# Fairness in Machine Learning as a Causal Question

Trenton Chang

EECS 598-009, Causality and Machine Learning, Fall 2023

October 23, 2023

# Table of Contents

# Table of Contents

# Table of Contents

# Table of Contents

# Table of Contents

# Disclaimer

- Fairness is an inherently normative topic
- Talking about fairness means covering some sensitive topics
- We're all from different backgrounds and probably won't agree on everything, and that's ok

# Goals & ground rules

# Goals & ground rules

- **Goal:** Expose you all to multiple ways to think about machine learning fairness in a causal context

# Goals & ground rules

- **Goal:** Expose you all to multiple ways to think about machine learning fairness in a causal context
- **Anti-goal:** Tell you the "right way" to think about fairness.

# Goals & ground rules

- **Goal:** Expose you all to multiple ways to think about machine learning fairness in a causal context
- **Anti-goal:** Tell you the "right way" to think about fairness.
- **Norms for discussion:** Assume good intent from others, and avoid making broad generalizations.

# Today's focus

# Today's focus

Remember the causal inference pipeline...

# Today's focus

Remember the causal inference pipeline...

- Define the estimand of interest

# Today's focus

Remember the causal inference pipeline...

- Define the estimand of interest
- Demonstrate identifiability

# Today's focus

Remember the causal inference pipeline...

- Define the estimand of interest
- Demonstrate identifiability
- Fit a model

# Today's focus

Remember the causal inference pipeline...

- Define the estimand of interest
- Demonstrate identifiability
- Fit a model

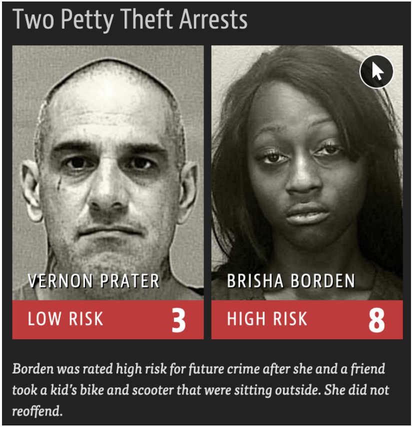
We'll be investigating how we can formulate problems of fairness in machine learning/decision-making as causal questions.

# What is fairness in machine learning?

# Fairness: A Motivating Example

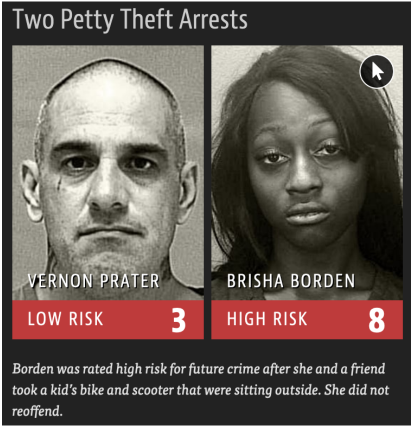
[1] Angwin et al. (2014), "Machine bias,"
`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

# Fairness: A Motivating Example




[1] Angwin et al. (2014), "Machine bias,"
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Fairness: A Motivating Example




An analysis of the COMPAS redicivism risk prediction algorithm highlighted racial bias in the algorithm's outputted risk scores.[1]

[1] Angwin et al. (2014), "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# COMPAS: reconstructed results for Broward County, FL data

[2] Since the algorithm outputs a number 1-10, as well as a bracket "Low," "Medium," and "High," the authors of this analysis treat "Low" as the negative label (did not recidivate) and "Medium/High" as positive. Data reproduced from Larson et al. (2016), "How We Analyzed the COMPAS Recidivism Algorithm," `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

# COMPAS: reconstructed results for Broward County, FL data

We show model[2] performance across groups:

- $\hat{Y} = 0$: predicted to not reoffend
- $\hat{Y} = 1$: predicted to reoffend

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

[2] Since the algorithm outputs a number 1-10, as well as a bracket "Low," "Medium," and "High," the authors of this analysis treat "Low" as the negative label (did not recidivate) and "Medium/High" as positive. Data reproduced from Larson et al. (2016), "How We Analyzed the COMPAS Recidivism Algorithm," `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

# ProPublica's argument

ProPublica argued that the model is discriminatory/unfair, because it makes disproportionate errors among Black defendants:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

[3] Similar results hold for false negative rate.

# ProPublica's argument

ProPublica argued that the model is discriminatory/unfair, because it makes disproportionate errors among Black defendants:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y}=0$ | $\hat{Y}=1$ | $\hat{Y}=0$ | $\hat{Y}=1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

- FPR among White defendants $= \frac{805}{805+990} \approx 44.85\%$
- FPR among Black defendants $= \frac{349}{349+1139} \approx 27.99\%$

[3] Similar results hold for false negative rate.

# ProPublica's argument

ProPublica argued that the model is discriminatory/unfair, because it makes disproportionate errors among Black defendants:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

- FPR among White defendants $= \frac{805}{805+990} \approx 44.85\%$
- FPR among Black defendants $= \frac{349}{349+1139} \approx 27.99\%$

44.85% vs. 27.99% is a pretty (subjectively) large gap![3]

[3] Similar results hold for false negative rate.

