Responsible Data Science

Fairness and Causality

Prof. George Wood

Center for Data Science
New York University

data RESPONSIBLY

NYU

Fairness and causality

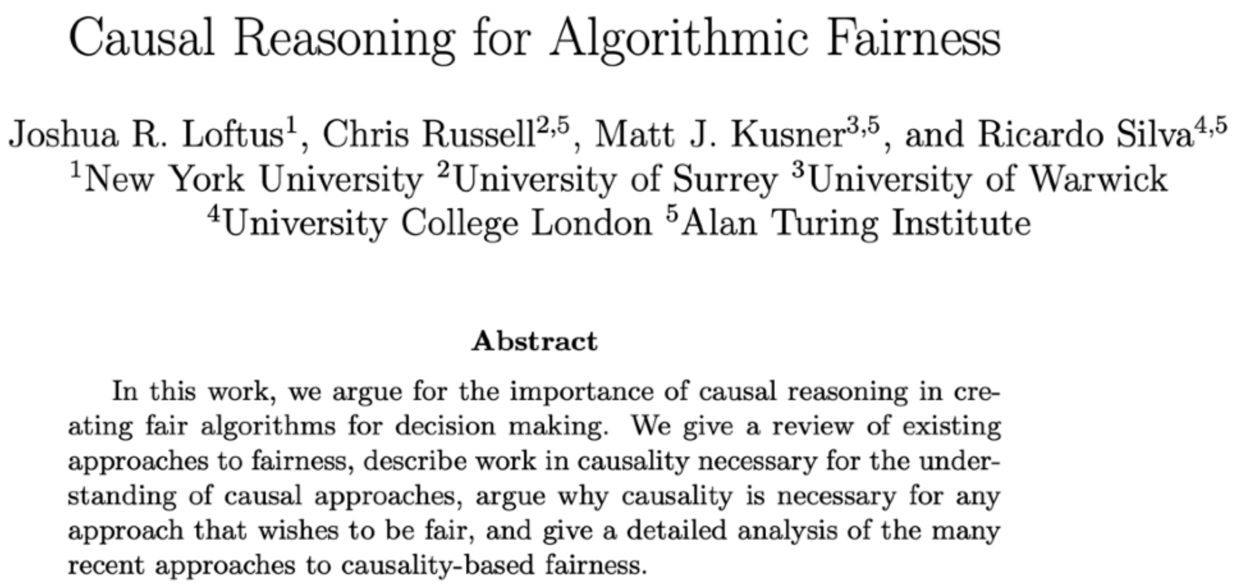
1. Review of fairness measures
2. Causal models
3. Causal models as a framework for fairness

Reading: Fairness and causality

The long road to fairer algorithms

Build models that identify and mitigate the causes of discrimination.

Matt J. Kusner & Joshua R. Loftus

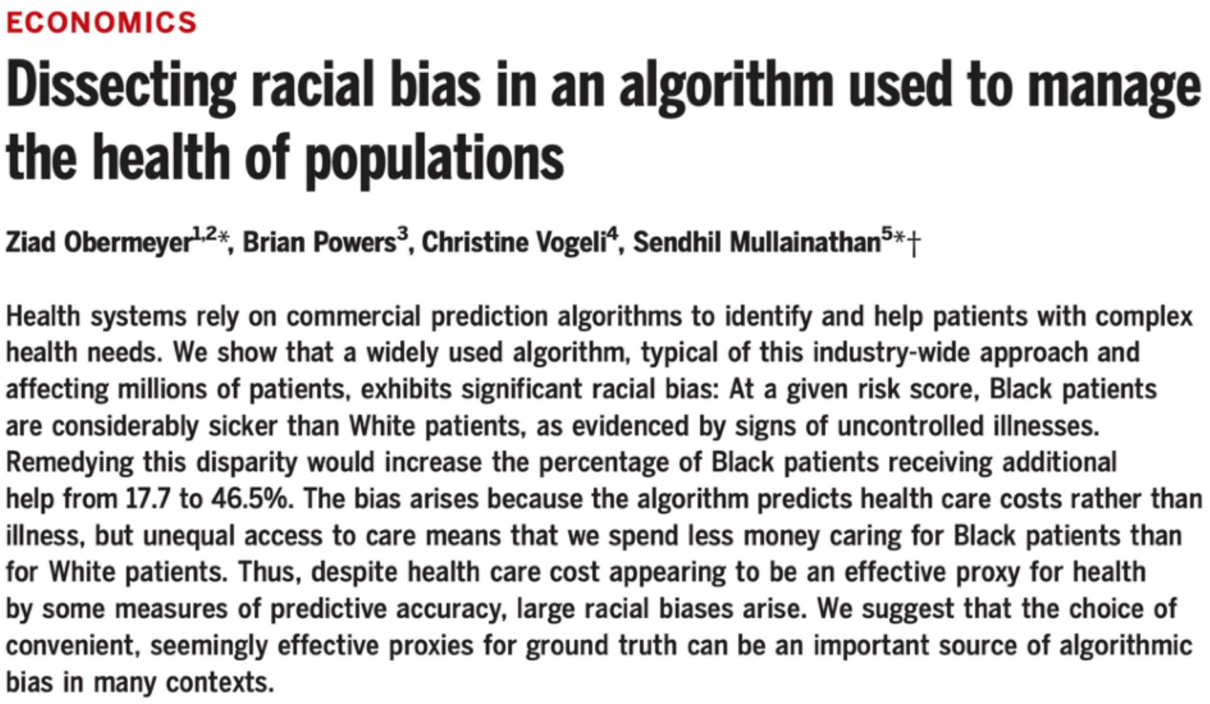


Causal Reasoning for Algorithmic Fairness

Joshua R. Loftus[1], Chris Russell[2,5], Matt J. Kusner[3,5], and Ricardo Silva[4,5]
[1]New York University [2]University of Surrey [3]University of Warwick
[4]University College London [5]Alan Turing Institute

Abstract

In this work, we argue for the importance of causal reasoning in creating fair algorithms for decision making. We give a review of existing approaches to fairness, describe work in causality necessary for the understanding of causal approaches, argue why causality is necessary for any approach that wishes to be fair, and give a detailed analysis of the many recent approaches to causality-based fairness.



ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

fairness measures, a review

Review of fairness measures

- Fairness through unawareness
- Individual fairness
- Demographic parity
- Equalized odds
- Calibration

Review of fairness measures

Notation

A: protected attributes

X: observable attributes

U: unobserved attributes

Y: outcome

Ŷ: predictor (produced by a machine learning algorithm as a prediction of Y)

Capital letters refer to features and lower case letters refer to a value that feature takes

e.g. suppose A is age, then **a** = old and **a**′ = young

Fairness through unawareness

A predictor $\hat{Y}$ satisfies fairness through unawareness if:

$$P(\hat{Y} = y | X = x)$$

- Predictions do not explicitly use protected attributes, **A**

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

Chief Justice Roberts

> "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."
>
> Chief Justice John Roberts (2017)

i.e. fairness through unawareness:

$$P(\hat{Y} = y \mid X = x)$$

- Do not explicitly use protected attributes, **A**

Individual fairness

A predictor $\hat{Y}$ satisfies individual fairness if:

$$P(\hat{Y}^i = y \mid X^i, A^i) \approx P(\hat{Y}^j = y \mid X^j, A^j)$$

When $\mathbf{d(i, j)} \approx 0$. Here, **d** is a task-specific metric that measures the similarity of individuals *i* and *j*.

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Demographic parity

A predictor $\hat{Y}$ satisfies demographic parity if:

$$P(\hat{Y} = y \,|\, A = a) = P(\hat{Y} = y \,|\, A = a')$$

- Predictions are independent of A

If this is not satisfied, we have disparate impact

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Demographic parity

In Lab 2, the predictor $\hat{Y}$ satisfied demographic parity after our in-processing fairness intervention:

$$P(Y \hat{=} y \mid A = \text{young}) = P(Y \hat{=} y \mid A = \text{old})$$

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Equalized odds

A predictor $\hat{Y}$ has equalized odds if:

$$P(\hat{Y} = y \,|\, A = a, Y = y) = P(\hat{Y} = y \,|\, A = a', Y = y)$$

- If a person truly has state **y**, the classifier will predict this at the same rate regardless of the value of **A**

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Equalized odds

The COMPAS predictor $\hat{Y}$ violated equalized odds. Specifically:

$$P(\hat{Y} = y \mid A = \text{Black}, Y = 0) \neq P(\hat{Y} = y \mid A = \text{White}, Y = 0)$$

- The prediction **y** for Black defendants who did not reoffend was higher than for White defendants who did not reoffend.
- Recall: FPR imbalance.

