Responsible Data Science

Fairness and Causality

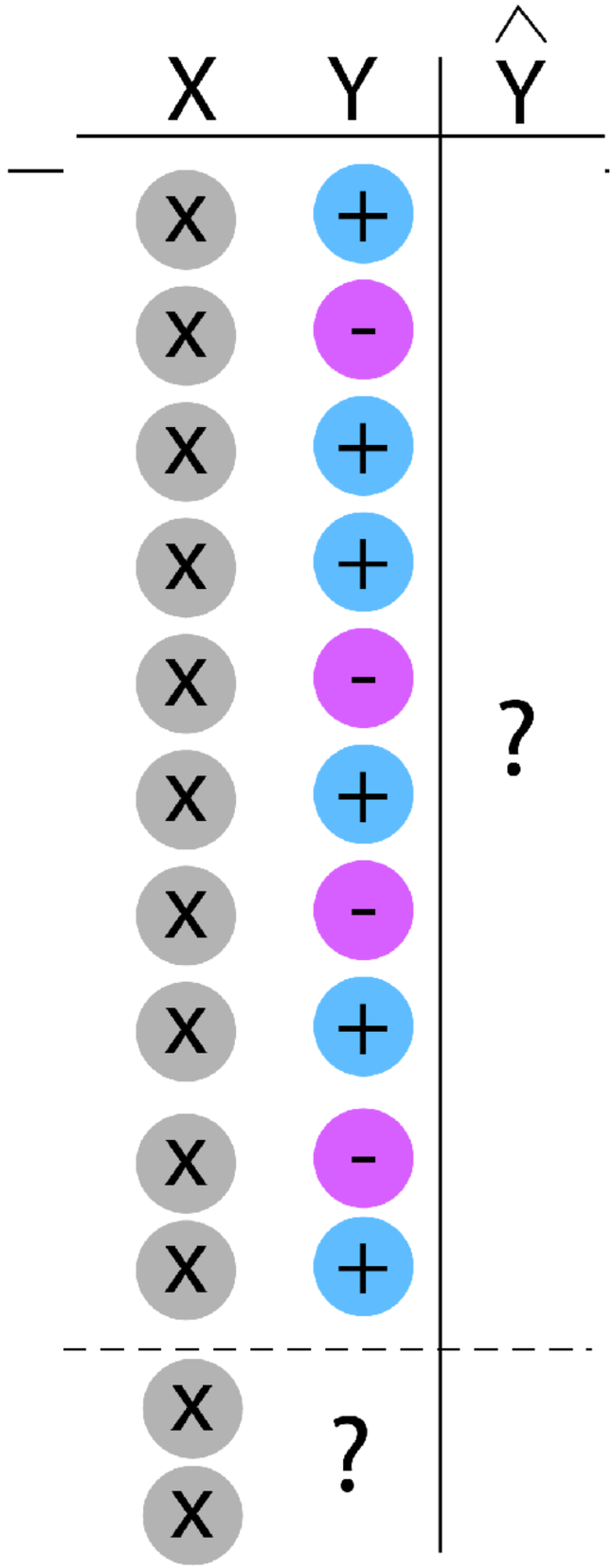
Prof. Elisha Cohen

Center for Data Science
New York University

data RESPONSIBLY

NYU

Review of Prediction



- We observe X and Y
- Want to predict $\hat{Y}$

- Why?
- We want to be able to predict $\hat{Y}$ out of sample when we do know observe Y

