcal{H}_{A=a} = y \mid Z = z, A = a\} = \mathbb{P}\{\mathcal{H}_{A=a'} = y \mid Z = z, A = a\}$



Related with causal graph => not an observational fairness criteria

In practice: hard to decide the graph, hard to decide use which feature

It is possible that a few individuals from a group may ***prefer another outcome*** than the one preferred by the majority of the group. $Y = 1$ is not always the best choice for all group.

It is possible that a few individuals from a group may ***prefer another outcome*** than the one preferred by the majority of the group. Y = 1 is not always the best choice for all group.

Table 1. The surveyed formalizations of fairness

	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

Definition 7: (Preferred treatment) A group-conditional predictor is said to satisfy preferred treatment if each group of the population receives more benefit from their respective predictor then they would have received from any other predictor i.e.

$$\mathbb{B}_S(\mathcal{H}_S) \geq \mathbb{B}_S(\mathcal{H}_T) \quad \text{for all } S, T \subset X$$

Group benefit: the expected proportion of individuals in the group for whom the predictor predicts the beneficial outcome,

i.e., $B = E(P(H_{\text{sub in group}}(S)=\text{beneficial outcome}))$.

$B_{\text{male}}(H_{\text{male}}(\text{male})) >= B_{\text{male}}(H_{\text{female}}(\text{male}))$

Definition 7: (Preferred treatment) A group-conditional predictor is said to satisfy preferred treatment if each group of the population receives more benefit from their respective predictor then they would have received from any other predictor i.e

$$\mathbb{B}_S(\mathcal{H}_S) \geq \mathbb{B}_S(\mathcal{H}_T) \quad \text{for all } S, T \subset X$$

Definition 8: (Preferred impact) A predictor H is said to have preferred impact as compared to another predictor H′ if H offers at-least as much benefit as H′ for all the groups.

$$\mathbb{B}_S(\mathcal{H}) \geq \mathbb{B}_S(\mathcal{H}') \quad \text{for all } S \subset X$$

Prospective notions of fairness

Definition 9: (Equality of resources) Unequal distribution of social benefits is only considered fair when it results from the intentional decisions and actions of the concerned individuals.

Ambition-sensitive: each individual's ambitions and choices that follow them ascertains the benefits they receive.

Endowment-insensitive: each individual's unchosen circumstances including the natural endowments should be offset.

Prospective notions of fairness

Definition 9: (Equality of resources) Unequal distribution of social benefits is only considered fair when it results from the intentional decisions and actions of the concerned individuals.

Definition 10: (Equality of capability of functioning) In order to equalize capabilities of "*being and doing*", people should be compensated for their unequal powers to convert opportunities into functionings. Call for addressing inequalities due to social/ natural endowments(gender/ sex).

Prospective notions of fairness

Definition 9: (Equality of resources) Unequal distribution of social benefits is only considered fair when it results from the intentional decisions and actions of the concerned individuals.

Definition 10: (Equality of capability of functioning) In order to equalize capabilities of "*being and doing*", people should be compensated for their unequal powers to convert opportunities into functionings. Call for addressing inequalities due to social/ natural endowments(gender/ sex).

Too subjective, the metric is still hard to define...

Any question?

We need to first formalizing the fairness…

- What's the PROTECTED attributes we care about in fariness?
- How we define FAIRNESS, for group or individual?
- How we REDUCE such discrimination in ML?

Avoiding Discrimination through Causal Reasoning

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, et al.

Jiawei Zhang
2021/11/18

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Motivation

- Most of these criteria (Demographic Parity/ Equalized odds/ Predictive Rate Parity, etc.) are observational: They depend only on the ***joint distribution*** of predictor, protected attribute, features, and outcome.

Motivation

- Most of these criteria (Demographic Parity/ Equalized odds/ Predictive Rate Parity, etc.) are observational: They depend only on the ***joint distribution*** of predictor, protected attribute, features, and outcome.

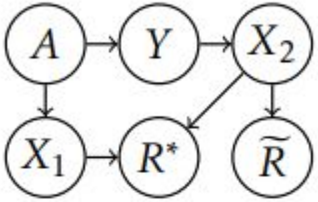


Figure 4: Graphical model for Scenario I.

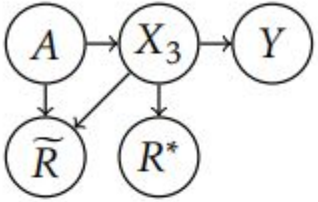


Figure 5: Graphical model for Scenario II.

Intuitively ***different social interpretations*** that admit ***identical joint disributions*** over (predictor, protected attribute, features, outcome).

Source: Moritz et al. Equality of Opportunity in Supervised Learning

Motivation

- Most of these criteria (Demographic Parity/ Equalized odds/ Predictive Rate Parity, etc.) are observational: They depend only on the ***joint distribution*** of predictor, protected attribute, features, and outcome.

