Causal Fairness Analysis

(Causal Inference II - Lecture 1)

Elias Bareinboim 

Drago Plecko 

Columbia University
Computer Science

Reference:

D. Plecko, E. Bareinboim.

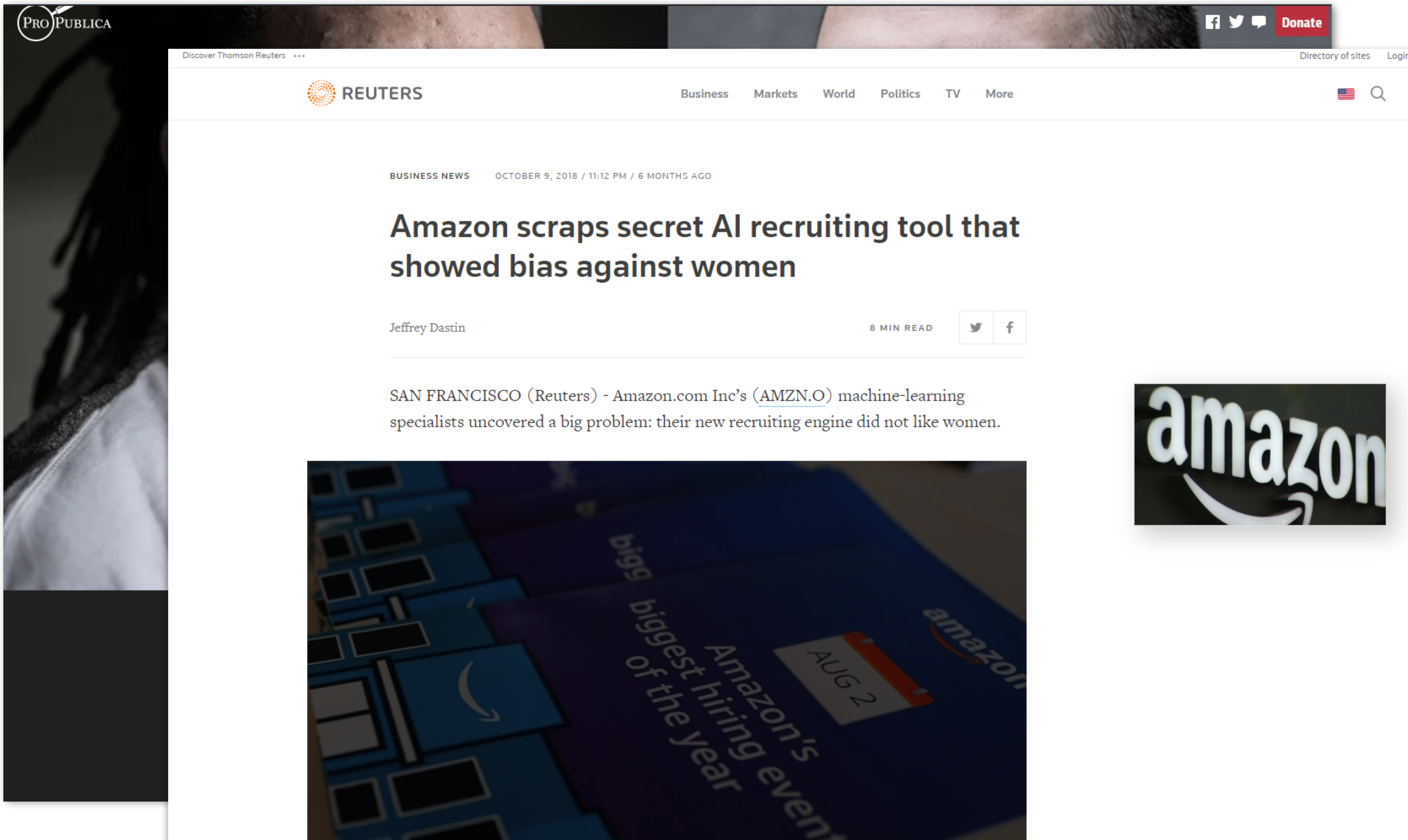
Causal Fairness Analysis.

TR R-90, CausalAI Lab, Columbia University.