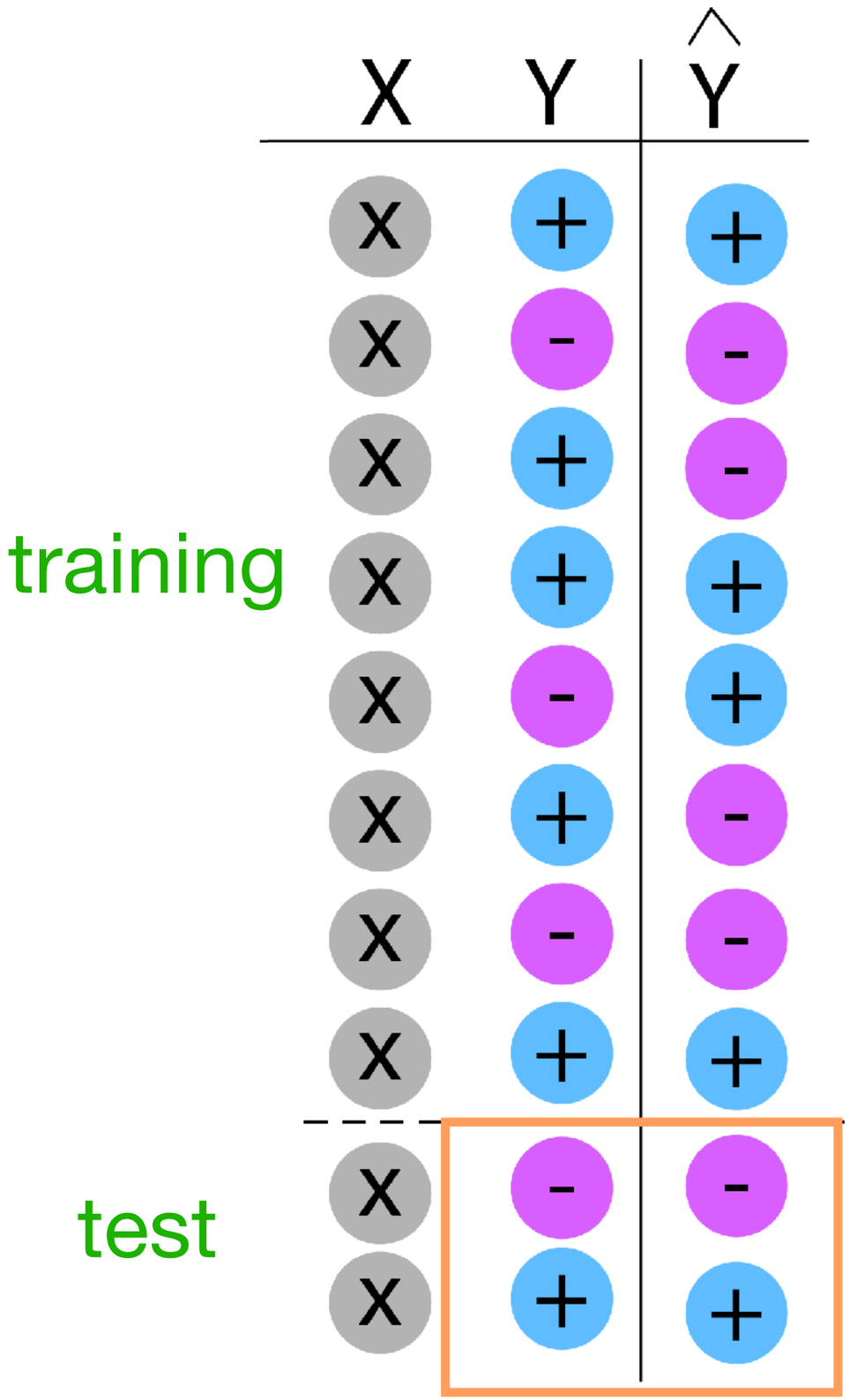
Review of Prediction

X	Y	$\hat{Y}$
x	+	+
x	-	-
x	+	+
x	+	+
x	-	-
x	+	+
x	-	-
x	+	+
x	-	-
x	+	+

- We observe X and Y
- Want to predict $\hat{Y}$
- What happens if we use all of our data to estimate a mapping ?
  $X \rightarrow Y$
- Overfit!

Bad out-of-sample fit!!!

Review of Prediction

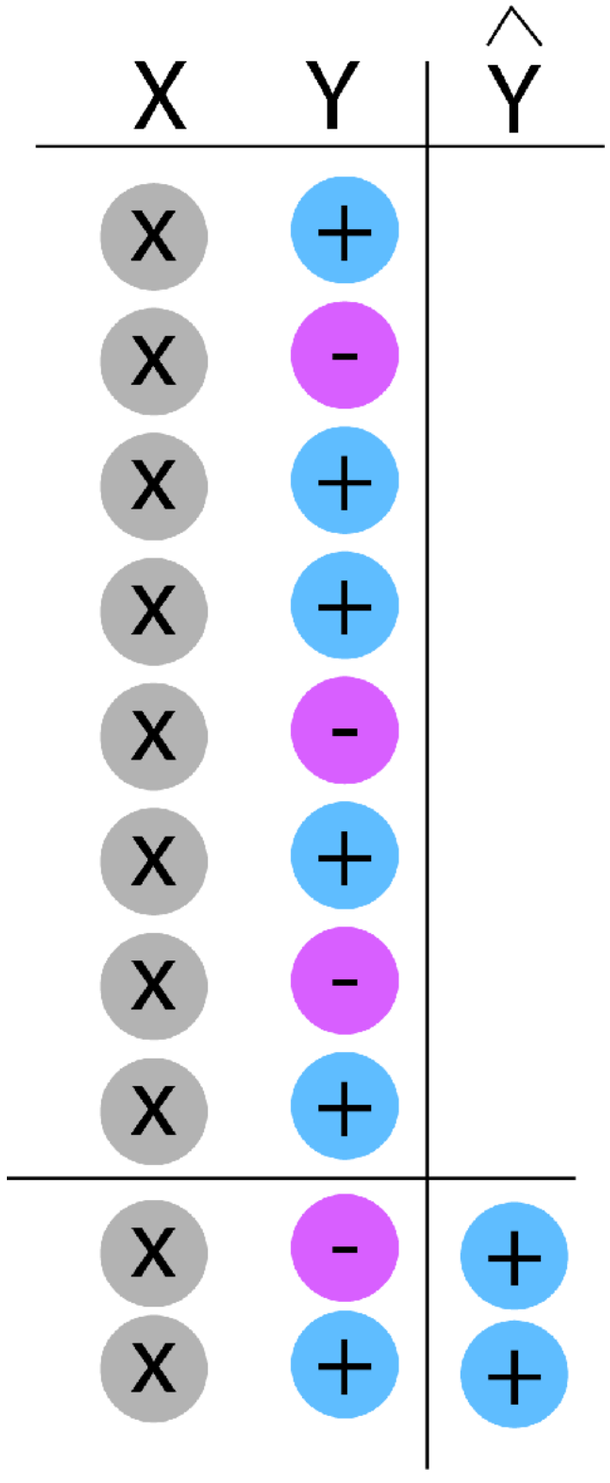


Procedure: split data into **training data** and **test data**

1. Fit model only on training data
2. Use model to make predictions $(\hat{Y})$ on test data
3. Compare true labels (Y) with predictions $(\hat{Y})$ on test data to evaluate accuracy

Evaluate accuracy

Review of Prediction



- Baseline classifier (very naive approach)
- Predict $\hat{Y}$ for test data to be majority label (no learning from training data)