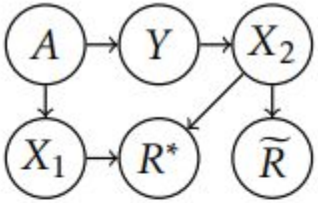


Figure 4: Graphical model for Scenario I.

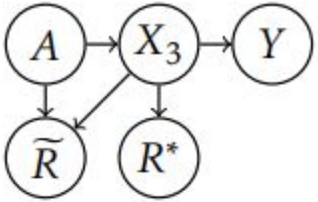


Figure 5: Graphical model for Scenario II.

Intuitively ***different social interpretations*** that admit ***identical joint disributions*** over (predictor, protected attribute, features, outcome).

=> No observational criterion can distinguish them

Source: Moritz et al. [Equality of Opportunity in Supervised Learning]

Motivation

What is the right fairness criterion?

Motivation

What is the right fairness criterion?

Causality,
which has beed mentioned in Counterfactual fairness.

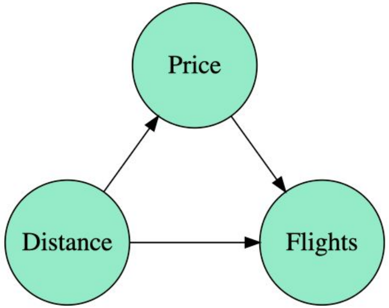
What do we want to assume about our model of the causal data generating process?

New Formal Setup

- A the protected attributes, e.g., race, gender.
- P a set of proxy variables, e.g, name, visual features.
- X features.
- R predictor.
- Y an observed outcome.
- $V_1 \to V_2 \to \ldots \to V_k$ directed path
- $V_1, V_k \notin Z$, if $V_i \in Z$ for some $i \in \{2, \ldots, k-1\}$

  Blocked by a set of nodes Z $V_1 \to Z_1 \to \ldots \to V_k$

Structural Equation Model



- $V_i = f_i\left(pa\left(V_i\right), N_i\right), \text{ for } i \in \{1, \ldots, n\}$

  $pa(V_i)$ are the parents of V_i, i.e., its *direct causes.*

  N_i are independent noise variables.
- Assume ***acyclicity*** => Recursively compute the other variables.
- Model R as a ***childless*** node, whose parents are its input variables.
- Given the noise variables => Entails a ***unique*** joint distribution.
- The ***same joint distribution*** can usually be entailed by ***multiple*** structural equation models, i.e., different causal structures.

Unresolved discrimination

Definition 1: (resolving variable) For any variable in the causal graph that is influenced by A in a manner that we accept as *non-discriminatory*, e.g., the GPA, Publication, Department choice.

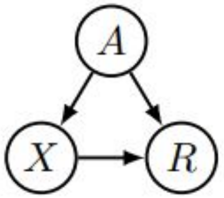
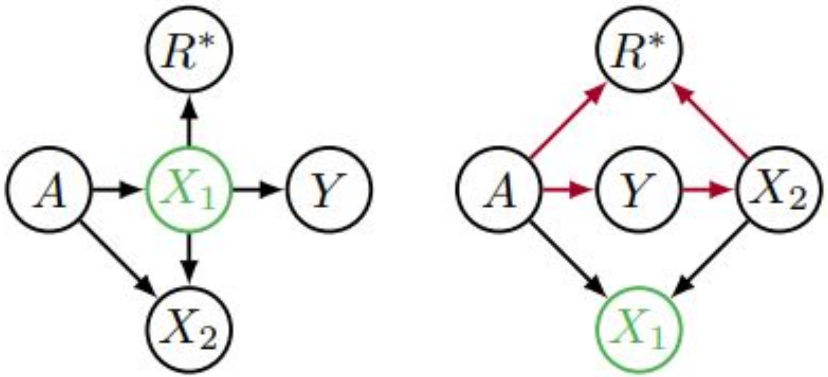


Figure 1: The admission decision R does not only directly depend on gender A, but also on department choice X, which in turn is also affected by gender A.

What matters is the direct effect of the protected attribute (here, gender A) on the decision (here, college admission R) that cannot be ascribed to a resolving variable such as department choice X.

Unresolved discrimination

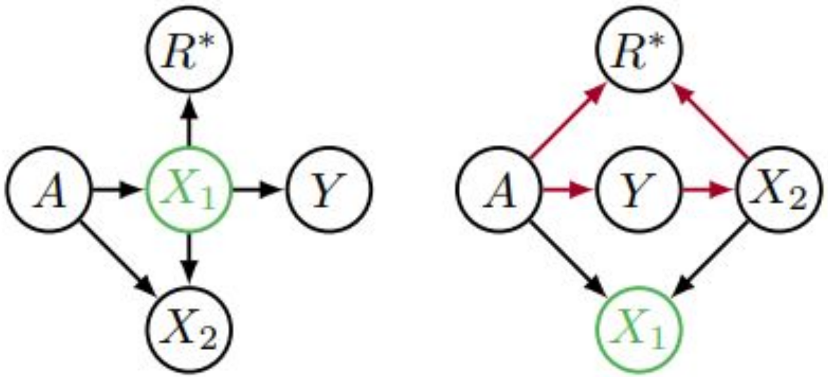
Definition 2: (Unresolved discrimination). A variable V in a causal graph exhibits *unresolved discrimination* if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.