# Wait, but is it?

# Wait, but is it?

Someone else might argue that the model is fair, because it has similar precision/PPV across groups:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

# Wait, but is it?

Someone else might argue that the model is fair, because it has similar precision/PPV across groups:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

- PPV among White defendants $= \frac{805}{805+1369} \approx 62.97\%$
- PPV among Black defendants $= \frac{505}{349+505} \approx 59.13\%$

# Wait, but is it?

Someone else might argue that the model is fair, because it has similar precision/PPV across groups:

| Ground truth | White defendants | | Black defendants | |
|---|---|---|---|---|
| | $\hat{Y} = 0$ | $\hat{Y} = 1$ | $\hat{Y} = 0$ | $\hat{Y} = 1$ |
| Did not reoffend | 990 | 805 | 1139 | 349 |
| Recidivated | 532 | 1369 | 461 | 505 |

Table: Confusion matrix by race of the COMPAS algorithm.

- PPV among White defendants $= \frac{805}{805+1369} \approx 62.97\%$
- PPV among Black defendants $= \frac{505}{349+505} \approx 59.13\%$

OK, 62.97% and 59.13% are relatively close...

# OK, now what?

This is a predictive model that has real impacts on peoples' lives. We'd like a better resolution to this discrepancy than "we added up different numbers and got different results."

# OK, now what?

This is a predictive model that has real impacts on peoples' lives. We'd like a better resolution to this discrepancy than "we added up different numbers and got different results."

**The trouble with fairness**

Clearly, defining "fairness" is subjective. We need some way to formalize our assumptions about what's "fair."

# Observational fairness: notation

There are three main categories of **observational fairness** definitions, which we will encode as [conditional] independence relationships between the following variables:

- $A$: sensitive attribute
- $Y$: any outcome of interest
- $\hat{Y}$: any prediction of the outcome of interest. Commonly assumed to be some function of a set of covariates $X$ (*i.e.*, a model).

# Sensitive attributes

**Definition: Sensitive attribute**

Informally, a **sensitive attribute** is any variable across which some notion of fairness is desirable. **Sensitive attributes** are (in general) implicitly assumed to be binary or categorical.

# Sensitive attributes

## Definition: Sensitive attribute

Informally, a **sensitive attribute** is any variable across which some notion of fairness is desirable. **Sensitive attributes** are (in general) implicitly assumed to be binary or categorical.

## Remark

There is no widely-accepted, mathematically-rigorous definition of a sensitive attribute. Its definition originates in anti-discrimination law (in a U.S. context, where it is called a *protected class*[4]), but is generally hand-waved.

[4] See the Civil Rights Act of 1964.

# Observational definitions of fairness

**Definition: Observationality (informal)**

A fairness criterion is **observational** if it can be written in the form $f(P(A, Y, \hat{Y}, X))$ for some functional $f$.

# Observational definitions of fairness

**Definition: Observationality (informal)**

A fairness criterion is **observational** if it can be written in the form $f(P(A, Y, \hat{Y}, X))$ for some functional $f$.

**Remark**

Intuitively, we can express **observational definitions of fairness** in terms of joint/conditional probability statements.

# The three categories of observational fairness criteria

# The three categories of observational fairness criteria

Most observational fairness definitions can be encoded as the following (conditional) independence conditions:

- Independence: $\hat{Y} \perp\!\!\!\perp A$
- Separation: $\hat{Y} \perp\!\!\!\perp A \mid Y$
- Sufficiency: $Y \perp\!\!\!\perp A \mid \hat{Y}$

# Independence

**Definition: Independence**

A set of predictions $\hat{Y}$ satisfies **independence** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A$.

# Independence

## Definition: Independence

A set of predictions $\hat{Y}$ satisfies **independence** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A$.

## Intuition

Consider a machine learning model for predicting the risk of a heart attack. **Independence** with respect to race that the model's outputs should be identical for Black and White patients.

# Independence

## Definition: Independence

A set of predictions $\hat{Y}$ satisfies **independence** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A$.

## Intuition

Consider a machine learning model for predicting the risk of a heart attack. **Independence** with respect to race that the model's outputs should be identical for Black and White patients.

## Other names in the literature

- demographic parity, statistical parity, group fairness, disparate impact

# How people measure independence

## One common empirical measurement of fairness

Follows from the statistical definition of independence; for some pre-specified threshold $\delta > 0$, we have that

$$\forall\, (a, a', \hat{y}). \quad \left| P(\hat{Y} = \hat{y} \mid A = a) - P(\hat{Y} = \hat{y} \mid A = a') \right| \leq \delta.$$

# How people measure independence

### One common empirical measurement of fairness

Follows from the statistical definition of independence; for some pre-specified threshold $\delta > 0$, we have that

$$\forall\, (a, a', \hat{y}). \quad \left| P(\hat{Y} = \hat{y} \mid A = a) - P(\hat{Y} = \hat{y} \mid A = a') \right| \leq \delta.$$

- Assumes we fully observe all $A$ and $\hat{Y}$.
- Empirical measurements of other fairness criteria proceed similarly for other definitions (adding the assumption that $Y$ is fully observed).
- There are other ways to measure fairness as well (less common in my observation), *e.g.*, MMD, $f$-divergences, mutual information.

# Separation

## Definition: Separation

A set of predictions $\hat{Y}$ satisfies **separation** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A \mid Y$.

# Separation

## Definition: Separation

A set of predictions $\hat{Y}$ satisfies **separation** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A \mid Y$.

## Intuition

For binary $Y$; this is equivalent to equal false positive/negative rates.

# Separation

## Definition: Separation

A set of predictions $\hat{Y}$ satisfies **separation** with respect to sensitive attribute $A$ if $\hat{Y} \perp\!\!\!\perp A \mid Y$.

## Intuition

For binary $Y$; this is equivalent to equal false positive/negative rates.

## Other names in the literature

- Error rate parity/equality of error rates, false positive/negative error rate balance, equalized odds. See Verma (2018) for more.[a]

[a] Verma, S., & Rubin, J. (2018). Fairness definitions explained.