causal models and fairness

r/ai

Review of notation

Notation

A: protected attributes

X: observable attributes

U: unobserved attributes

Y: outcome

$\hat{\mathbf{Y}}$: 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

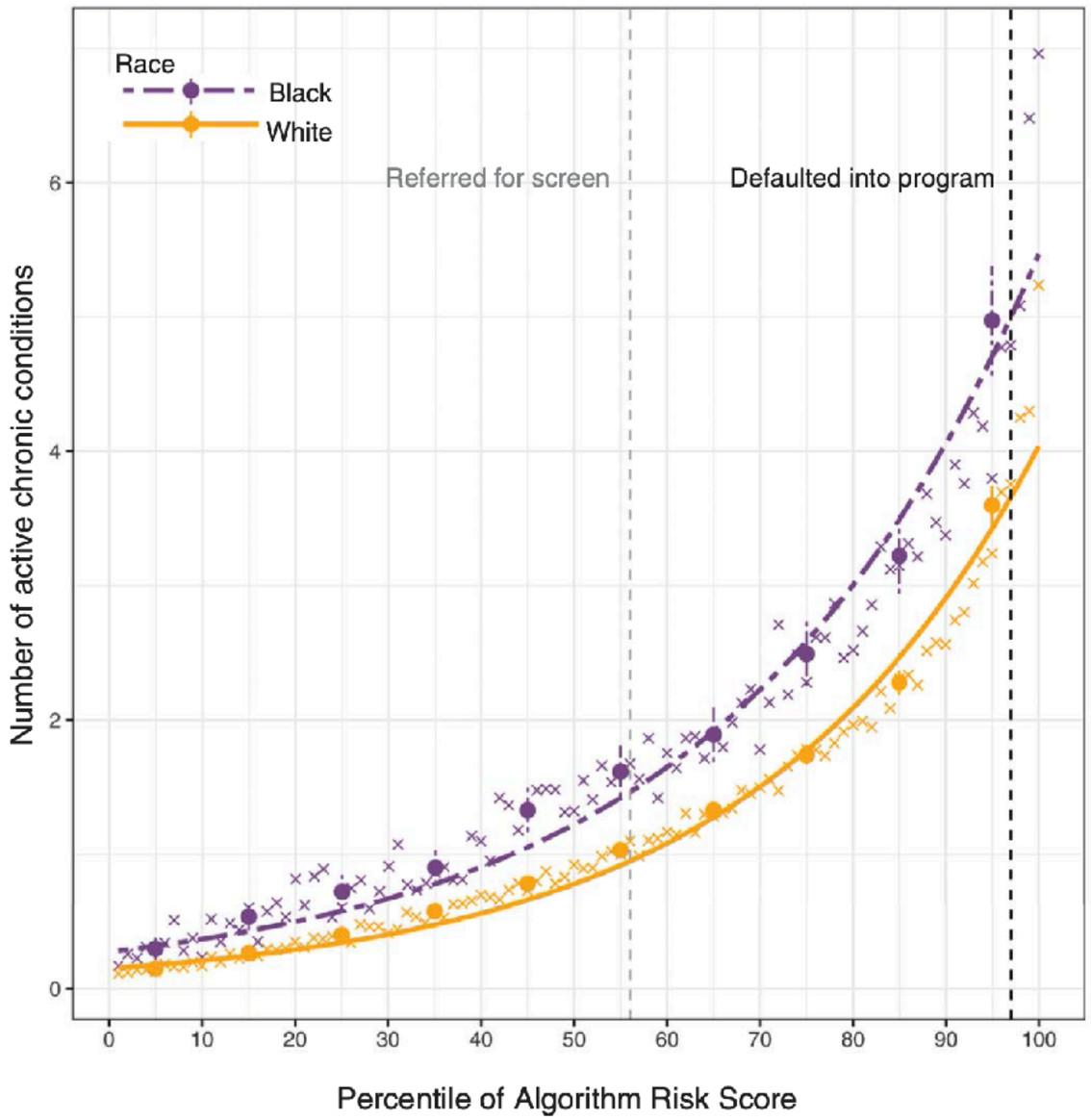
e.g. suppose **A** is age, then **a** = old and **a′** = young

How does this relate to fairness?

- Many ideas in (algorithmic) fairness rely on causal reasoning
- Consider health disparities example. Why might we consider it unfair?

How does this relate to fairness?

- Many ideas in (algorithmic) fairness rely on causal reasoning
- Consider health disparities example. Why might we consider it unfair?



- A: Race (protected characteristic)
- X: Medical expenditures
- Y: Future healthcare needs

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

How does this relate to fairness?

- Many ideas in (algorithmic) fairness rely on causal reasoning
- Consider health disparities example. Why might we consider it unfair?
  - Patient's referral for screening was based on past medical expenditures. Patient without health insurance may not go to the doctor even though have poor health
  - 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*,

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

for all a'

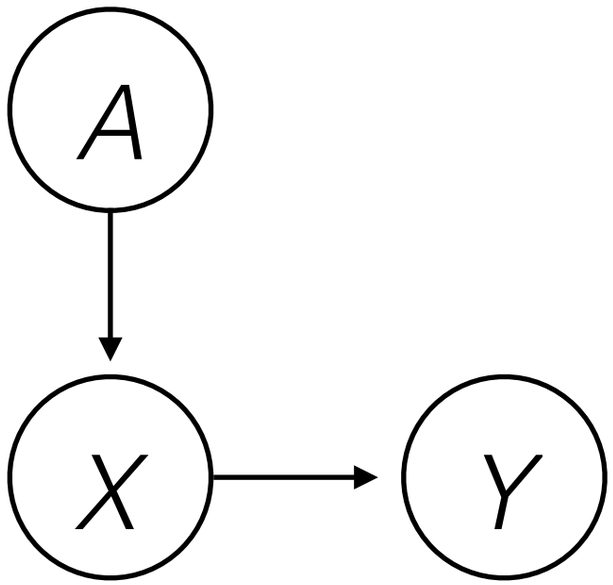
Capital letters represent random variables.

Lower case letters denote particular values of a random variable.

We denote an “intervention” (i.e. a change in value *a*) on A by the notation: $A \leftarrow a$.

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

Is COMPAS counterfactually fair?

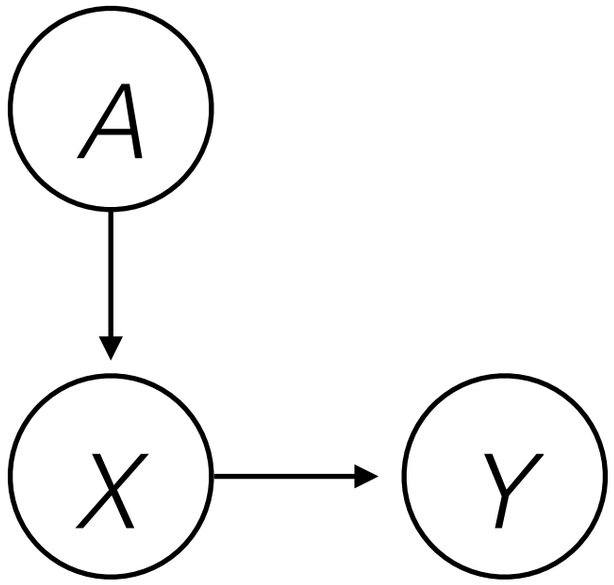


- **A**: protected attribute, race
- **X**: predictors, e.g. previous charges, contact with criminal justice system
- **Y**: recidivism