Two graphs that may generate the same joint distribution for the Bayes optimal unconstrained predictor R*. If X1 is a resolving variable, R* exhibits unresolved discrimination in the right graph (along the red paths), but not in the left one.

Unresolved discrimination

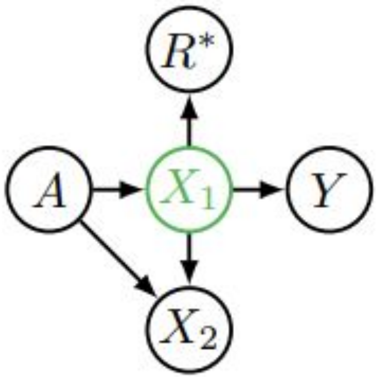
Definition 2: (Unresolved discrimination). A variable V in a causal graph exhibits <u>*unresolved discrimination*</u> if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.

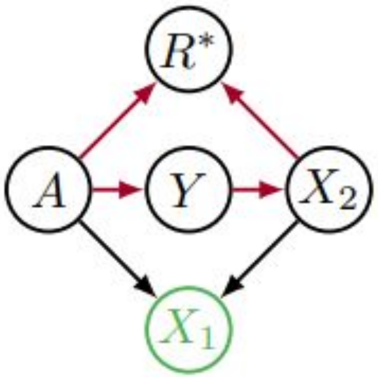


Q1: what if the set of resolving variables is empty?

Unresolved discrimination

Definition 2: (Unresolved discrimination). A variable V in a causal graph exhibits *unresolved discrimination* if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.

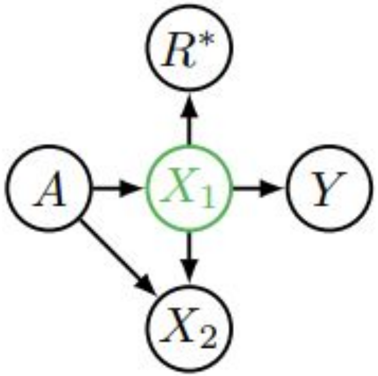


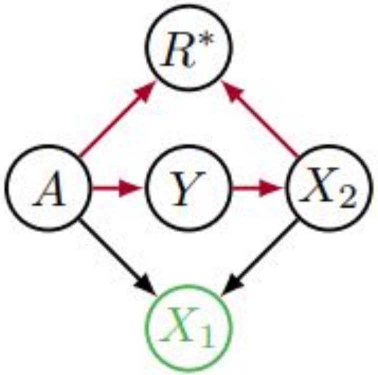


Q1: what if the set of resolving variables is empty?
=> No directed paths from A to R are allowed, get a causal analog of demographic parity.

Unresolved discrimination

Definition 2: (Unresolved discrimination). A variable V in a causal graph exhibits *unresolved discrimination* if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.

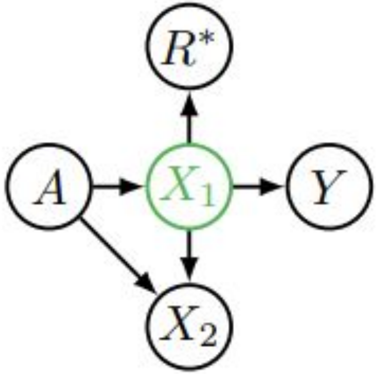


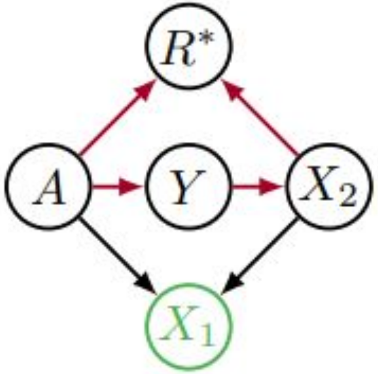


Q2: what if the set of resolving variables is {Y}?

Unresolved discrimination

Definition 2: (Unresolved discrimination). A variable V in a causal graph exhibits *unresolved discrimination* if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.