# Sufficiency

**Definition: Sufficiency**

A set of predictions $\hat{Y}$ satisfies **sufficiency** with respect to sensitive attribute $A$ if $Y \perp\!\!\!\perp A \mid \hat{Y}$.

# Sufficiency

## Definition: Sufficiency

A set of predictions $\hat{Y}$ satisfies **sufficiency** with respect to sensitive attribute $A$ if $Y \perp\!\!\!\perp A \mid \hat{Y}$.

## Intuition

This is equivalent to enforcing identical calibration curves across groups.

# Sufficiency

## Definition: Sufficiency

A set of predictions $\hat{Y}$ satisfies **sufficiency** with respect to sensitive attribute $A$ if $Y \perp\!\!\!\perp A \mid \hat{Y}$.

## Intuition

This is equivalent to enforcing identical calibration curves across groups.

## Other names in the literature

- Statistical calibration, group calibration, predictive parity.

# The limits of observational definitions

# The 1973 Berkeley admissions case study

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applied | Admitted (%) | Applied | Admitted (%) |
| **Total** | 2651 | 44 | 1835 | 30 |

[5] Reproduced from Barocas, Hardt, and Narayanan (2019).

# The 1973 Berkeley admissions case study

**Question**

Given the information we have, which definition of fairness could we apply to this example?

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applied | Admitted (%) | Applied | Admitted (%) |
| **Total** | 2651 | 44 | 1835 | 30 |

[5] Reproduced from Barocas, Hardt, and Narayanan (2019).

# The 1973 Berkeley admissions case study

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applied | Admitted (%) | Applied | Admitted (%) |
| **Total** | 2651 | 44 | 1835 | 30 |

[5] Reproduced from Barocas, Hardt, and Narayanan (2019).

# The 1973 Berkeley admissions case study

**Question**

What happens when we apply our fairness definition at the department level?

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applied | Admitted (%) | Applied | Admitted (%) |
| **Total** | 2651 | 44 | 1835 | 30 |
| A | 825 | 62 | 108 | 82 |
| B | 520 | 60 | 25 | 68 |
| C | 325 | 37 | 593 | 34 |
| D | 417 | 33 | 375 | 35 |
| E | 191 | 28 | 393 | 24 |
| F | 373 | 6 | 341 | 7 |

Table: UC Berkeley admissions data from 1973[5]

[5] Reproduced from Barocas, Hardt, and Narayanan (2019).

# Observational definitions of fairness are not explanations

- When we tried to apply a naive fairness definition to evaluate the fairness of UC Berkeley admissions decisions from 1973, we ran into *Simpson's paradox*.

# Observational definitions of fairness are not explanations

- When we tried to apply a naive fairness definition to evaluate the fairness of UC Berkeley admissions decisions from 1973, we ran into *Simpson's paradox*.
- There are a bunch of potential explanations for why this difference occurs, but it is impossible to tell from the table if these are true.
- Observational definitions of fairness can tell us whether a disparity exists, but are *not* explanations.

# DAGs to the rescue? Graphical discrimination analysis

# Why DAGS?

- DAGs encode beliefs about "how the world works" (*i.e.*, counterfactuals help us model *what-if* scenarios).

# Why DAGS?

- DAGs encode beliefs about "how the world works" (*i.e.*, counterfactuals help us model *what-if* scenarios).
- DAGs imply a set of [conditional] independencies.

# Why DAGS?

- DAGs encode beliefs about "how the world works" (*i.e.*, counterfactuals help us model *what-if* scenarios).
- DAGs imply a set of [conditional] independencies.
- Hey, wait a minute, fairness definitions are also encoded as [conditional] independencies.

# Why DAGS?

- DAGs encode beliefs about "how the world works" (*i.e.*, counterfactuals help us model *what-if* scenarios).
- DAGs imply a set of [conditional] independencies.
- Hey, wait a minute, fairness definitions are also encoded as [conditional] independencies.

# Why DAGS?

**The punchline**

Thus, **our prior beliefs in “how the world works/should work”** can help us **choose a fairness definition**—and in turn figure out what constraints we can impose on estimation/modeling.


BUILD - A - DAG
$X_1$ SEASON
SPRINKLER $X_3$
$X_2$ RAIN
$X_4$ WET
$X_5$ SLIPPERY
Causal inference
EST. 2023!

Turn to your neighbor and discuss a potential DAG for the Berkeley admissions case study. Use (at least) these variables:

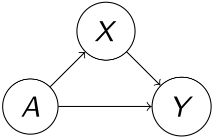
$$A = \text{gender as reported on the application form}$$
$$X = \text{department choice}$$
$$Y = \text{admission decision}$$

# Posing discrimination as a causal question

Let's use this causal DAG to model the Berkeley example:

# Posing discrimination as a causal question

Let's use this causal DAG to model the Berkeley example:



# Posing discrimination as a causal question

Let's use this causal DAG to model the Berkeley example:

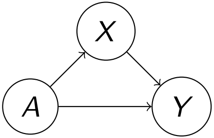
$$
\begin{tikzcd}
 & X \arrow[dr] & \\
A \arrow[ur] \arrow[rr] & & Y
\end{tikzcd}
$$

**Question**

In causal language, what is one way we could argue that there is/isn't any discrimination? *Hint: Think about hypothetical values of causal effects.*

# Posing discrimination as a causal question

Let's use this causal DAG to model the Berkeley example:




**Question**

In causal language, what is one way we could argue that there is/isn't any discrimination? *Hint: Think about hypothetical values of causal effects.*

The causal effect of $A$ on $Y$ is zero.