Calibration

A predictor $\hat{Y}$ is calibrated if:

$$P(Y = y \mid A = a, \hat{Y} = y) = P(Y = y \mid A = a', \hat{Y} = y)$$

- If the classifier predicts that a person has state **y**, their probability of actually having state **y** should be the same for all values of **A**

[J. Loftus, C. Russell, M.J. Kusner, R. Silva, arXiv:1805.05859 2018]

Calibration

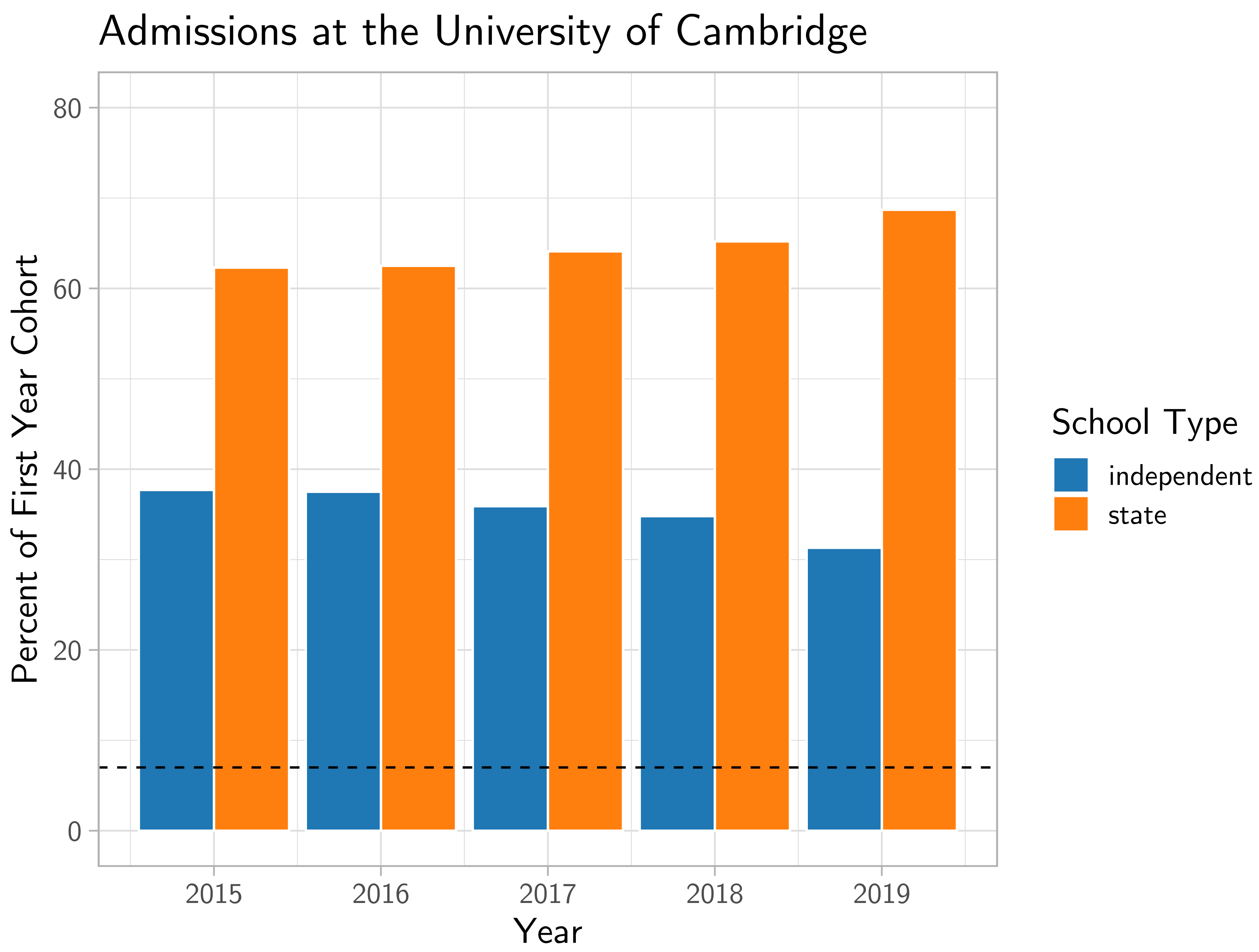
The COMPAS $\hat{Y}$ is calibrated:

$$P(Y = y \mid A = \text{Black}, \hat{Y} = 0.8) = P(Y = y \mid A = \text{White}, \hat{Y} = 0.8)$$

- This sounds similar to equalized odds. But they are fundamentally incompatible
- In nearly all real cases, we cannot satisfy calibration **and** equalized odds at the same time

causal models

What is a causal model?



What is a causal model?

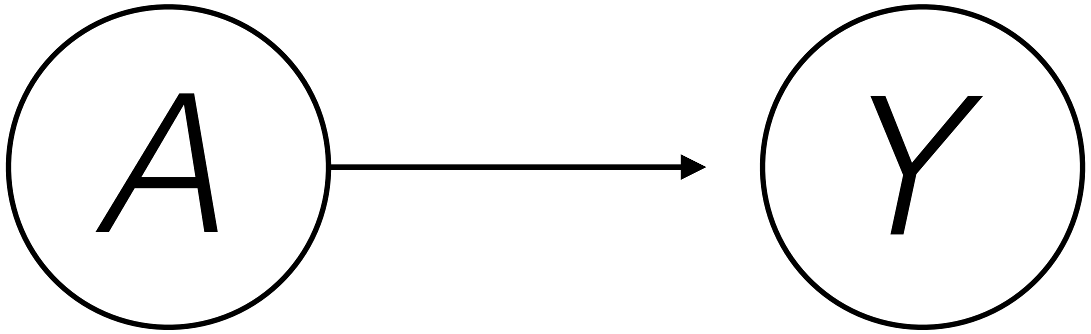
A, **Intervention**:

Student attends an independent school

→

Y, **Outcome**:

Student gets a place at Cambridge



We can represent this causal structure using a Directed Acyclic Graph (DAG)

Cause and counterfactuals

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well.

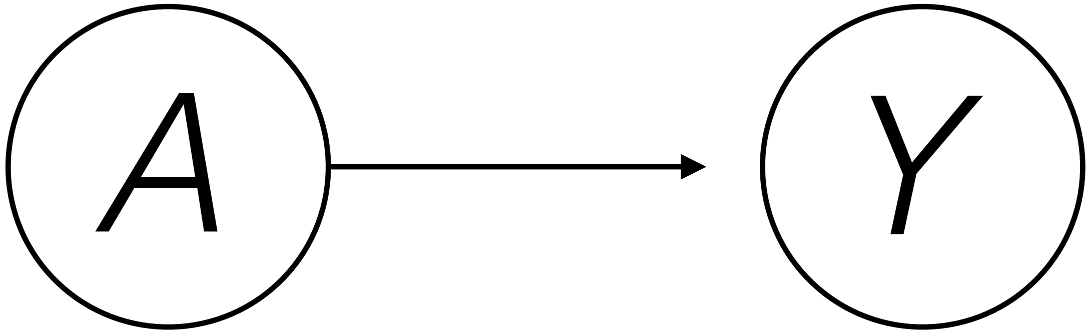
David Lewis, *Journal of Philosophy* (1973)

Counterfactuals

A causal model presupposes a counterfactual

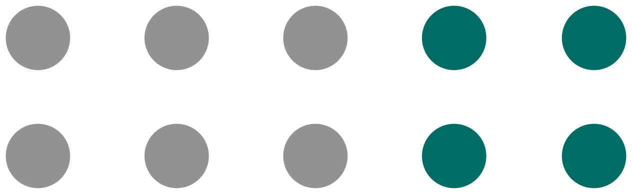
A, **Intervention**: Student attends an independent school → Y, **Outcome**: Student gets a place at Cambridge

A', **Intervention**: Student does not attend an independent school → Y, **Outcome**: Student does not get a place at Cambridge?