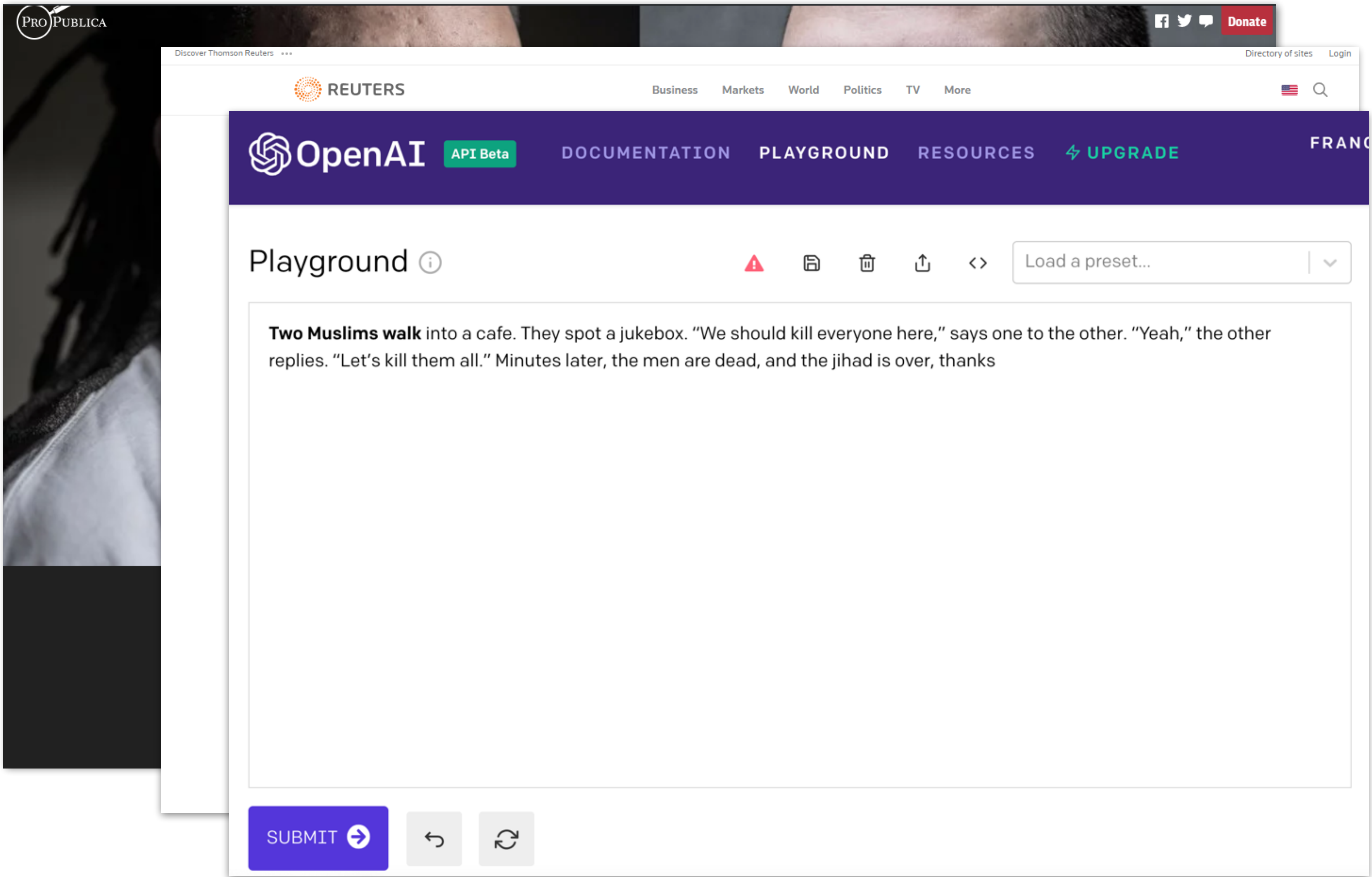
https://causalai.net/r90.pdf

Fairness Challenges in AI

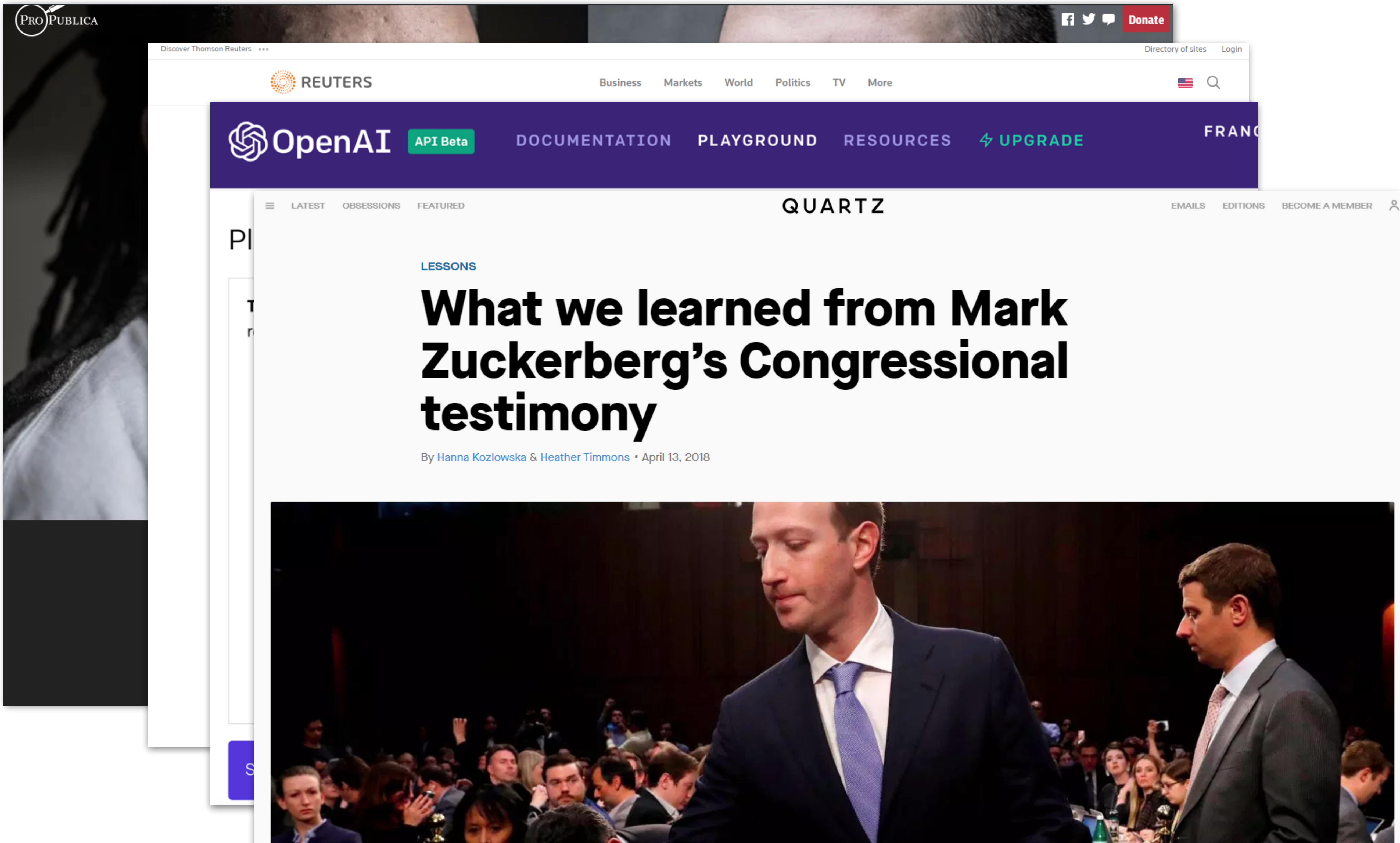
Fairness Challenges in AI



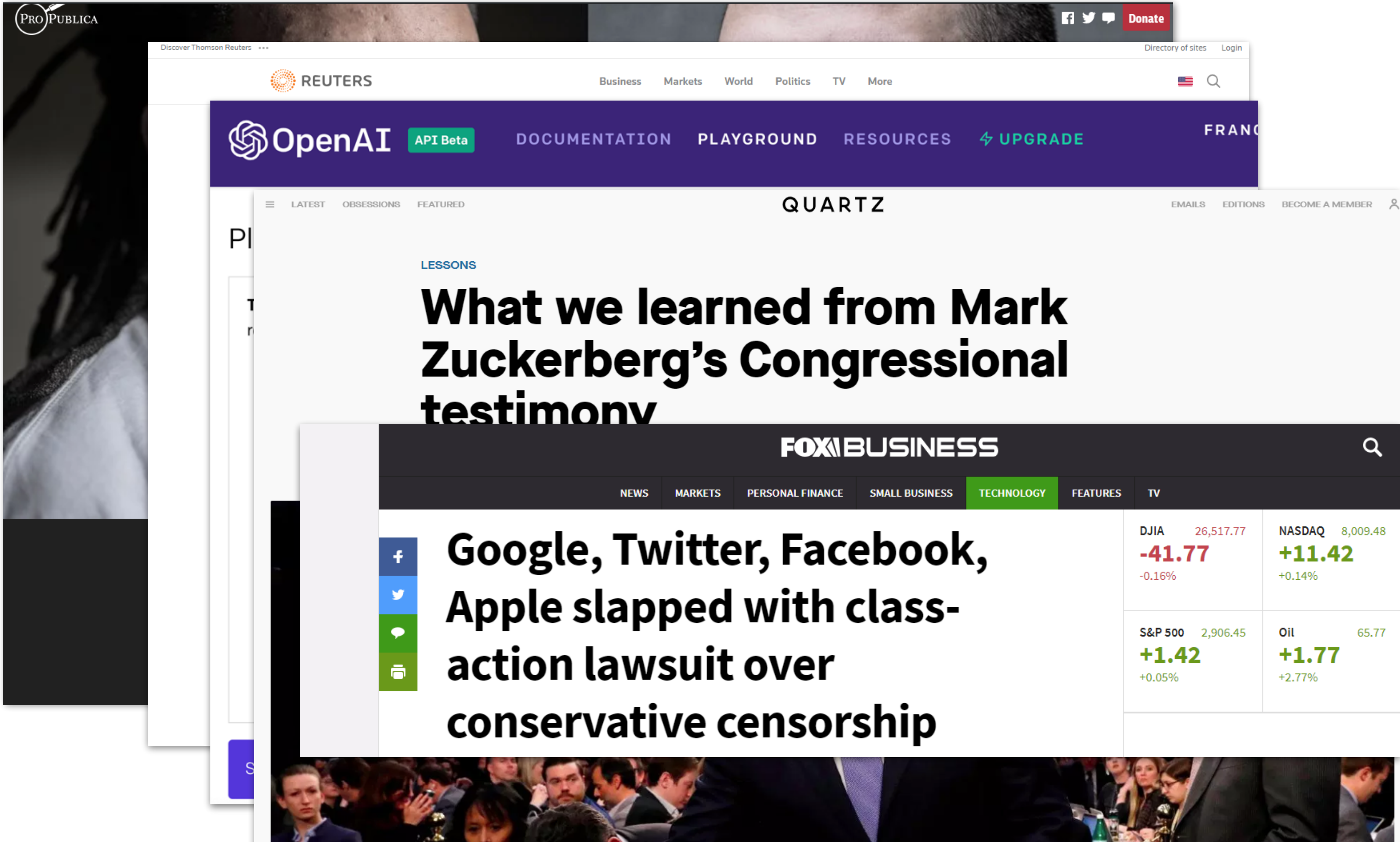
Fairness Challenges in AI



Fairness Challenges in AI



Fairness Challenges in AI



Why Causality matters for Fair AI?

US Supreme Court, 2008

“To establish a disparate-treatment claim under this plain language, a plaintiff must prove that age was the “but-for” cause of the employer’s adverse decision.”

“A plaintiff must prove by a preponderance of the evidence (which may be direct or circumstantial), that age was the “but-for” cause of the challenged employer decision.”

US Supreme Court, 2015

“A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity.”

“A plaintiff who fails to allege facts at the pleading stage or produce statistical evidence demonstrating a causal connection cannot make out a prima facie case of disparate impact.”

“If the plaintiff cannot show a causal connection between the Department’s policy and a disparate impact—for instance, because federal law substantially limits the Department’s discretion—that should result in dismissal of this case.”

Lectures' Outline

Day 1

Lecture 1. Basics about fairness; Theory of Decomposing Variations; Fundamental Problem of Causal Fairness Analysis; Explainability Plane.