# Counterfactual definitions of fairness

Formally, we might come up with counterfactual versions of observational fairness definitions by replacing $Y$ with $Y(a)$ in our existing observational definitions of fairness, *e.g.*,

$$Y(a) \perp\!\!\!\perp A \tag{1}$$

for counterfactual independence (if the applicant's gender had been different from what was observed, their admission status should not change).

[6] Corbett-Davies, Gaebler and Nilforoshan (2018). "The Measure and Mismeasure of Fairness."

# Counterfactual definitions of fairness

Formally, we might come up with counterfactual versions of observational fairness definitions by replacing $Y$ with $Y(a)$ in our existing observational definitions of fairness, *e.g.*,

$$Y(a) \perp\!\!\!\perp A \tag{1}$$

for counterfactual independence (if the applicant's gender had been different from what was observed, their admission status should not change).

Similar extensions can be applied to the other definitions.[6]

[6] Corbett-Davies, Gaebler and Nilforoshan (2018). "The Measure and Mismeasure of Fairness."

# Counterfactual definitions of fairness

Formally, we might come up with counterfactual versions of observational fairness definitions by replacing $Y$ with $Y(a)$ in our existing observational definitions of fairness, *e.g.*,

$$Y(a) \perp\!\!\!\perp A \tag{1}$$

for counterfactual independence (if the applicant's gender had been different from what was observed, their admission status should not change).

Similar extensions can be applied to the other definitions.[6]

**Takeaway**

Counterfactual fairness asserts that "**If a sensitive attribute had been different, there would be no effect on the outcome.** (potentially conditional on other information)"

[6] Corbett-Davies, Gaebler and Nilforoshan (2018). "The Measure and Mismeasure of Fairness."

# Conducting graphical discrimination analysis

# Conducting graphical discrimination analysis

- First, construct a DAG representing our *belief* in the mechanism of discrimination

# Conducting graphical discrimination analysis

- First, construct a DAG representing our *belief* in the mechanism of discrimination
- Using the DAG, we turn identifying **fairness or discrimination** into **identifying a causal effect**. We know how to do that!

# Conducting graphical discrimination analysis

- First, construct a DAG representing our *belief* in the mechanism of discrimination
- Using the DAG, we turn identifying **fairness or discrimination** into **identifying a causal effect**. We know how to do that!

*(one hypothetical fairness measurement in our setting)*

$$\left| \frac{1}{N} \sum_{i=1}^{N} P(Y \mid A = \text{male}) - P(Y \mid A = \text{female}) \right|$$

# Interpreting graphical discrimination analysis

- Under our causal assumptions, “fairness” is defined in terms of a null causal effect.
  - **Counterfactual interpretation:** “Intervening” to change a sensitive attribute should not affect the outcome.

# So, we identified a causal effect. Can we go home?

# So, we identified a causal effect. Can we go home?

- Under consistency, no unmeasured confounders, and positivity; the causal effect of gender on acceptance to graduate school at Berkeley in 1973 is *identifiable*.

# So, we identified a causal effect. Can we go home?

- Under consistency, no unmeasured confounders, and positivity; the causal effect of gender on acceptance to graduate school at Berkeley in 1973 is *identifiable*.
- But this doesn't really explain why/how discrimination arises...

# So, we identified a causal effect. Can we go home?

- Under consistency, no unmeasured confounders, and positivity; the causal effect of gender on acceptance to graduate school at Berkeley in 1973 is *identifiable*.
- But this doesn't really explain why/how discrimination arises...
- And, doesn't help us with our original issue—even with causal assumptions, it's not clear how we can *explain* discrimination (yet)!

# Structural discrimination: University of Adversaria

# Structural discrimination: University of Adversaria

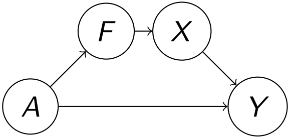
- The University of Adversaria systematically reduces funding to programs that attract more female applicants.
- This artificially reduces acceptance rates in such departments.

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...

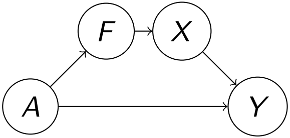
# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...



# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...

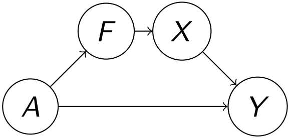



## Question

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...



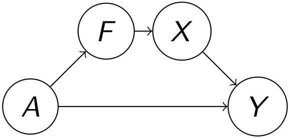

## Question

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

1. Capturing the "strength" of the path $A \to Y$?

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...



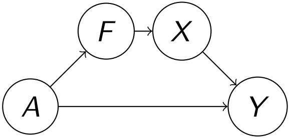

## Question

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

1. Capturing the "strength" of the path $A \to Y$?
2. Capturing the "strength" of the path $A \to F \to X \to Y$?

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...



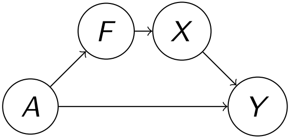

**Question**

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

1. Capturing the "strength" of the path $A \to Y$?
2. Capturing the "strength" of the path $A \to F \to X \to Y$?
3. Some combination of both?

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...



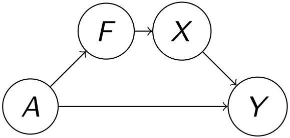

**Question**

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

1. Capturing the "strength" of the path $A \to Y$?
2. Capturing the "strength" of the path $A \to F \to X \to Y$?
3. Some combination of both?

# Structural discrimination, continued

Let's add a "funding mechanism" node $F$ to our DAG...




## Question

In this DAG, when we estimate the causal effect of $A$ on $Y$, do we...