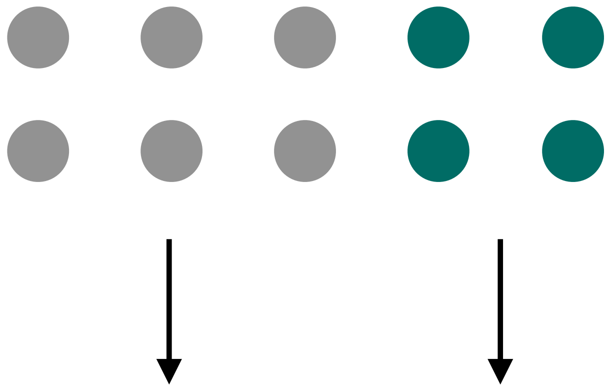
Association and causation

Population



a a$'$

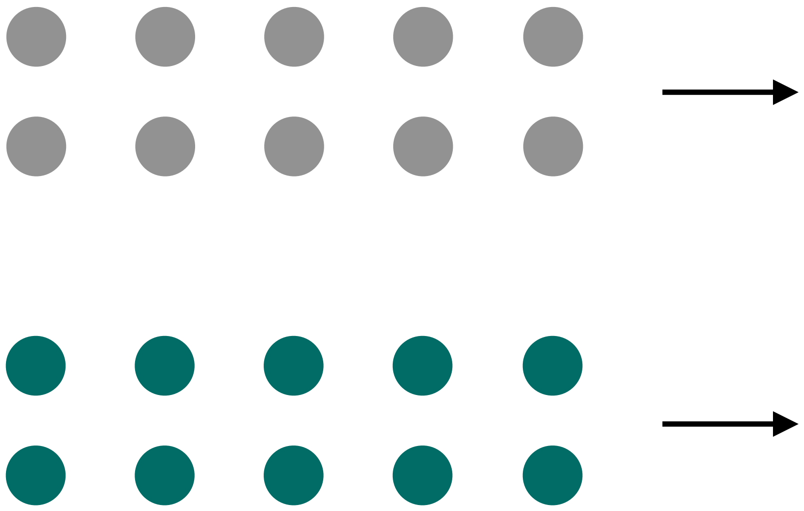
Association



a a$'$

$E[Y^{A=a}]$ $E[Y^{A=a'}]$

Causation

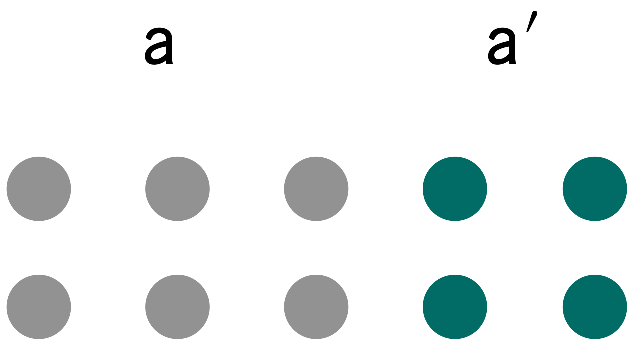


a $\longrightarrow$ $E[Y \mid A = a]$

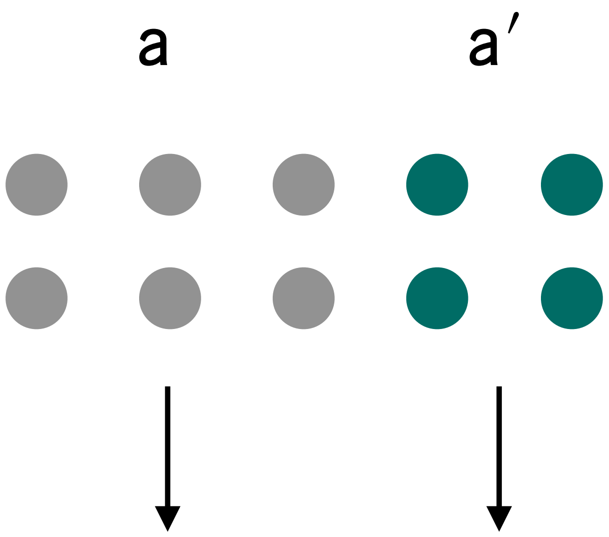
a$'$ $\longrightarrow$ $E[Y \mid A = a']$

Association and causation

Population



Association



$E[Y^{A=a}]$ $E[Y^{A=a'}]$

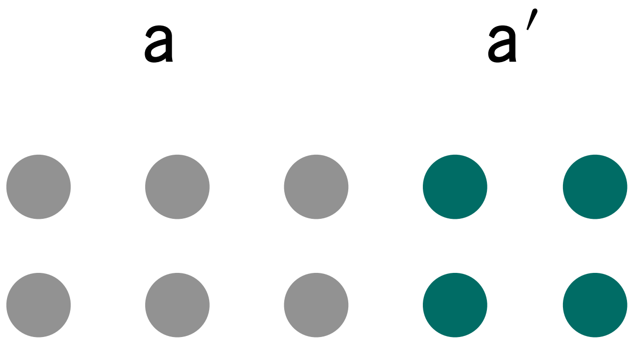
What is the probability of going to Cambridge for students at public schools

What is the probability of going to Cambridge for students at private schools

⇒ The world as it is

Association and causation

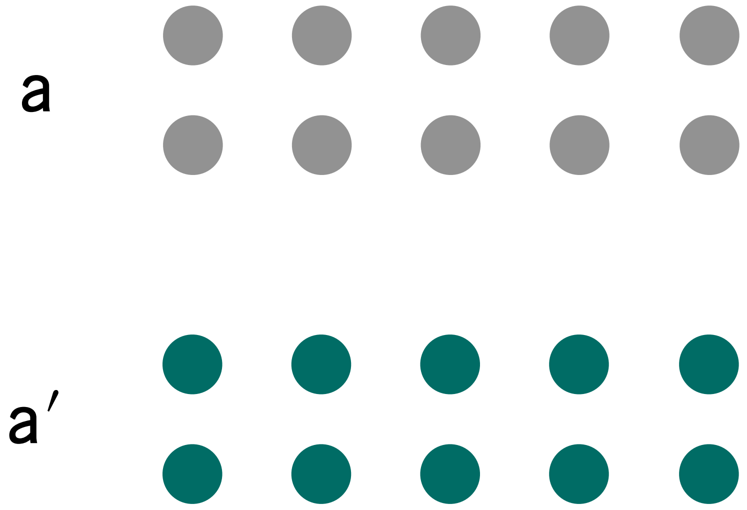
Population



What if a student at the private school had attended a public school?

What if a student at the public school had attended a private school?

Causation



$\longrightarrow E[Y \mid A = a]$

$\Rightarrow$ A counterfactual world

$\longrightarrow E[Y \mid A = a']$

Fundamental problem of causal inference

We cannot observe the counterfactual

Building blocks of causal modeling

Notation

A: intervention

X: observable attributes

U: unobserved attributes

Y: outcome

Capital letters refer to features and lower case letters refer to a value that feature takes

e.g. suppose A is school type, then **a** = public and **a**′ = private

Confounders

X, **Confounding variable**:

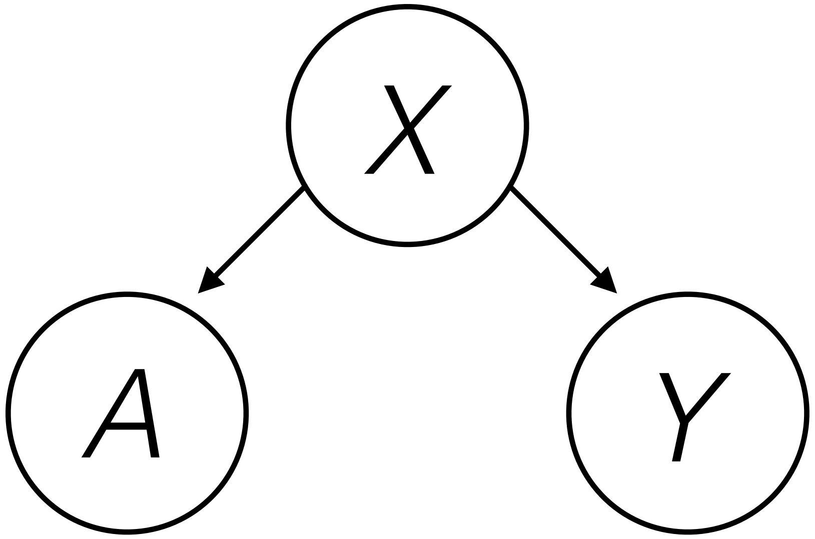
?