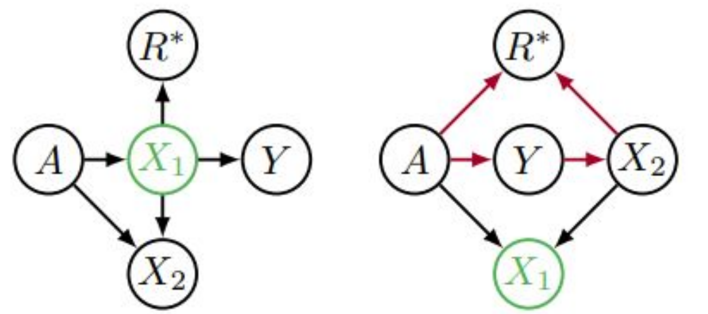




Q2: what if the set of resolving variables is {Y}?
=> A causal analog of equalized odds where strict independence is not necessary.

The limitaion of observational criterion

Theorem 1: Given a joint distribution over the protected attribute A, the true label Y , and some features X1, . . . , Xn, in which we have already specified the resolving variables, ***no observational criterion*** can generally determine whether the Bayes optimal unconstrained predictor or the Bayes optimal equal odds predictor exhibit unresolved discrimination.



Proof omitted.

Potential proxy discrimination

Definition 3: (Potential proxy discrimination). A variable V in a causal graph exhibits ***potential proxy discrimination***, if there exists a directed path from A to V that is blocked by a proxy variable and V itself is not a proxy.

But why we need to care about proxy…?

Potential proxy discrimination

Definition 3: (Potential proxy discrimination). A variable V in a causal graph exhibits ***potential proxy discrimination***, if there exists a directed path from A to V that is blocked by a proxy variable and V itself is not a proxy.

But why we need to care about proxy…?

- Determining causal effects in general requires modeling ***interventions***.
- Interventions on deeply rooted individual properties such as gender or race are notoriously ***difficult to conceptualize***.
- Intervention based on proxy variables(name, visual featurues) poses a ***more manageable*** problem.
- By deciding on a suitable proxy we can find ***an adequate mounting point*** for determining and removing its influence on the prediction. **How?**

Proxy discrimination

Definition 4: (Potential proxy discrimination). A predictor R exhibits no proxy discrimination based on a proxy P if for all p, p′

$$\mathbb{P}(R \mid do(P = p)) = \mathbb{P}\left(R \mid do\left(P = p'\right)\right)$$

Proxy discrimination

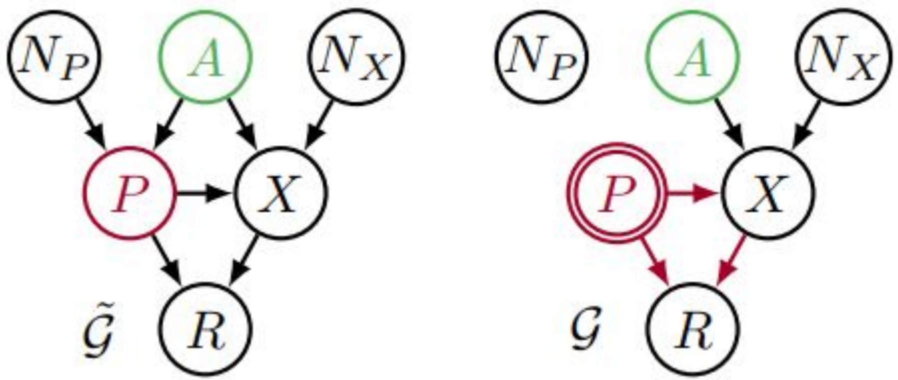
Definition 4: (Potential proxy discrimination). A predictor R exhibits no proxy discrimination based on a proxy P if for all p, p′

$$\mathbb{P}(R \mid do(P = p)) = \mathbb{P}\left(R \mid do\left(P = p'\right)\right)$$

Proposition 1: If there is no directed path from a proxy to a feature, unawareness avoids proxy discrimination.

We are ready to avoid discrimination.

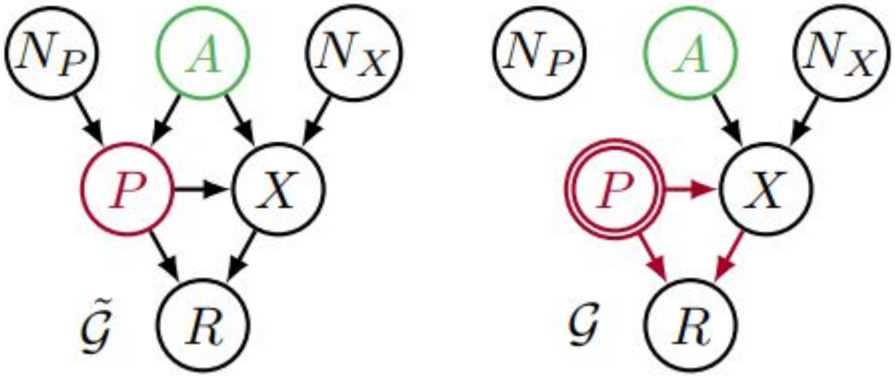
Avoiding proxy discrimination



$$P = \alpha_P A + N_P, \qquad X = \alpha_X A + \beta P + N_X, \qquad R_\theta = \lambda_P P + \lambda_X X$$

We will refer to the ***terminal ancestors*** of a node V in a causal graph D, denoted by *taD(V)*, which are those ancestors of V that are also root nodes of D.