1. Capturing the "strength" of the path $A \to Y$?
2. Capturing the "strength" of the path $A \to F \to X \to Y$?
3. Some combination of both?

# Conflating sources of discrimination

Our naive estimate of the effect of $A$ and $Y$ can conflate two sources of discrimination:

# Conflating sources of discrimination

Our naive estimate of the effect of $A$ and $Y$ can conflate two sources of discrimination:

- $A \to Y$: systematic, direct gender discrimination (*taste-based discrimination*)

# Conflating sources of discrimination

Our naive estimate of the effect of $A$ and $Y$ can conflate two sources of discrimination:

- $A \to Y$: systematic, direct gender discrimination (*taste-based discrimination*)
- $A \to F \to X \to Y$: indirect gender discrimination due to structural factors (*structural discrimination*)

# Conflating sources of discrimination

Our naive estimate of the effect of $A$ and $Y$ can conflate two sources of discrimination:

- $A \rightarrow Y$: systematic, direct gender discrimination (*taste-based discrimination*)
- $A \rightarrow F \rightarrow X \rightarrow Y$: indirect gender discrimination due to structural factors (*structural discrimination*)

**Question**

Why might separating out these two sources of discrimination be useful? *I.e.*, isn't all discrimination bad?

# Conflating sources of discrimination

Our naive estimate of the effect of $A$ and $Y$ can conflate two sources of discrimination:

- $A \rightarrow Y$: systematic, direct gender discrimination (*taste-based discrimination*)
- $A \rightarrow F \rightarrow X \rightarrow Y$: indirect gender discrimination due to structural factors (*structural discrimination*)

**Question**

Why might separating out these two sources of discrimination be useful? *I.e.*, isn't all discrimination bad?

Yes, but if we want to design policies that target the underlying *causes* of discrimination, separating these out could be useful.

# Discrimination analysis with direct & indirect effects

# The natural direct effect

If we care about discrimination along $A \to Y$, we might want to measure, across *fixed* levels of $X$ (or $F$), the effect of $A$ on $Y$. Perhaps we can fix the levels to $X(a)$, and aggregate.

[7] Pearl, J. (2011) The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms.

# The natural direct effect

If we care about discrimination along $A \to Y$, we might want to measure, across *fixed* levels of $X$ (or $F$), the effect of $A$ on $Y$. Perhaps we can fix the levels to $X(a)$, and aggregate.

For an arbitrary mediator $M$ (*i.e.*, $M \in \{X, F\}$), this is the **natural direct effect**:[7]

[7] Pearl, J. (2011) The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms.

# The natural direct effect

If we care about discrimination along $A \to Y$, we might want to measure, across *fixed* levels of $X$ (or $F$), the effect of $A$ on $Y$. Perhaps we can fix the levels to $X(a)$, and aggregate.

For an arbitrary mediator $M$ (*i.e.*, $M \in \{X, F\}$), this is the **natural direct effect**:[7]

**Natural Direct Effect (NDE)**

$$\begin{aligned} &\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a', M(a'))] \\ &= \sum_m [\mathbb{E}[Y \mid m, a] - \mathbb{E}[Y \mid m, a']] P(m \mid a) \end{aligned}$$

[7] Pearl, J. (2011) The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms.

# The natural direct effect

If we care about discrimination along $A \to Y$, we might want to measure, across *fixed* levels of $X$ (or $F$), the effect of $A$ on $Y$. Perhaps we can fix the levels to $X(a)$, and aggregate.

For an arbitrary mediator $M$ (*i.e.*, $M \in \{X, F\}$), this is the **natural direct effect**:[7]

## Natural Direct Effect (NDE)

$$\mathbb{E}[Y(a, M(a'))] - \mathbb{E}[Y(a', M(a'))]$$
$$= \sum_m [\mathbb{E}[Y \mid m, a] - \mathbb{E}[Y \mid m, a']] P(m \mid a)$$

[7] Pearl, J. (2011) The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms.

# NDE Intuition

Let $a, a'$ correspond to "male, female," respectively (without loss of generality). Set $M := X$ (we are analyzing department choice as a mediator). The NDE is the difference between two terms:

[8] Here, "if" is shorthand for a counterfactual; *i.e.*, what would have happened if we intervened such that some variable takes on a specific value.

# NDE Intuition

Let $a, a'$ correspond to "male, female," respectively (without loss of generality). Set $M := X$ (we are analyzing department choice as a mediator). The NDE is the difference between two terms:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status ($Y$) if[8] male and department choice $X$ is what it would be if individual had been male

[8] Here, "if" is shorthand for a counterfactual; *i.e.*, what would have happened if we intervened such that some variable takes on a specific value.

# NDE Intuition

Let $a, a'$ correspond to "male, female," respectively (without loss of generality). Set $M := X$ (we are analyzing department choice as a mediator). The NDE is the difference between two terms:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status ($Y$) if[8] male and department choice $X$ is what it would be if individual had been male
- $\mathbb{E}[Y(a', X(a))]$: Acceptance status if female and department choice is what it would be for a male applicant

[8] Here, "if" is shorthand for a counterfactual; *i.e.*, what would have happened if we intervened such that some variable takes on a specific value.

# NDE Intuition

Let $a, a'$ correspond to "male, female," respectively (without loss of generality). Set $M := X$ (we are analyzing department choice as a mediator). The NDE is the difference between two terms:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status ($Y$) if[8] male and department choice $X$ is what it would be if individual had been male
- $\mathbb{E}[Y(a', X(a))]$: Acceptance status if female and department choice is what it would be for a male applicant

**Intuition:** The NDE "turns off" the effect of the mediator on the outcome *by fixing it given an intervention*, such that we only capture the effect of gender *directly* on admission.