A, Intervention:

Student attends a private school

Y, **Outcome**:

Student admitted to Cambridge



Confounders

X, **Confounding variable**:

Parent's income

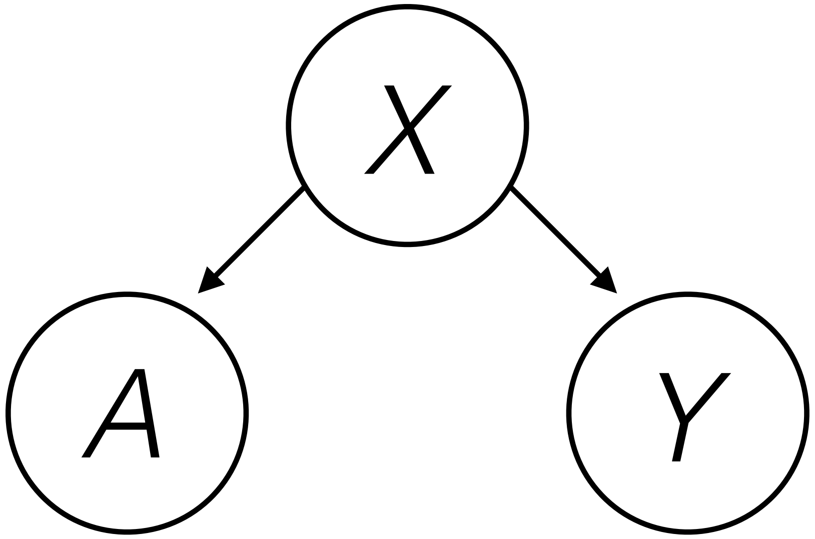
A, **Intervention**:

Student attends a private school

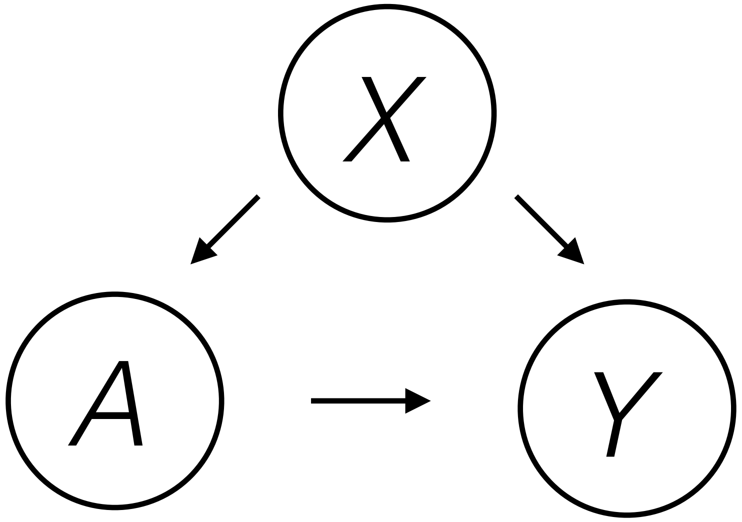
Y, **Outcome**:

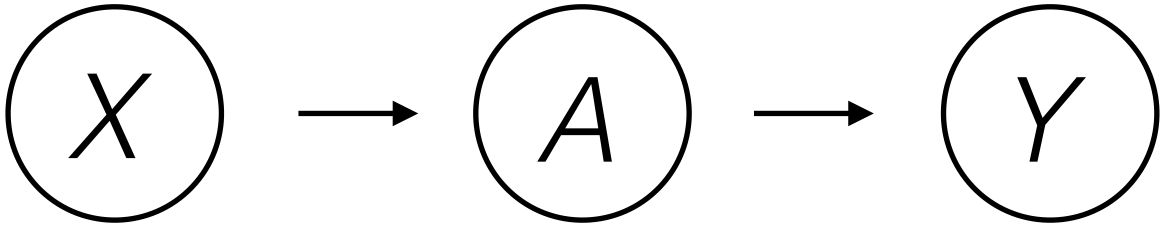
Student admitted to Cambridge

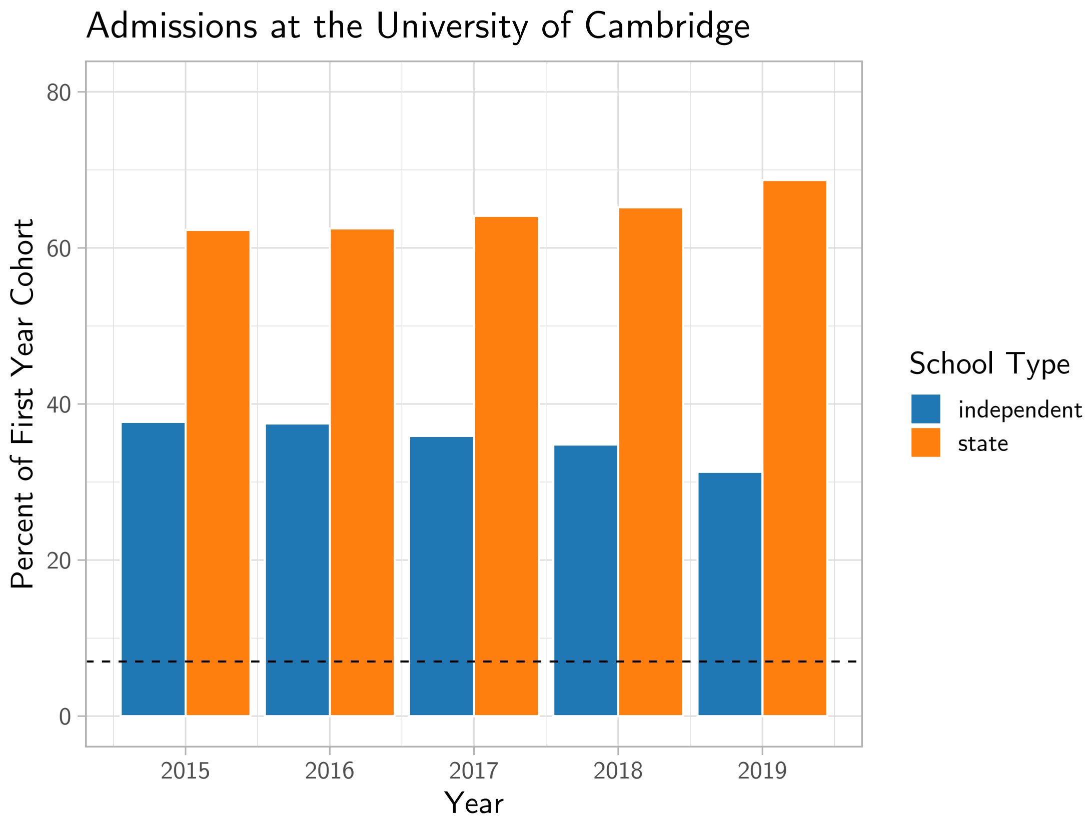
We could postulate many confounders

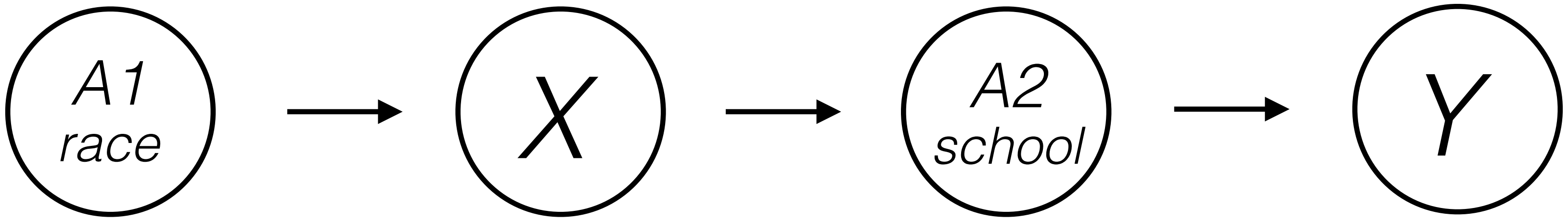


We can postulate many causal models

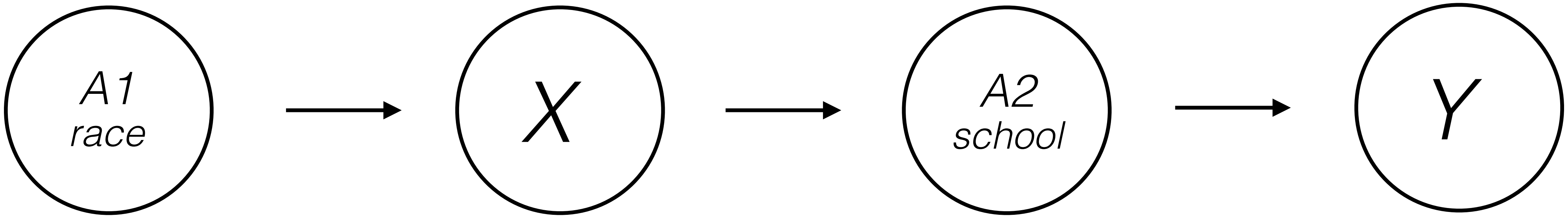








Ancestors in DAGs



causal models and fairness

How does this relate to fairness?

- Many ideas in (algorithmic) fairness rely on causal reasoning

- Consider admissions to Cambridge. Why might we consider it unfair?

- Student’s chance of admission is lower if they attend a public school, and these circumstances are *morally arbitrary* (outside of a student’s control)

- We often invoke a counterfactual when discussing fairness, e.g. bank loans; what if the person had been old instead of young…

Counterfactual fairness

A predictor $\hat{\mathbf{Y}}$ is counterfactually fair if under any context X = *x* and A = *a*,

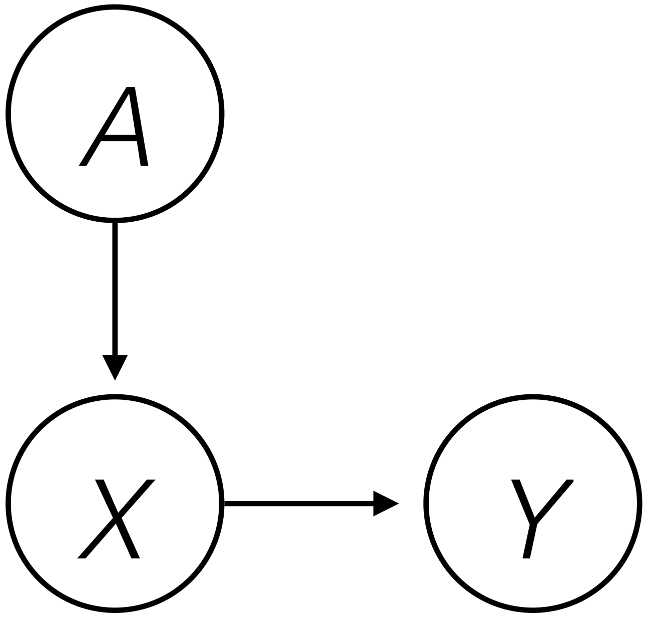
$$P(\hat{\mathbf{Y}}_{a} \,|\, \mathbf{X} = x, \mathbf{A} = a) = P(\hat{\mathbf{Y}}_{a'} \,|\, \mathbf{X} = x, \mathbf{A} = a)$$

for all a'

Capital letters represent random variables. Lower case letters denote particular values of a random variable.

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

Is COMPAS counterfactually fair?

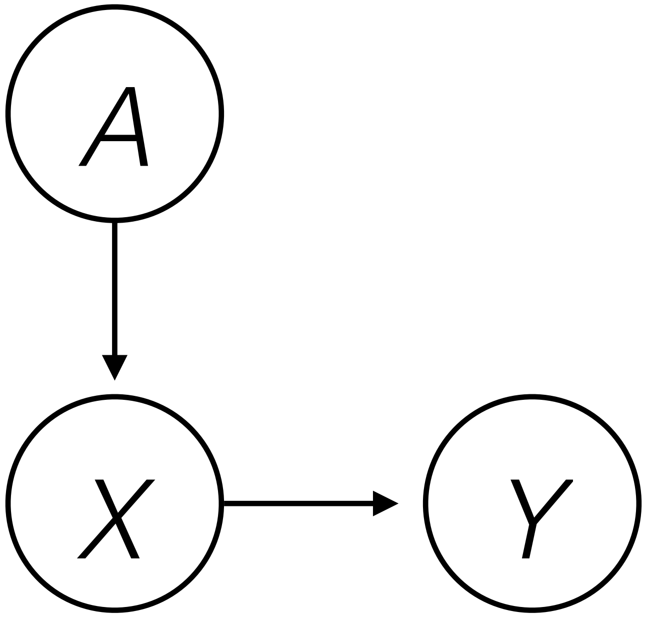


- **A**: protected attribute, race
- **X**: predictors, e.g. previous charges, contact with criminal justice system
- **Y**: recidivism

Y = f(X), X = f(A)

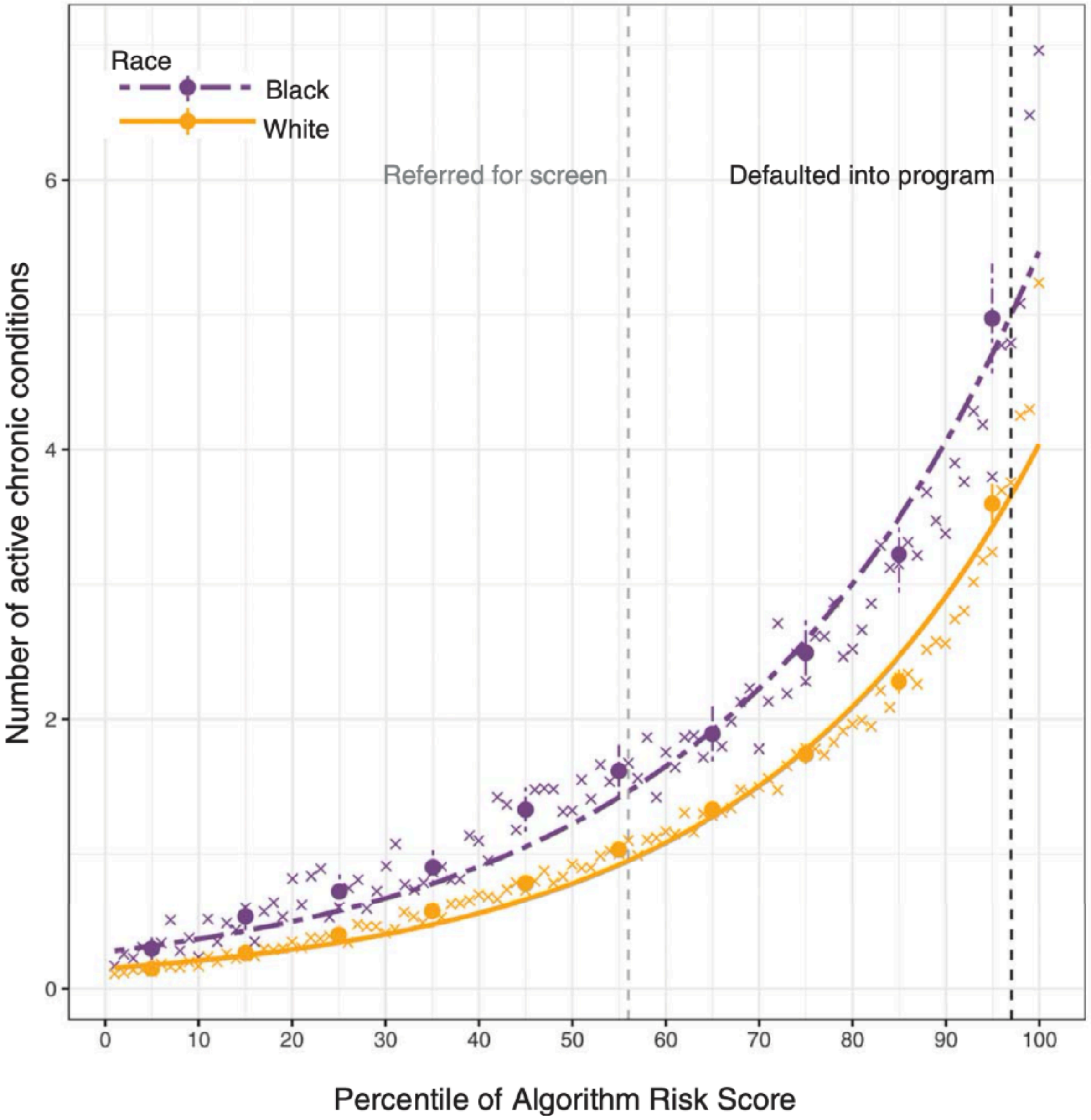
X is descendant (downstream) of A; Y = f(X, A)