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

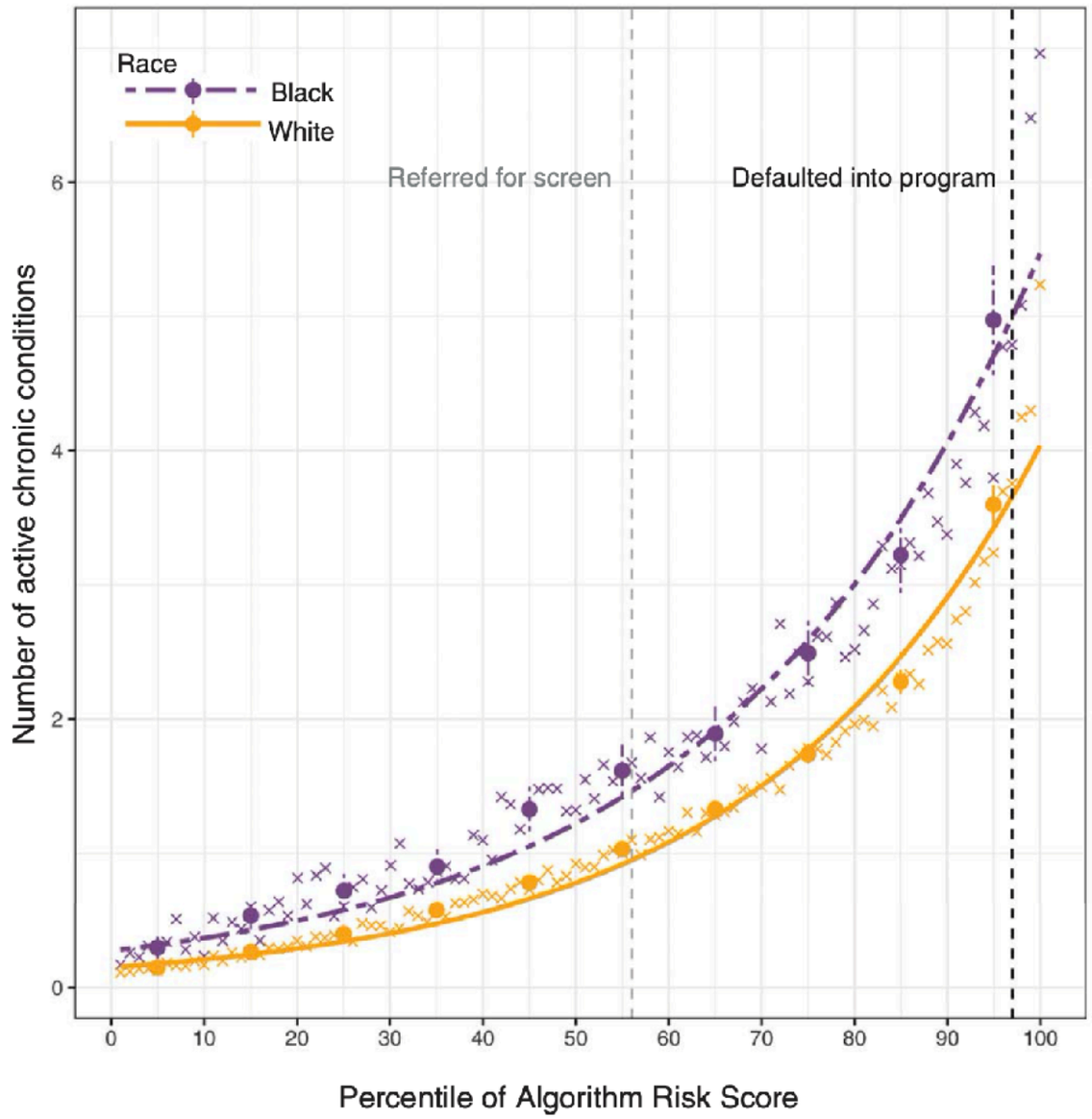
X is descendant (downstream) of A; $Y = f(X, A)$

Counterfactual fairness in healthcare



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

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



[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

[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.

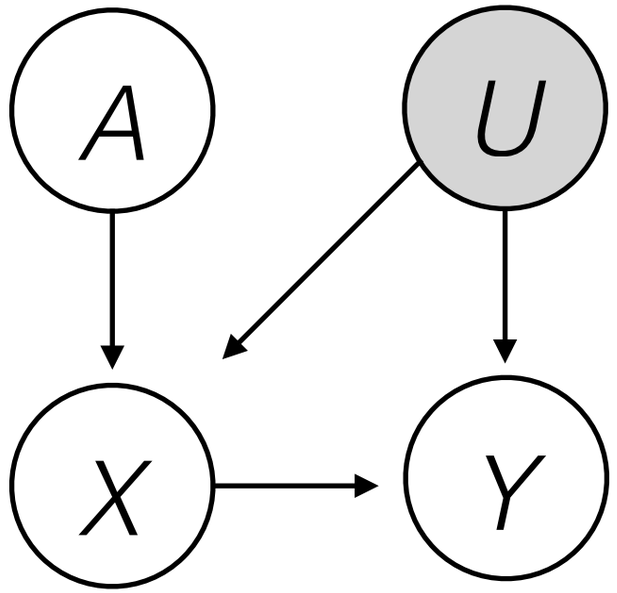
So one of the unintended consequences is people say, "Oh my God, you are arresting kids for marijuana that are all minorities." Yes, that's true. Why? Because we put all the cops in minority neighborhoods. Yes, that's true. Why do we do it? Because that's where all the crime is."