Avoiding proxy discrimination

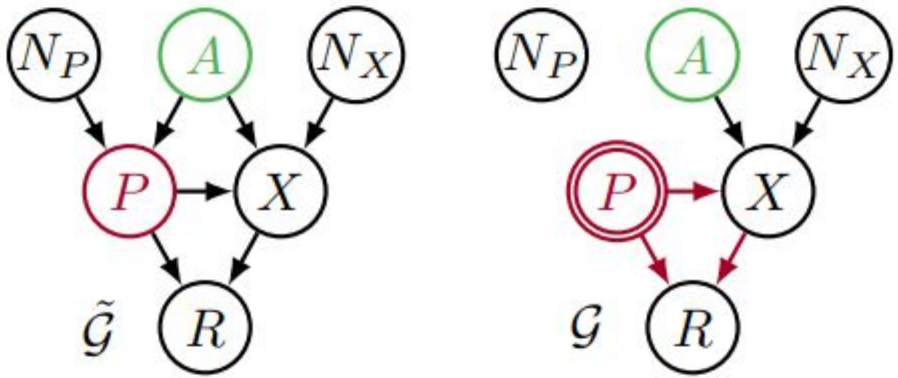


$$P = \alpha_P A + N_P, \qquad X = \alpha_X A + \beta P + N_X, \qquad R_\theta = \lambda_P P + \lambda_X X$$

We will refer to the ***terminal ancestors*** of a node V in a causal graph D, denoted by *taD(V)*, which are those ancestors of V that are also root nodes of D.

benevolent viewpoint: we allow any path from A to R unless it passes through a proxy variable, which we consider worrisome.

Avoiding proxy discrimination



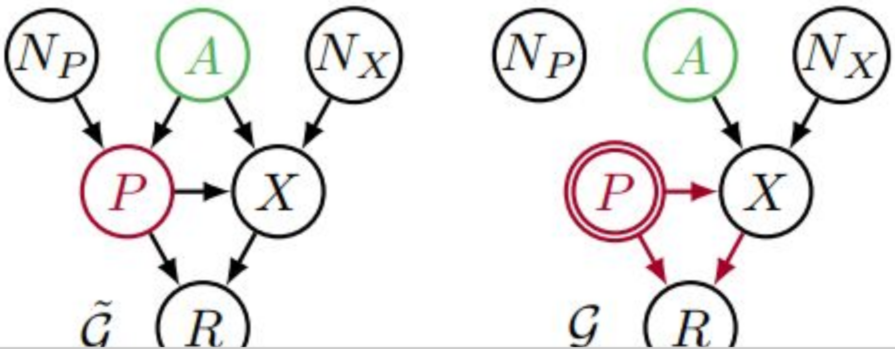
$$P = p, \qquad X = \alpha_X A + \beta P + N_X, \qquad R_\theta = \lambda_P P + \lambda_X X\,. \tag{2}$$

$$R_\theta = (\lambda_P + \lambda_X \beta)p + \lambda_X(\alpha_X A + N_X)\,. \tag{3}$$

$$\mathbb{P}((\lambda_P + \lambda_X \beta)p + \lambda_X(\alpha_X A + N_X)) = \mathbb{P}((\lambda_P + \lambda_X \beta)p' + \lambda_X(\alpha_X A + N_X))\,. \tag{4}$$

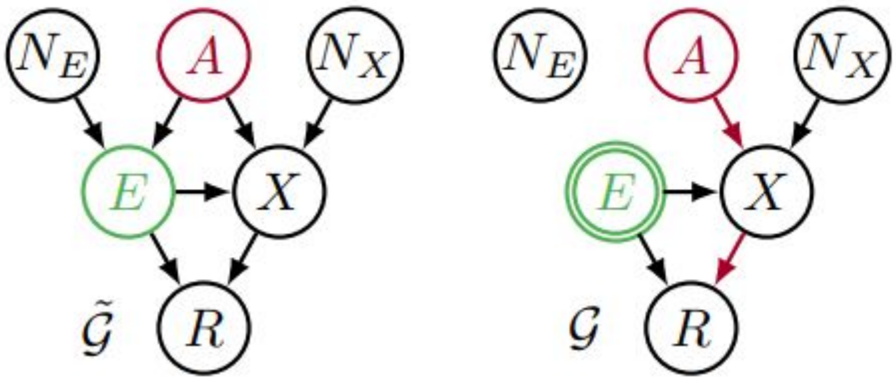
$$R_\theta = -\lambda_X \beta P + \lambda_X X = \lambda_X(X - \beta P)$$

Avoiding proxy discrimination



Proposition 2: If there is a choice of parameters θ_0 such that $R_{\theta_0}(P, X)$ is constant with respect to its first argument and the structural equations are expressible, the before procedure returns a predictor from the given hypothesis class that exhibits no proxy discrimination and is non-trivial in the sense that it can ***make use of features that exhibit potential proxy discrimination***.