Counterfactual fairness in healthcare



- A: protected attribute, race
- X: medical expenditures
- Y: future healthcare needs

$$P(\hat{\mathbf{Y}}_{a}\,|\,\mathbf{X} = \$50{,}000) \neq P(\hat{\mathbf{Y}}_{a'}\,|\,\mathbf{X} = \$50{,}000)$$



[Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Science *2019*]

Counterfactual fairness

- The prediction/outcome should not be a causal descendant of an individual's protected attribute*
- This is contingent on the postulated causal model representing the world as it is; what if the model is a poor representation?
- Promotes transparency: causal model must be postulated
- Idea: many (competing) worlds can be postulated

* we'll revisit this in a moment

[C. Russell, M.J. Kusner, J.R., Loftus, R. Silva, NeurIPS (2017)

Bloomberg’s World

“So you want to spend the money on a lot of cops in the streets. Put those cops where the crime is, which means in minority neighborhoods.

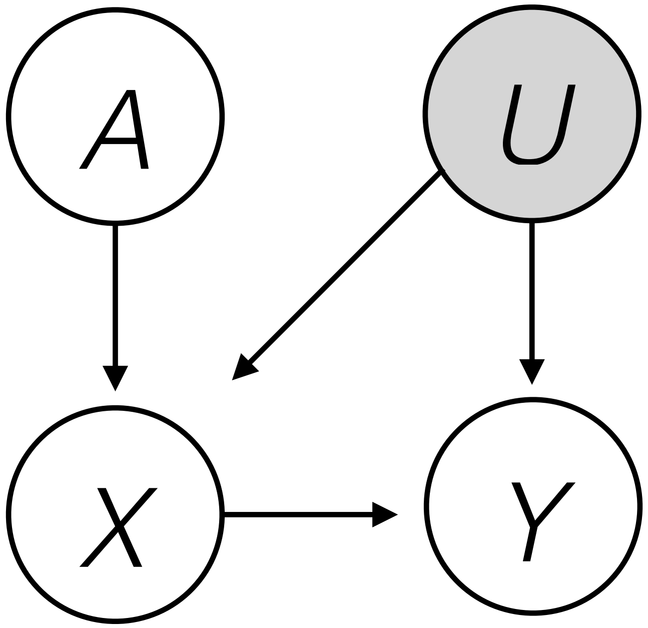
So one of the unintended consequences is people say, “Oh my God, you are arresting kids for marijuana that are all minorities.” Yes, that’s true. Why? Because we put all the cops in minority neighborhoods. Yes, that’s true. Why do we do it? Because that’s where all the crime is.”

Michael Bloomberg (2015)

Bloomberg’s World

To make a thief, make an owner; to create crime, create laws.

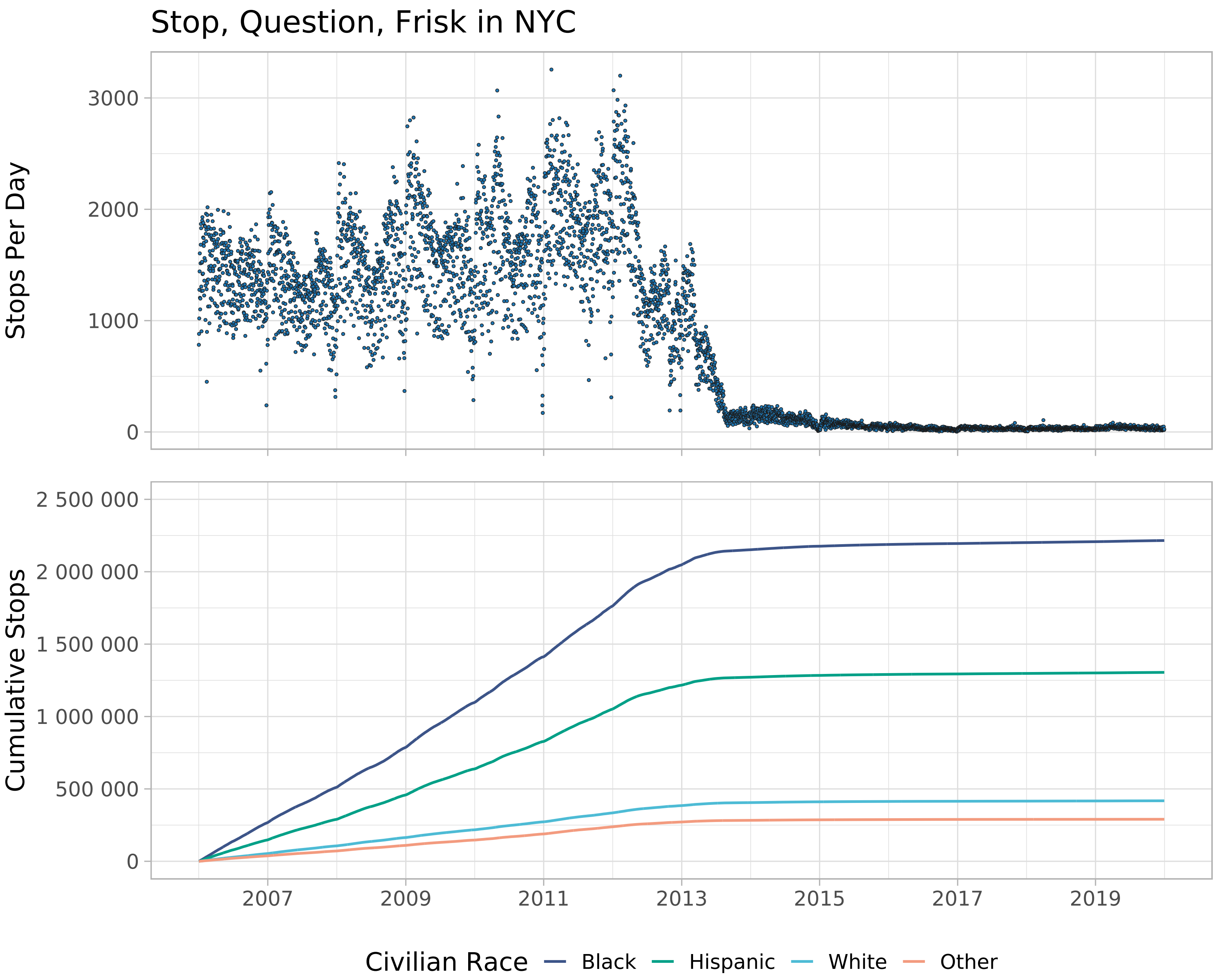
Ursula K. Le Guin, The Dispossessed



- **A**: racial composition of neighborhood
- **X**: police deployment rate
- **U**: other factors influencing enforcement patterns and charge rate
- **Y**: criminal charge

Bloomberg argued the city should determine X based on Y, encoding the targeted policing of minorities.