[8] Here, "if" is shorthand for a counterfactual; *i.e.*, what would have happened if we intervened such that some variable takes on a specific value.

# The natural indirect effect

Whew! That takes care of the path $A \to Y$. What about the path $A \to F \to X \to Y$?

For this one, we can turn to the **natural indirect effect. For an arbitrary mediator** $M \in \{F, X\}$**:**

## Natural Indirect Effect (NIE)

$$\begin{aligned} &\mathbb{E}[Y(a, M(a))] - \mathbb{E}[Y(a, M(a'))] \\ &= \sum_m \mathbb{E}[Y \mid m, a][P(m \mid a') - P(m \mid a)] \end{aligned}$$

# NIE Intuition

Setting $M := X$ again, let's break down each term of the NIE:

# NIE Intuition

Setting $M := X$ again, let's break down each term of the NIE:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status (Y) if male and department choice $X$ is what it would be if individual had been male (same term as NDE)

# NIE Intuition

Setting $M := X$ again, let's break down each term of the NIE:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status (Y) if male and department choice $X$ is what it would be if individual had been male (same term as NDE)
- $\mathbb{E}[Y(a, X(a'))]$: Acceptance status (Y) if male and department choice $X$ is what it would be if individual had been female

What's different from before?

# NIE Intuition

Setting $M := X$ again, let's break down each term of the NIE:

- $\mathbb{E}[Y(a, X(a))]$: Acceptance status (Y) if male and department choice $X$ is what it would be if individual had been male (same term as NDE)
- $\mathbb{E}[Y(a, X(a'))]$: Acceptance status (Y) if male and department choice $X$ is what it would be if individual had been female

What's different from before? For the NDE, the first component of the $Y(\cdot, \cdot)$ counterfactual is different; for the NIE, the $X(a)$ (2nd) component differs.

**Intuition:** The NIE "turns off" the effect of the treatment directly on the outcome, only allowing it to affect $Y(\cdot, \cdot)$ via changes to the mediator (*i.e.*, to $X(\cdot)$).

# The Mediation Formula (Identifiability of the NDE/NIE)

$C$: confounder(s) (cfd.), $M$: mediator (med.)[9]

## Assumptions

1. $\forall a.\ Y(a, M(a)) = Y(a)$ **(composition)**
2. $\forall (a, m).\ A \perp\!\!\!\perp Y(a, m) \mid C$ **(no treatment-outcome cfd.)**
3. $\forall (a, m).\ M \perp\!\!\!\perp Y(a, m) \mid (C, A)$ **(no med.-outcome cfd.)**
4. $\forall a.\ A \perp\!\!\!\perp M(a) \mid C$ **(no treatment-med. cfd.)**
5. $\forall (a, a', m).\ Y(a, m) \perp\!\!\!\perp M(a') \mid C$ **(no "cross-world" confounding)**

[9] Further reading: Ding (2023), A First Course in Causal Inference, Ch. 27. https://arxiv.org/pdf/2305.18793.pdf

# Theorem: Identifiability of the NDE

**Theorem**

Under the previous assumptions (2-5), we have that $\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y \mid A = a, M = m] P(M = m \mid A = a')$.

# Theorem: Identifiability of the NDE

**Theorem**

Under the previous assumptions (2-5), we have that $\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y \mid A = a, M = m]P(M = m \mid A = a')$.

$$\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y(a, M(a')) \mid M(a') = m]P(M(a') = m)$$

# Theorem: Identifiability of the NDE

## Theorem

Under the previous assumptions (2-5), we have that
$\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y \mid A = a, M = m] P(M = m \mid A = a')$.

$$
\begin{aligned}
\mathbb{E}[Y(a, M(a'))] &= \sum_m \mathbb{E}[Y(a, M(a')) \mid M(a') = m] P(M(a') = m) \\
&= \sum_m \mathbb{E}[Y(a, m) \mid M(a') = m] \underbrace{P(M = m \mid A = a')}_{A \perp\!\!\!\perp M(a) | C + \text{consistency}}
\end{aligned}
$$

# Theorem: Identifiability of the NDE

**Theorem**

Under the previous assumptions (2-5), we have that $\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y \mid A = a, M = m]P(M = m \mid A = a')$.

$$\begin{aligned}
\mathbb{E}[Y(a, M(a'))] &= \sum_m \mathbb{E}[Y(a, M(a')) \mid M(a') = m]P(M(a') = m) \\
&= \sum_m \mathbb{E}[Y(a, m) \mid M(a') = m] \underbrace{P(M = m \mid A = a')}_{A \perp\!\!\!\perp M(a) \mid C + \text{consistency}} \\
&= \sum_m \underbrace{\mathbb{E}[Y(a, m)]}_{Y(a, M(a)) \perp\!\!\!\perp M(a') \mid C} P(M = m \mid A = a')
\end{aligned}$$

# Theorem: Identifiability of the NDE

## Theorem

Under the previous assumptions (2-5), we have that
$\mathbb{E}[Y(a, M(a'))] = \sum_m \mathbb{E}[Y \mid A = a, M = m] P(M = m \mid A = a')$.

$$
\begin{aligned}
\mathbb{E}[Y(a, M(a'))] &= \sum_m \mathbb{E}[Y(a, M(a')) \mid M(a') = m] P(M(a') = m) \\
&= \sum_m \mathbb{E}[Y(a, m) \mid M(a') = m] \underbrace{P(M = m \mid A = a')}_{A \perp\!\!\!\perp M(a) \mid C + \text{consistency}} \\
&= \sum_m \underbrace{\mathbb{E}[Y(a, m)]}_{Y(a, M(a)) \perp\!\!\!\perp M(a') \mid C} P(M = m \mid A = a') \\
&= \sum_m \underbrace{\mathbb{E}[Y \mid A = a, M = m]}_{A \perp\!\!\!\perp Y(a,m) \mid C \wedge M \perp\!\!\!\perp Y(a,m) \mid (C,A)} P(M = m \mid A = a').
\end{aligned}
$$