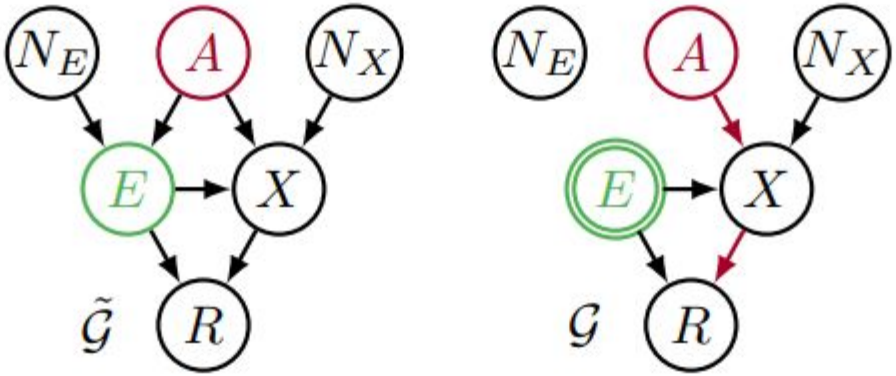
Avoiding unresolved discrimination



$$E = \alpha_E A + N_E, \qquad X = \alpha_X A + \beta E + N_X, \qquad R_\theta = \lambda_E E + \lambda_X X .$$

skeptic viewpoint: all paths from the protected attribute A to R are problematic, unless they are justified by a resolving variable.

Avoiding unresolved discrimination

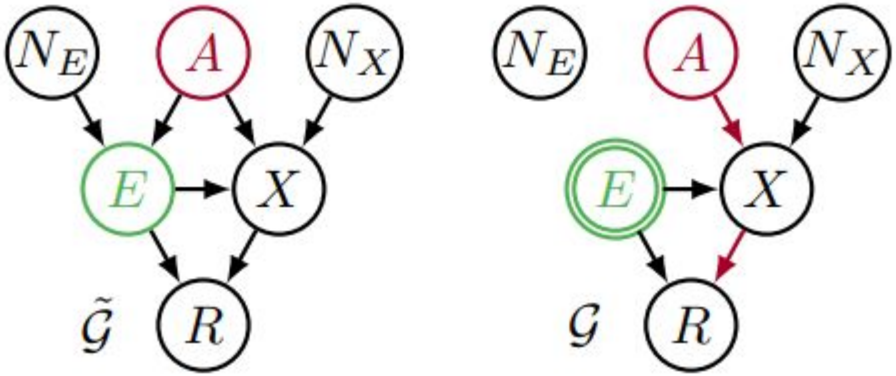


$$E = \eta, \qquad X = \alpha_X A + \beta E + N_X, \qquad R_\theta = \lambda_E E + \lambda_X X \,. \tag{5}$$

$$R_\theta = (\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X A + \lambda_X N_X \,. \tag{6}$$

$$\mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a + \lambda_X N_X)) = \mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a' + \lambda_X N_X)) \,. \tag{7}$$

Avoiding unresolved discrimination



$$E = \eta, \qquad X = \alpha_X A + \beta E + N_X, \qquad R_\theta = \lambda_E E + \lambda_X X \,. \tag{5}$$

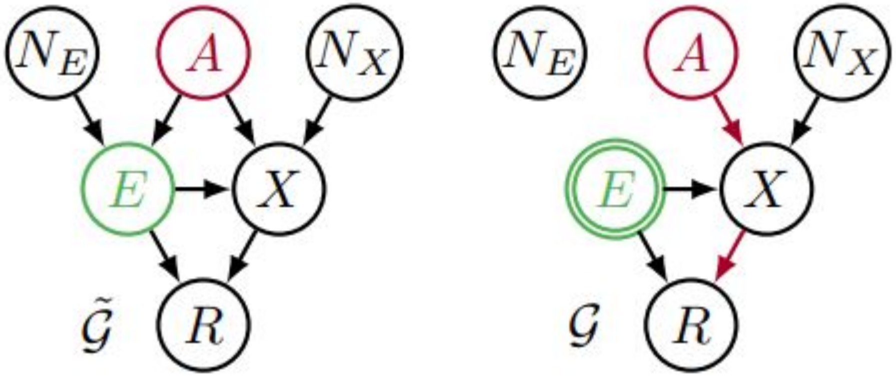
$$R_\theta = (\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X A + \lambda_X N_X \,. \tag{6}$$

$$\mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a + \lambda_X N_X)) = \mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a' + \lambda_X N_X)) \,. \tag{7}$$