Michael Bloomberg (2015)

Bloomberg's World

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

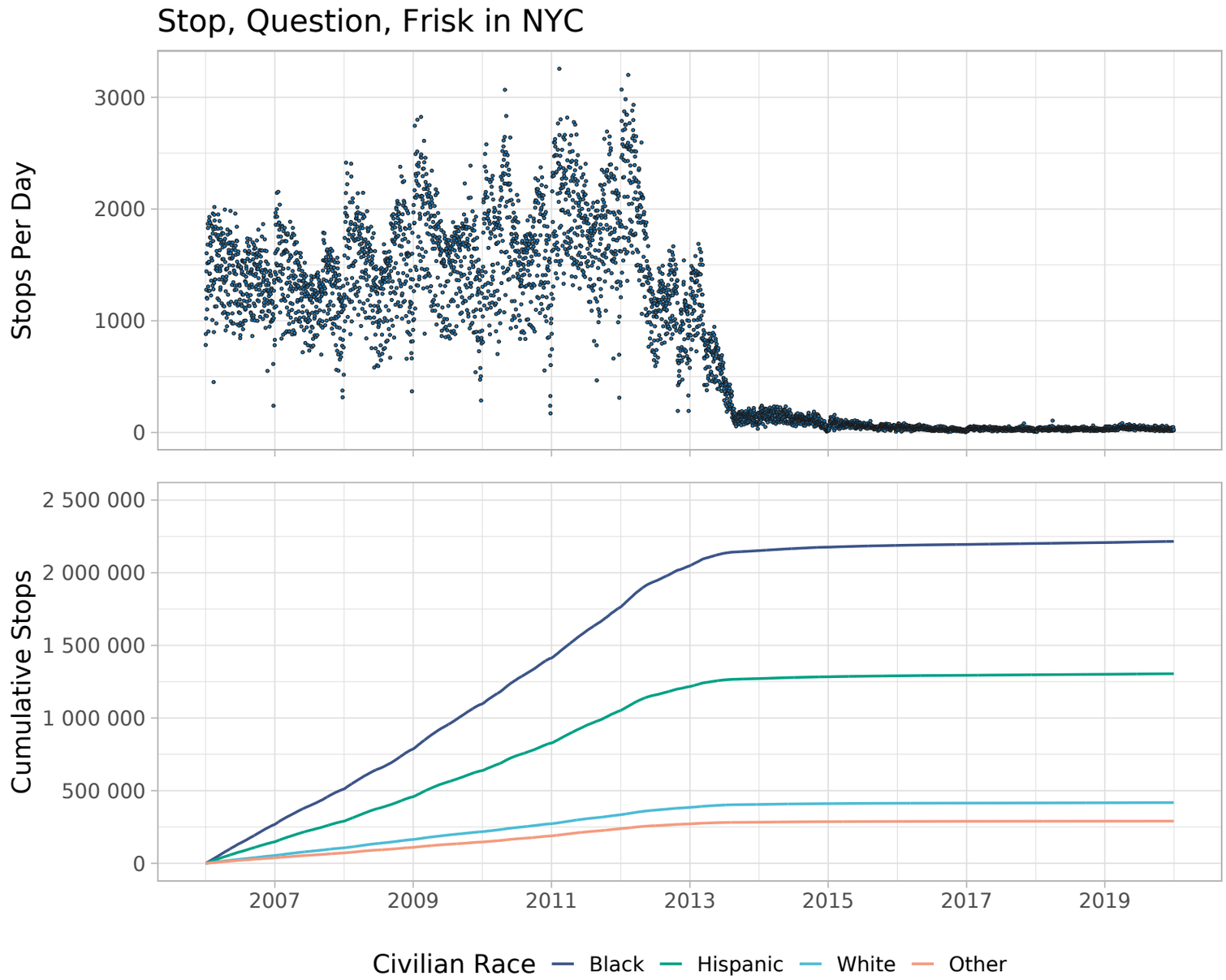
Ursula K. Le Guin, The Dispossessed



- **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

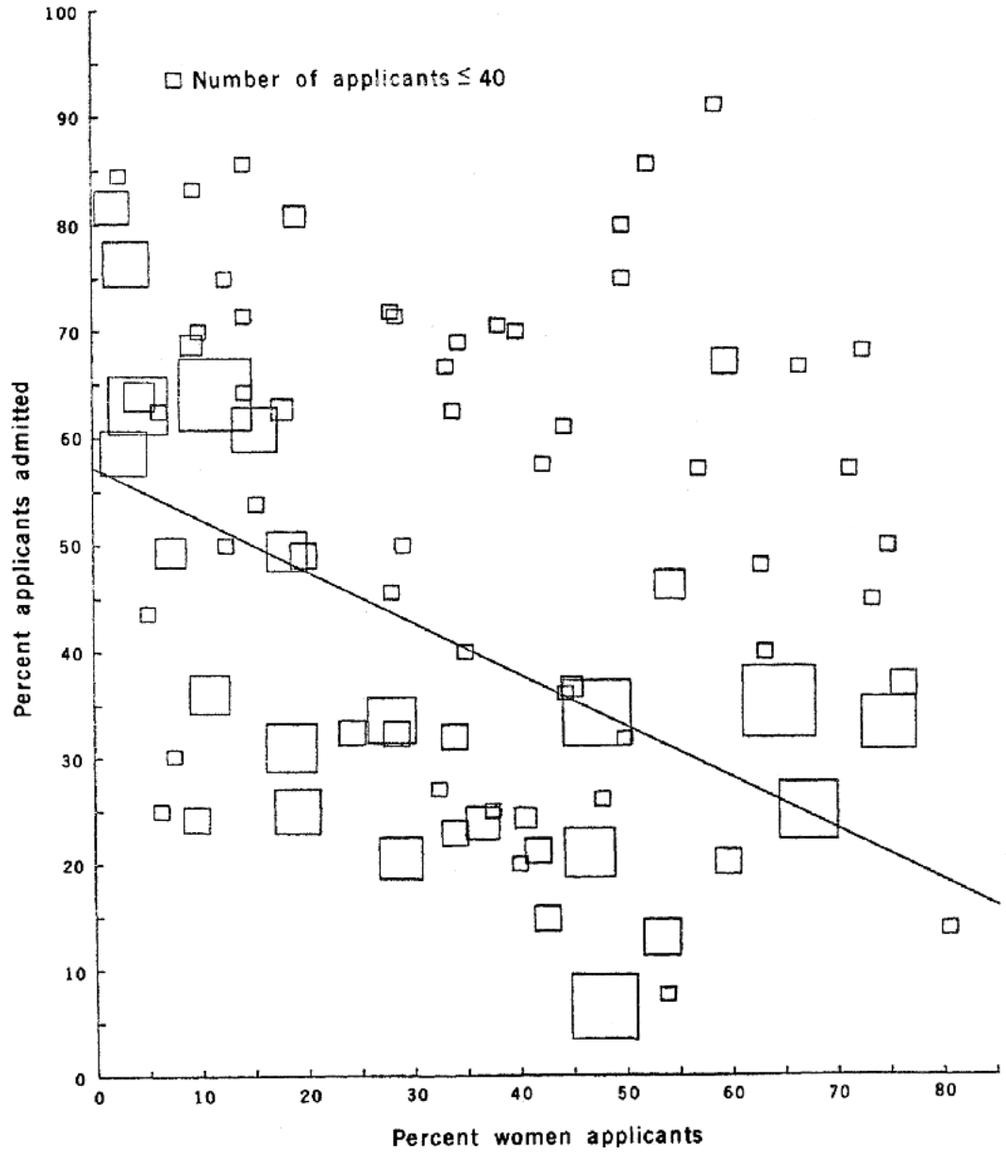


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

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

Individual vs group fairness

Recall:

- Individual fairness is satisfied if two individuals—who are similar with respect to a task—have the same probability of the positive outcome
- Demographic parity (group fairness) is satisfied when the probability of the positive outcome is the same for all groups

Can Berkeley satisfy both? How?