$\square$

# How focusing on the ATE conflates NDE and NIE

Recall that the motivation for the NDE and NIE was that we were conflating two sources of discrimination:

# How focusing on the ATE conflates NDE and NIE

Recall that the motivation for the NDE and NIE was that we were conflating two sources of discrimination:

- $A \rightarrow Y$: systematic, direct gender discrimination (*taste-based discrimination*)

# How focusing on the ATE conflates NDE and NIE

Recall that the motivation for the NDE and NIE was that we were conflating two sources of discrimination:

- $A \rightarrow Y$: systematic, direct gender discrimination (*taste-based discrimination*)
- $A \rightarrow F \rightarrow X \rightarrow Y$: indirect gender discrimination due to structural factors (*structural discrimination*)

It turns out, we can write that the $\text{ATE} = \text{NDE} + \text{NIE}$:

$$\begin{aligned}\text{ATE} = \mathbb{E}[Y(a)] - \mathbb{E}[Y(a')] &= \mathbb{E}[Y(a, X(a))] - \mathbb{E}[Y(a', X(a'))] \\ &= \mathbb{E}[Y(a, M(a))] - \mathbb{E}[Y(a, M(a'))] + \mathbb{E}[Y(a, M(a'))] \\ &\quad - \mathbb{E}[Y(a', M(a'))] = \text{NIE} + \text{NDE}.\end{aligned}$$

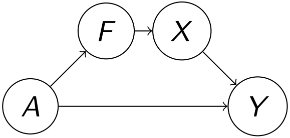
# Beyond NIE (informal)

Recall our DAG for this problem..

[10] The identifiability conditions here get somewhat involved; for more details, consult Nabi and Shipster (2017), "Fair Inference on Outcomes" and Pearl (2005), "Direct and Indirect Effects," Section 3.7.

# Beyond NIE (informal)

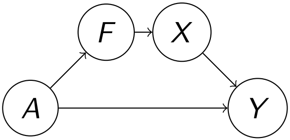
Recall our DAG for this problem..




[10] The identifiability conditions here get somewhat involved; for more details, consult Nabi and Shipster (2017), "Fair Inference on Outcomes" and Pearl (2005), "Direct and Indirect Effects," Section 3.7.

# Beyond NIE (informal)

Recall our DAG for this problem..

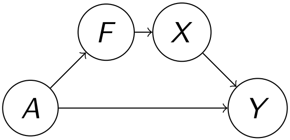



- There are only two paths from $A \to Y$: $A \to Y$ itself and $A \to F \to X \to Y$

[10] The identifiability conditions here get somewhat involved; for more details, consult Nabi and Shipster (2017), "Fair Inference on Outcomes" and Pearl (2005), "Direct and Indirect Effects," Section 3.7.

# Beyond NIE (informal)

Recall our DAG for this problem..

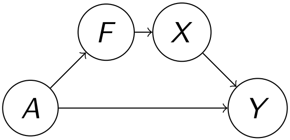



- There are only two paths from $A \to Y$: $A \to Y$ itself and $A \to F \to X \to Y$
- If we have more than two paths, we can generalize the NIE to **path-specific effects**

[10] The identifiability conditions here get somewhat involved; for more details, consult Nabi and Shipster (2017), "Fair Inference on Outcomes" and Pearl (2005), "Direct and Indirect Effects," Section 3.7.

# Beyond NIE (informal)

Recall our DAG for this problem..




- There are only two paths from $A \to Y$: $A \to Y$ itself and $A \to F \to X \to Y$
- If we have more than two paths, we can generalize the NIE to **path-specific effects**
- This is done by "turning off" causal effects along all paths—except the one we care about.[10]

[10] The identifiability conditions here get somewhat involved; for more details, consult Nabi and Shipster (2017), "Fair Inference on Outcomes" and Pearl (2005), "Direct and Indirect Effects," Section 3.7.

# (Bonus) Pitfalls of using sensitive attributes in causal inference

Recall: in causal fairness analysis, what do we usually define as the "**treatment**?"

Recall: in causal fairness analysis, what do we usually define as the "**treatment**?"

A **sensitive attribute!**

Recall: in causal fairness analysis, what do we usually define as the "**treatment**?"

A **sensitive attribute!**

But a treatment is an *intervention*, which raises questions:

- How can we intervene on someone's demographics?
- Is causal inference even well-defined when treatment is defined as a (presumably immutable) sensitive attribute?
- Is this *purely* a philosophical problem, or can it have real implications on causal effect estimation?