Rθ is not explicitly a function of A, we cannot cancel implicit influences of A through X…

$$\lambda_X = 0\ ?$$

Avoiding unresolved discrimination



$$E = \eta, \qquad X = \alpha_X A + \beta E + N_X, \qquad R_\theta = \lambda_E E + \lambda_X X \,. \tag{5}$$

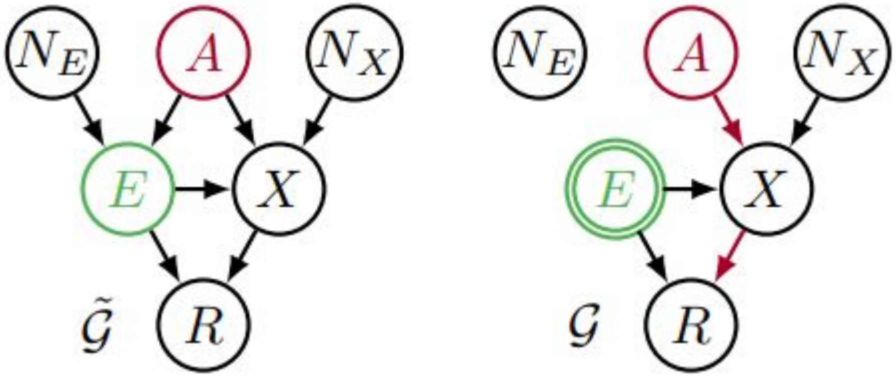
$$R_\theta = (\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X A + \lambda_X N_X \,. \tag{6}$$

$$\mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a + \lambda_X N_X)) = \mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a' + \lambda_X N_X)) \,. \tag{7}$$

Rθ is not explicitly a function of A, we cannot cancel implicit influences of A through X…

$\lambda_X = 0\,?$ ⟶ cancel $A \to E \to X \to R$

Avoiding unresolved discrimination

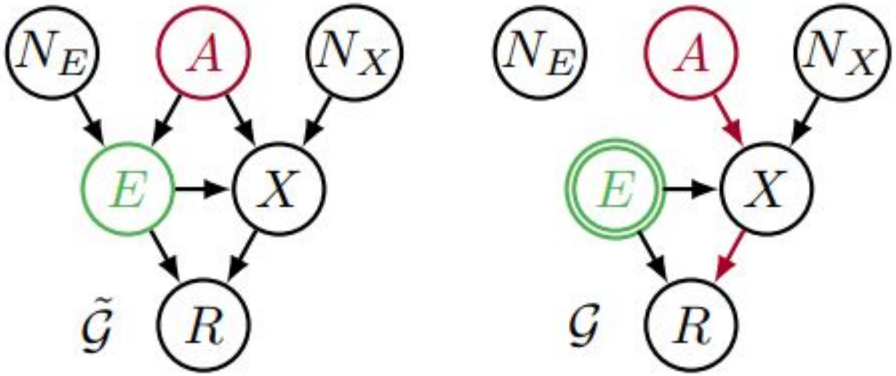


$$E = \eta, \qquad X = \alpha_X A + \beta E + N_X, \quad R_\theta = \lambda_E E + \lambda_X X + \lambda_A A \tag{5}$$

$$R_\theta = (\lambda_E + \lambda_X \beta)\eta + (\lambda_X \alpha_X + \lambda_A)A + \lambda_X N_X$$

$$\lambda_A = -\lambda_X \alpha_X$$

Avoiding unresolved discrimination



$$E = \eta, \qquad X = \alpha_X A + \beta E + N_X, \quad R_\theta = \lambda_E E + \lambda_X X + \lambda_A A \tag{5}$$

$$R_\theta = (\lambda_E + \lambda_X \beta)\eta + (\lambda_X \alpha_X + \lambda_A)A + \lambda_X N_X$$

$$\lambda_A = -\lambda_X \alpha_X$$

In general, if Rθ does not have access to A, we can not adjust for unresolved discrimination without also removing resolved influences from A on Rθ.

Relating proxy discriminations to other notions of fairness

Figure 5: *Left:* A generic graph $\tilde{\mathcal{G}}$ to describe proxy discrimination. *Right:* The graph corresponding to an intervention on P. The circle labeled "DAG" represents any sub-DAG of $\tilde{\mathcal{G}}$ and $\mathcal{G}$ containing an arbitrary number of variables that is compatible with the shown arrows. Dashed arrows can, but do not have to be present in a given scenario.

Relating proxy discriminations to other notions of fairness

Figure 5: *Left:* A generic graph $\tilde{\mathcal{G}}$ to describe proxy discrimination. *Right:* The graph corresponding to an intervention on P. The circle labeled "DAG" represents any sub-DAG of $\tilde{\mathcal{G}}$ and $\mathcal{G}$ containing an arbitrary number of variables that is compatible with the shown arrows. Dashed arrows can, but do not have to be present in a given scenario.