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

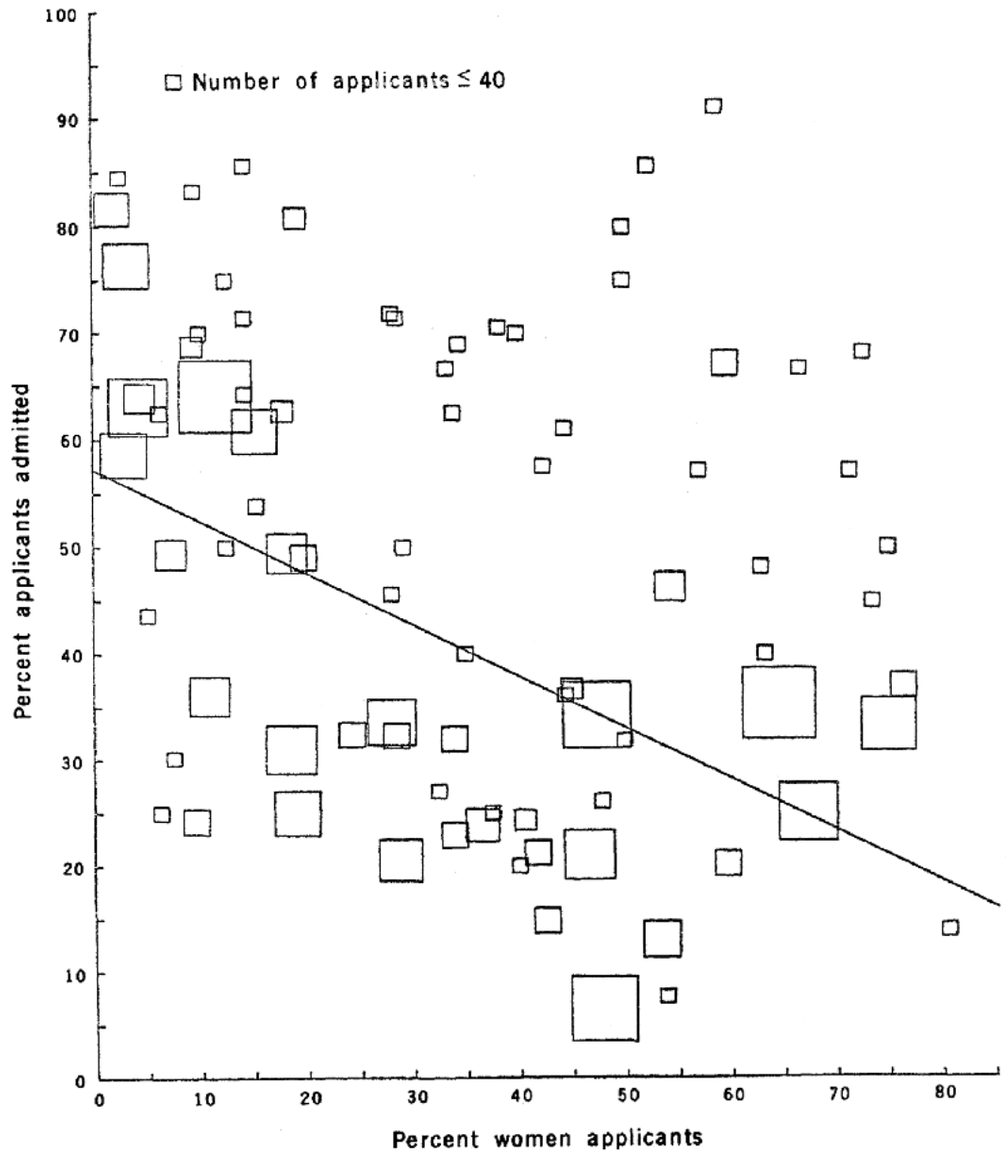
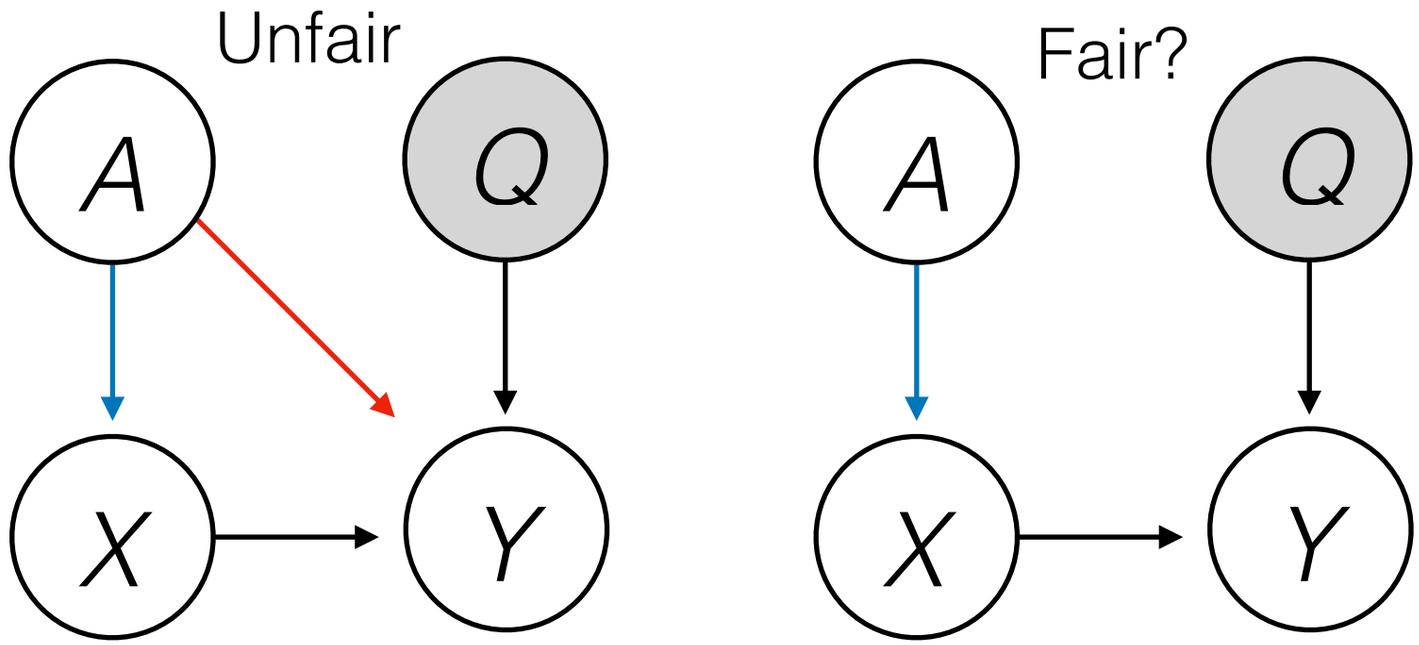


Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

Path-specific counterfactual fairness

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



- 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

1. Constrain ML methods so predictions satisfy a chosen notion of fairness

2. Interventions from a new policy

   A. Maximize positive effects

   B. Improve fairness overall

- A: protected characteristic
- X: features
- Z: Intervention
- Y(z): counterfactual outcome with z intervention
- Y(0): outcome with no intervention

[M.J. Kusner, J. Loftus, C. Russell, R. Silva. “Causal Interventions for Fairness” arXiv:1806.02380 2018]

Counterfactual privilege

1. Want $Y_i(a, \mathbf{z}) = Y_i(a', \mathbf{z})$
   - hard to guarantee without perfect intervention
2. Is $Y_i(a, \mathbf{z})$ preferable to $Y_i(a,0)$?
   - Need to consider which outcomes are desirable

Define privilege as having a better outcome *because* of ones value of A i.e.

$$\mathbb{E}[Y_i(a,0)] > \mathbb{E}[Y_i(a',0)]$$

[M.J. Kusner, J. Loftus, C. Russell, R. Silva. “Causal Interventions for Fairness” arXiv:1806.02380 2018]

Counterfactual privilege

Suppose a vendor wants to implement a policy *z*. We can constrain "counterfactual privilege" such that:

$$E[\hat{Y}_i(a_i, \mathbf{z})] - E[\hat{Y}_i(a_i', \mathbf{z})] \leq \tau$$

- Exclude policies 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]