# Counterfactuals: parallel universes, kinda (sorry, physics!)

# Counterfactuals: parallel universes, kinda (sorry, physics!)

Consider the counterfactuals $Y(a)$ and $Y(a')$. Imagine two parallel universes that "split off" at the time of intervention (setting $A = a$ or $a'$:

# Counterfactuals: parallel universes, kinda (sorry, physics!)

Consider the counterfactuals $Y(a)$ and $Y(a')$. Imagine two parallel universes that "split off" at the time of intervention (setting $A = a$ or $a'$:

- The universe where $Y = Y(a)$ is our world (wlog)

# Counterfactuals: parallel universes, kinda (sorry, physics!)

Consider the counterfactuals $Y(a)$ and $Y(a')$. Imagine two parallel universes that "split off" at the time of intervention (setting $A = a$ or $a'$:

- The universe where $Y = Y(a)$ is our world (wlog)
- The counterfactual universe is the one where $Y = Y(a')$

# Counterfactuals: parallel universes, kinda (sorry, physics!)

Consider the counterfactuals $Y(a)$ and $Y(a')$. Imagine two parallel universes that "split off" at the time of intervention (setting $A = a$ or $a'$:

- The universe where $Y = Y(a)$ is our world (wlog)
- The counterfactual universe is the one where $Y = Y(a')$

**The only thing different about these universes is the intervention!**

# Counterfactuals: parallel universes, kinda (sorry, physics!)

Consider the counterfactuals $Y(a)$ and $Y(a')$. Imagine two parallel universes that "split off" at the time of intervention (setting $A = a$ or $a'$:

- The universe where $Y = Y(a)$ is our world (wlog)
- The counterfactual universe is the one where $Y = Y(a')$

**The only thing different about these universes is the intervention!** This sounds good for something like a clinical trial with an RCT design...

# The paradox of sensitive attributes as interventions

Suppose that we have a study to determine the effect of race (White vs. Black) on hiring decisions. We use submitted resumes to collect data.[11]

[11] Loosely based on Bertrand and Mullainathan (2004), "Are Emily and Greg more employable than Lakisha and Jamal?"

# The paradox of sensitive attributes as interventions

Suppose that we have a study to determine the effect of race (White vs. Black) on hiring decisions. We use submitted resumes to collect data.[11]

In "counterfactual land," we are asking "If candidate X had been the other race, *all else being equal*, what would be the causal effect?"

[11] Loosely based on Bertrand and Mullainathan (2004), "Are Emily and Greg more employable than Lakisha and Jamal?"

# The paradox of sensitive attributes as interventions

Suppose that we have a study to determine the effect of race (White vs. Black) on hiring decisions. We use submitted resumes to collect data.[11]

In "counterfactual land," we are asking "If candidate X had been the other race, *all else being equal*, what would be the causal effect?"

**"All else being equal"**

In this case, what do you think "all else being equal" means?

[11] Loosely based on Bertrand and Mullainathan (2004), "Are Emily and Greg more employable than Lakisha and Jamal?"

# The paradox of sensitive attributes as interventions

Suppose that we have a study to determine the effect of race (White vs. Black) on hiring decisions. We use submitted resumes to collect data.[11]

In "counterfactual land," we are asking "If candidate X had been the other race, *all else being equal*, what would be the causal effect?"

**"All else being equal"**

In this case, what do you think "all else being equal" means?

Potential positivity violation—not clear if such an individual exists, nor is *intervening* on race well-defined!

[11] Loosely based on Bertrand and Mullainathan (2004), "Are Emily and Greg more employable than Lakisha and Jamal?"

# A potential resolution through social constructivism

[12] Further reading with respect to race: Omi and Winant (1985), Racial Formation in the United States, Ch. 4

[13] Here, I slightly disagree with Kasirzadeh and Smart (2021); see their paper for a counterpoint.

# A potential resolution through social constructivism

**Main idea:** Social categories such as *race* do not have inherent physical grounding, but rather physical/real-world objects are "given" extraneous meaning via societal norms, policies, or laws.[12]

[12] Further reading with respect to race: Omi and Winant (1985), Racial Formation in the United States, Ch. 4

[13] Here, I slightly disagree with Kasirzadeh and Smart (2021); see their paper for a counterpoint.

# A potential resolution through social constructivism

**Main idea:** Social categories such as *race* do not have inherent physical grounding, but rather physical/real-world objects are "given" extraneous meaning via societal norms, policies, or laws.[12]

**"We get it Trenton, you're a humanities kid; what does this mean for causal inference?"**

[12] Further reading with respect to race: Omi and Winant (1985), Racial Formation in the United States, Ch. 4

[13] Here, I slightly disagree with Kasirzadeh and Smart (2021); see their paper for a counterpoint.

# A potential resolution through social constructivism

**Main idea:** Social categories such as *race* do not have inherent physical grounding, but rather physical/real-world objects are "given" extraneous meaning via societal norms, policies, or laws.[12]

**"We get it Trenton, you're a humanities kid; what does this mean for causal inference?"** When we say we want to measure the causal effect of *race*, *gender*, or some other social category on an outcome—*race/gender* are simply *shorthand*/abbreviations for some *aspect* of race/gender/etc.[13]

[12] Further reading with respect to race: Omi and Winant (1985), Racial Formation in the United States, Ch. 4

[13] Here, I slightly disagree with Kasirzadeh and Smart (2021); see their paper for a counterpoint.

# Revisiting racial bias in resume screening

We can resolve this issue for the resume screening example as follows:

- "race" $\rightarrow$ "racial perception of name by evaluator"

# Takeaways

**Be precise!**

- Clearly state what *effect* we're trying to measure when we treat a sensitive attribute as a variable.
- This means clearly defining what aspect of a sensitive attribute that you care about (*e.g.*, a decision-maker's *perception* of race, someone's *self-reported* gender, biological sex)

# Conclusion: What we learned today

**Closing share-out**

Turn to a neighbor and discuss what you learned today!

# Conclusion: What we learned today

**Closing share-out**

Turn to a neighbor and discuss what you learned today!

My takeaways:

- We motivated ways to define fairness from a non-causal and causal perspective
- We discussed how causal fairness is a matter of testing for a null causal effect (if a person had different characteristics, the outcome shouldn't change)
- We highlight different causal effects (vanilla ATE, NDE, and NIE) to estimate when thinking about causal fairness