Definition 5: A predictor R exhibits no individual proxy discrimination, if for all x and all p, p′:

$$\mathbb{P}(R \mid do(P = p), X = x) = \mathbb{P}\left(R \mid do\left(P = p'\right), X = x\right)$$

Relating proxy discriminations to other notions of fairness

Definition 5: A predictor R exhibits no individual proxy discrimination, if for all x and all p, p′:

$$\mathbb{P}(R \mid do(P = p), X = x) = \mathbb{P}\left(R \mid do\left(P = p'\right), X = x\right)$$

Definition 6: A predictor R exhibits no proxy discrimination in expectation, if for all p, p′:

$$\mathbb{E}[R \mid do(P = p)] = \mathbb{E}\left[R \mid do\left(P = p'\right)\right]$$

Analysis of proxy discrimination

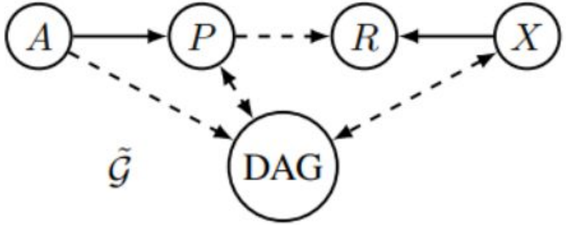
$$P = \hat{f}_P(pa(P))$$

$$X = \hat{f}_X(pa(X)) = f_X\left(P, ta^{\mathcal{G}}(X)\backslash\{P\}\right)$$

$$R = \hat{f}_R(P, X) = f_R\left(P, ta^{\mathcal{G}}(R)\backslash\{P\}\right)$$

$$ta_P^{\mathcal{G}}(X) := ta^{\mathcal{G}}(X)\backslash\{P\}$$

Analysis of proxy discrimination



$$P = \hat{f}_P(pa(P))$$

$$X = \hat{f}_X(pa(X)) = f_X\left(P, ta^{\mathcal{G}}(X)\backslash\{P\}\right)$$

$$R = \hat{f}_R(P, X) = f_R\left(P, ta^{\mathcal{G}}(R)\backslash\{P\}\right)$$

$$ta_P^{\mathcal{G}}(X) := ta^{\mathcal{G}}(X)\backslash\{P\}$$

Theorem 2: Let the influence of P on X be additive and linear, i.e.

$$X = f_X\left(P, ta_P^{\mathcal{G}}(X)\right) = g_X\left(ta_P^{\mathcal{G}}(X)\right) + \mu_X P$$

for some function g_X and $\mu_X \in \mathbb{R}$. Then any predictor of the form

$$R = r(X - \mathbb{E}[X \mid do(P)])$$

for some function *r* exhibits *no proxy discrimination.*

Theorem 2: Let the influence of P on X be additive and linear, i.e.

$$X = f_X\left(P, ta_P^{\mathcal{G}}(X)\right) = g_X\left(ta_P^{\mathcal{G}}(X)\right) + \mu_X P$$

for some function g_X and $\mu_X \in \mathbb{R}$. Then any predictor of the form

$$R = r(X - \mathbb{E}[X \mid do(P)])$$

for some function *r* exhibits *no proxy discrimination.*

Corollary 1. Under the assumptions of Theorem 2, if all directed paths from any ancestor of *P* to *X* in the graph *G* are blocked by *P*, then any predictor based on the adjusted features $\tilde{X} := X - \mathbb{E}[X \mid P]$ exhibits no proxy discrimination and can be learned from the observational distribution $\mathbb{P}(P, X, Y)$ when target labels Y are available.

Proposition 3: Any predictor of the form $R = \lambda(X - \mathbb{E}[X \mid do(P)]) + c$ for $\lambda, c \in \mathbb{R}$ exhibits no proxy discrimination in expectation.

Proposition 3: Any predictor of the form $R = \lambda(X - \mathbb{E}[X \mid do(P)]) + c$ for λ, c ∈ |R exhibits no proxy discrimination in expectation.

From this and the proof of Corollary 1 we conclude the following Corollary:

Corollary 2. If all directed paths from any ancestor of *P* to *X* are blocked by *P*, any predictor of the form $R = r(X - \mathbb{E}[X \mid P])$ for linear *r* exhibits no proxy discrimination in expectation and can be learned from the observational distribution |P(*P*, *X*, *Y*) when target labels *Y* are available.

Conclusion

- the concept of resolving variables and proxy variables.
- the procedure to remove proxy discrimination given linear assumption.

Conclusion

- the concept of resolving variables and proxy variables.
- the procedure to remove proxy/unresolved discrimination given linear assumption.

  LIMITS:
- Stong assumption about we can construct a valid causal graph.
- Most theorems are based on linear case => less expressivity, accuracy?
- Usually, the causal relation is non-linear in ML.
- ....

Any question?