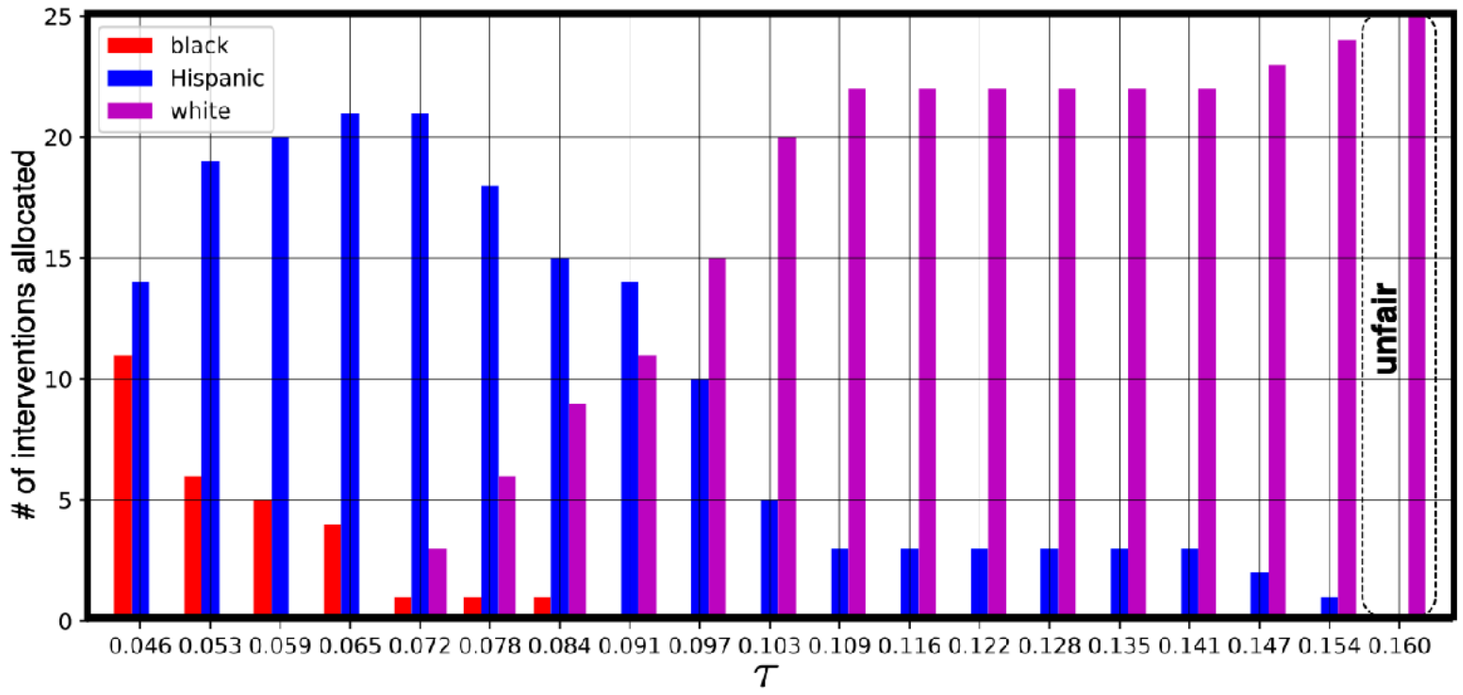
Counterfactual privilege

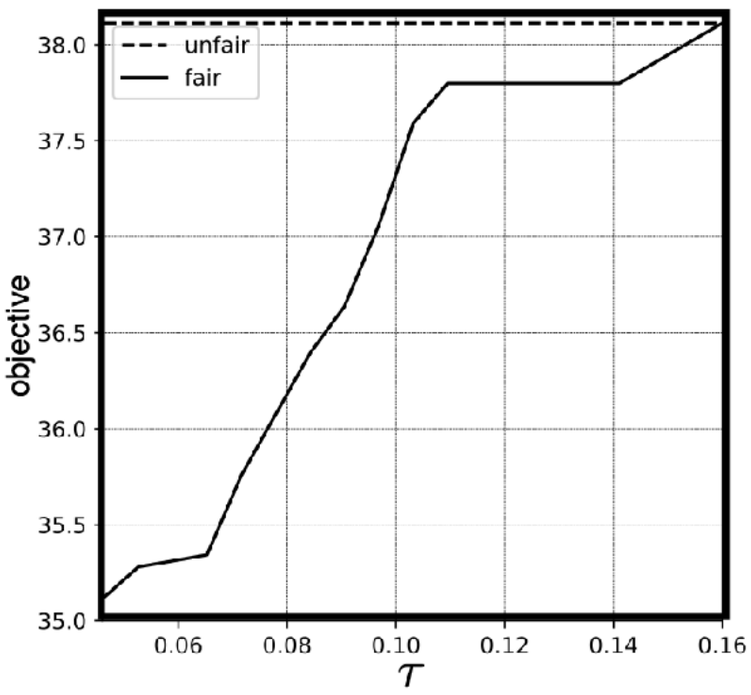
- We can estimate the expected number of additional college applicants for all feasible allocations of z

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

- 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

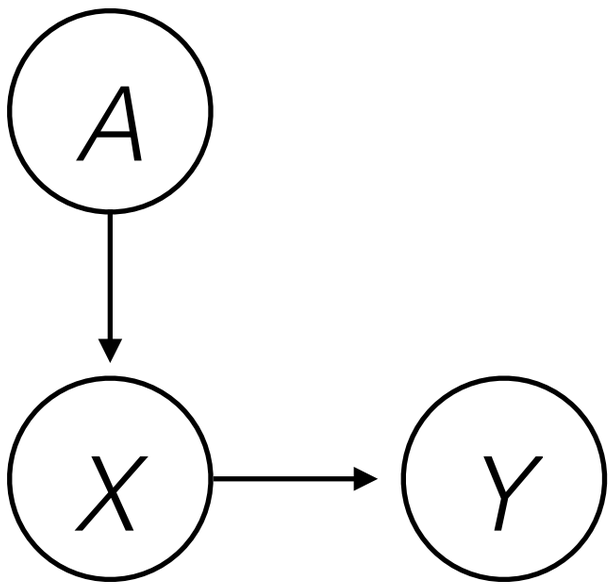
Counterfactual privilege





Revisiting Chief Justice John Roberts

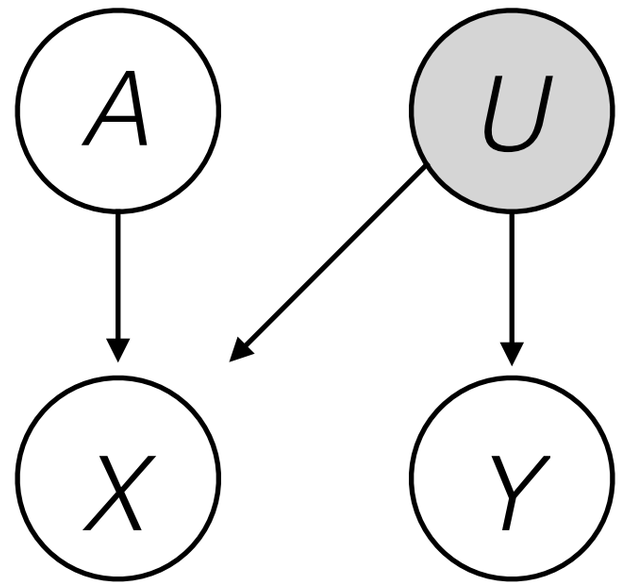
> "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."
>
> Chief Justice John Roberts (2017)



- 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]

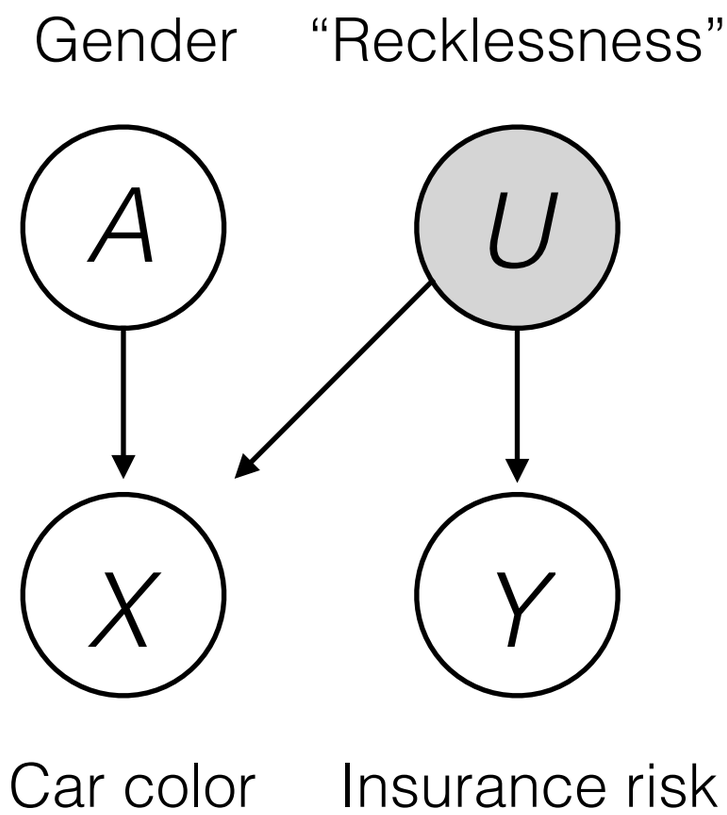


Note that X doesn't cause Y in this model!

- 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

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

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