The consequences



Causal ancestors, the case of Berkeley Admissions

- An early paper on fairness studied graduate admissions at Berkeley
- Women applicants were admitted at lower rates
- However, women applied to more competitive departments, on average
- At the department-level, women were slightly favored in admissions

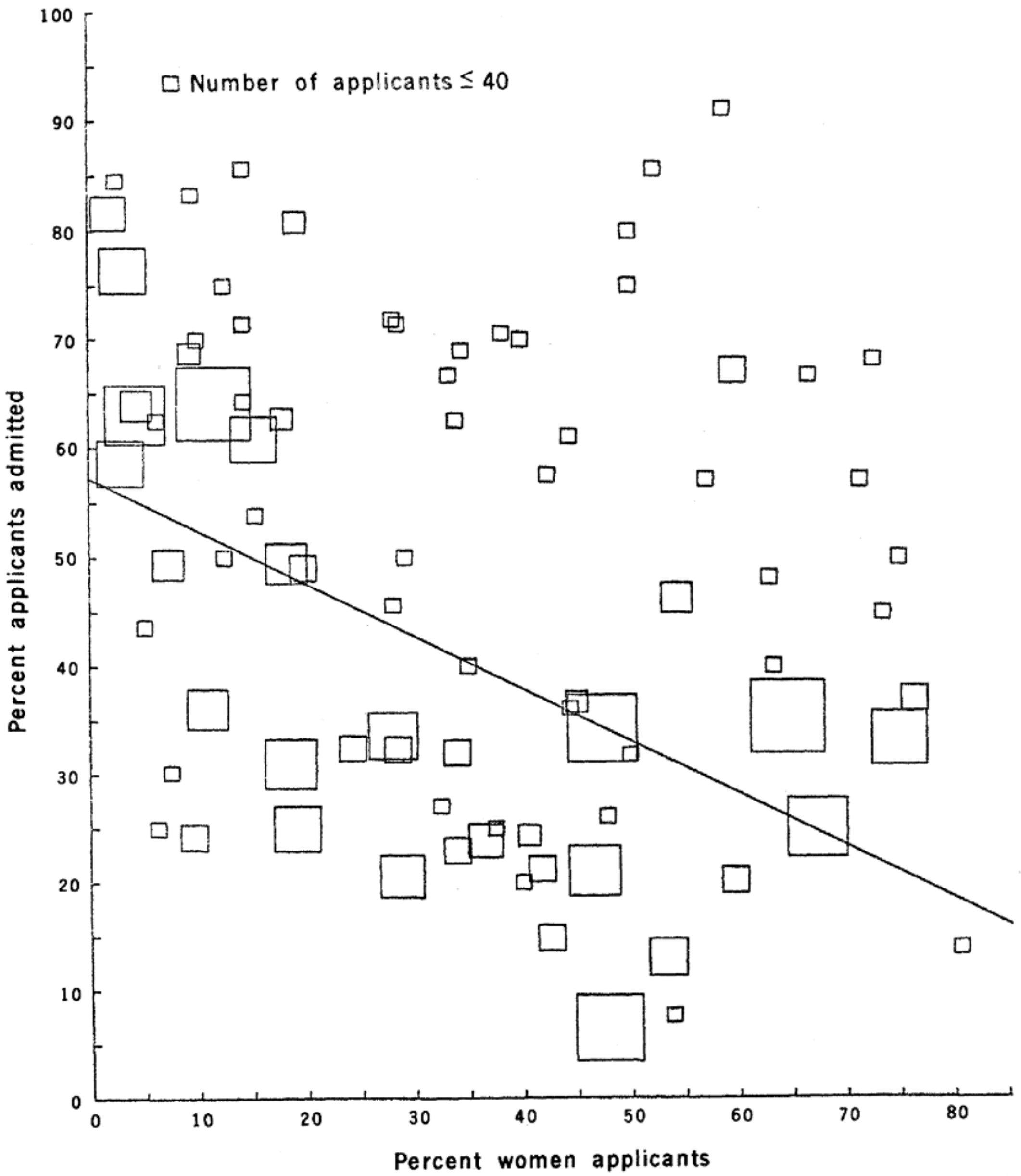
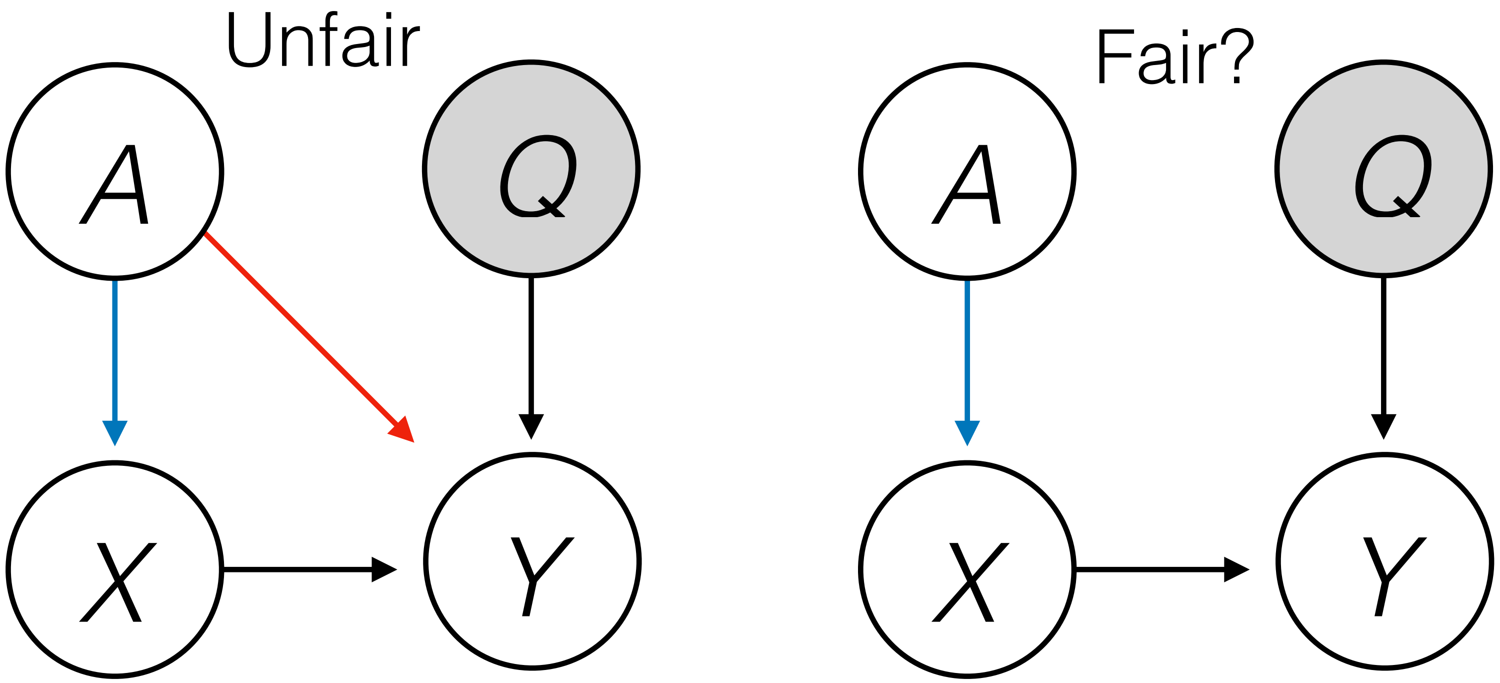


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

[E.A. Bickel, J. Hammel, W. O'Connell, Science *1975]*

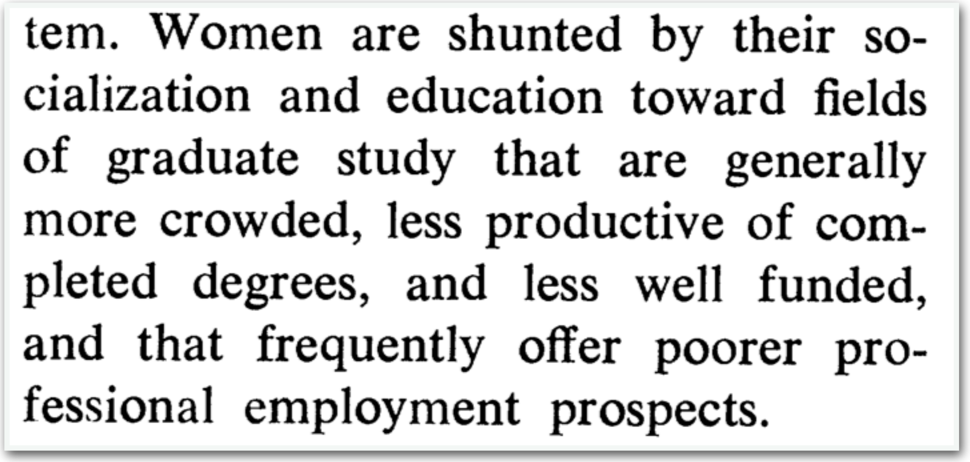
Path-specific counterfactual fairness

- A: gender
- X: department choice
- Q: qualification
- Y: admission


Unfair

Fair?

- Fair at what decision point? For which decision maker?
- Berkeley (the vendor) might say "you can't expect us to resolve sexism in broader society!"


tem. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

[S. Chiappa and T.P.S Gillam, arXiv:1802.08139 (2018)]

Counterfactual privilege

Suppose a vendor wants to implement a policy z. Proposes constraining "counterfactual privilege" such that:

$$\mathrm{E}[\hat{\mathrm{Y}}_{\mathrm{i}}(\mathrm{a}_{\mathrm{i}}, \mathbf{z})] - \mathrm{E}[\hat{\mathrm{Y}}_{\mathrm{i}}(\mathrm{a}_{\mathrm{i}}', \mathbf{z})] \leq \tau$$

- Exclude intervention assignments that allow an individual i to become more than τ units better off in expectation due to the interaction of z and A
- Anything $\geq \tau$ is considered unfair privilege

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

Counterfactual privilege

- Suppose US Department of Education wants to increase college attendance
- Proposes an intervention that will provide financial assistance for 25 schools in NYC to hire a Calculus tutor
- Which schools should receive financial assistance?

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

Counterfactual privilege

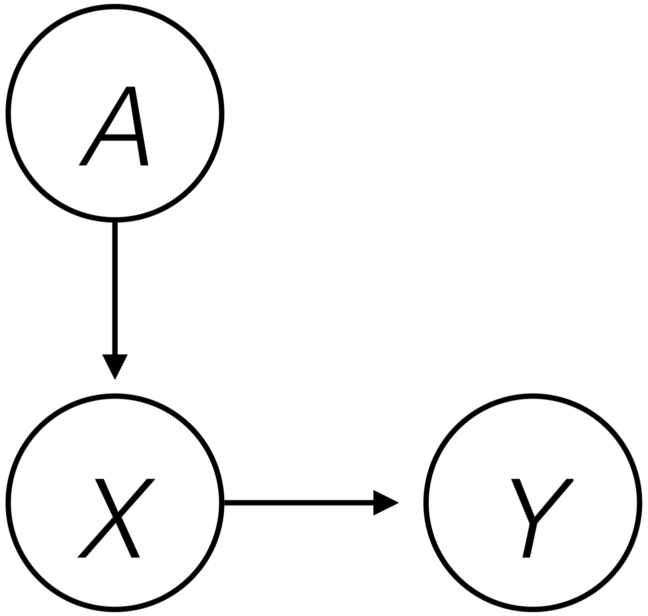
- We can estimate the expected number of additional college applicants for all feasible allocations of z

$$\sum_{i=1}^{n} E[Y_i(\mathbf{z}) \,|\, A_i = a_i, X_i = x_i)$$

- Under each allocation, we can assess how much “better off” group a would be relative to group a′
  - This quantity is τ
- We have a solution path of possible values of τ and trade-offs with respect to the expected number of additional applicants

Revisiting Chief Justice John Roberts

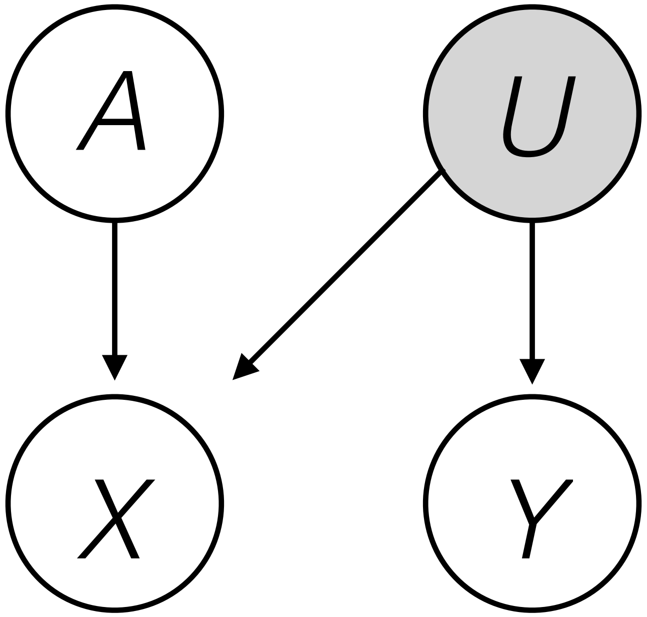
> "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."
>
> Chief Justice John Roberts (2017)



- Fairness through unawareness
- But A cannot be disentangled from X
- This is a common pattern of counterfactual unfairness

Chief Justice Roberts' view can introduce unfairness

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]

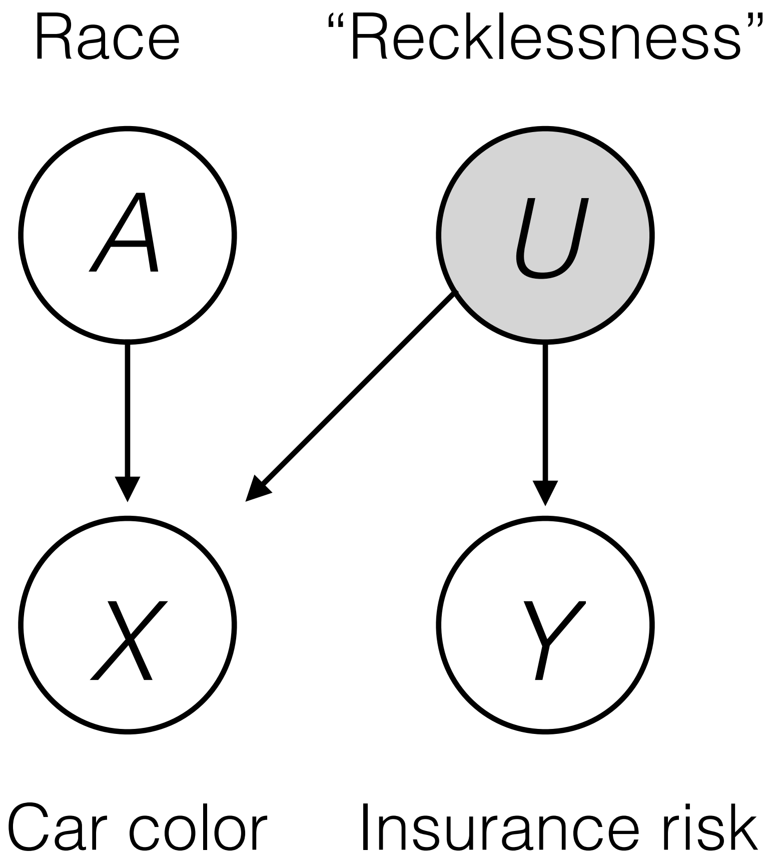


Note that X doesn't cause Y in this model!

- The variable **X** is a descendant of **U**
- **X** is also a descendant of **A**, i.e. **X** = f(**A**, **U**)
- If we use **X** to predict Y, we are using **U** and **A**

Chief Justice Roberts' view can introduce unfairness

[M.J. Kusner, J. Loftus, C. Russell, R. Silva, arXiv:1703.06856v3 2018]



- "Fairness through unawareness" introduces unfairness
- X = f(A, U); by ignoring A we cannot adjust for its influence on X
- This would be counterfactually unfair

$$P(\hat{Y}_{A \leftarrow a} \mid X = \text{red}) \neq P(\hat{Y}_{A \leftarrow a'} \mid X = \text{red})$$

A causal framework for fairness

- Causal reasoning and counterfactual fairness clarifies what is at stake in a particular data science task (e.g. risk assessment)

- Enhances transparency by requiring the specification of a causal model

However,

- It is a framework for assessing and enhancing fairness given causal model(s) + data, not a “solution to fairness”

- Underlying moral and ethical concerns around risk-assessment tools and other ML tasks do not go away, nor do problems of data bias