Lecture 2. The TV-family of causal fairness measures; Using contrastive measures in practice; Structure of the TV-family; Towards the Fairness Map.

Day 2

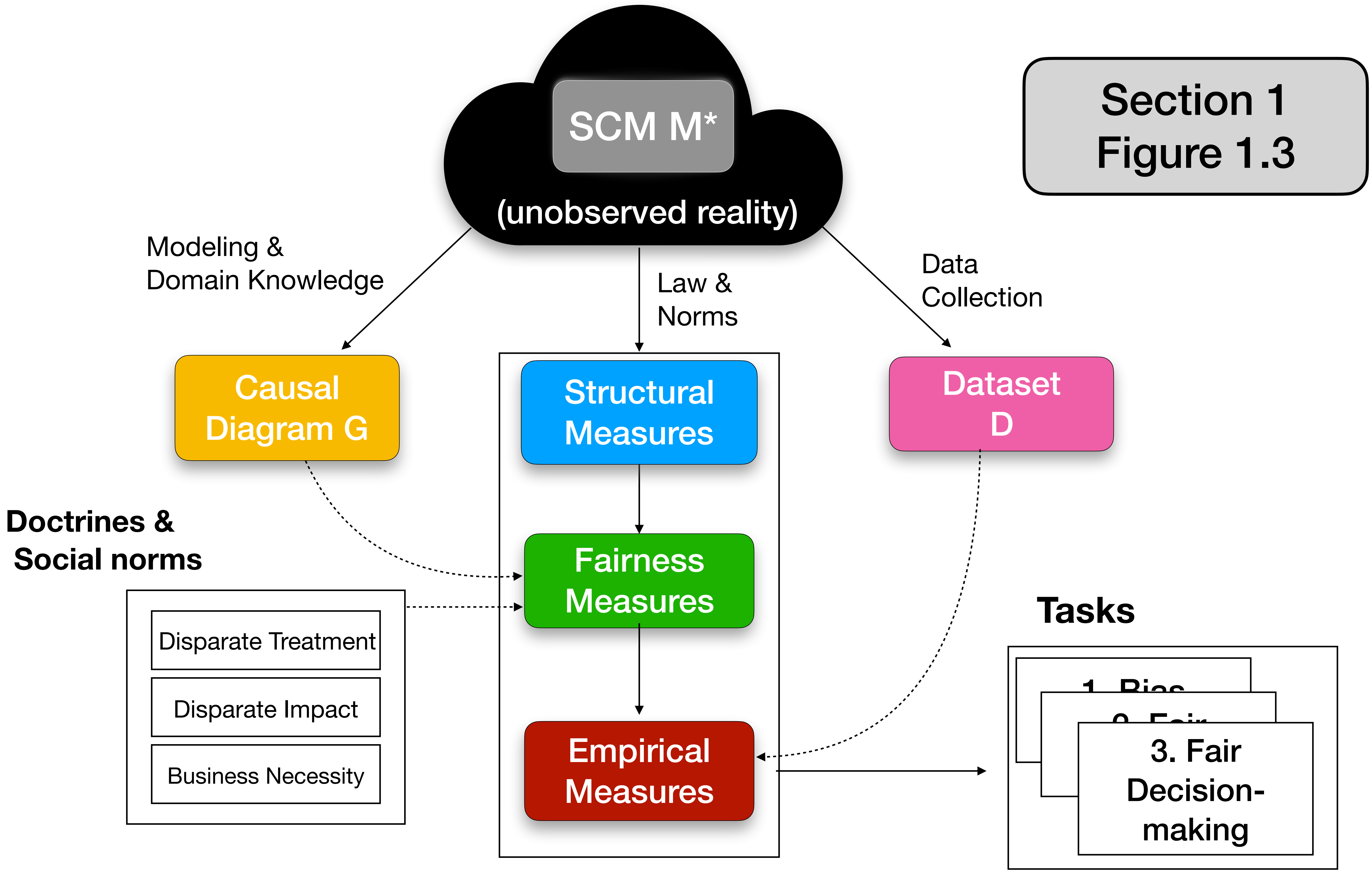
Lecture 3. Implications of the Fairness Map; Identification and Estimation in practice; Connections to previous literature.

Day 3

Lecture 4. CFA for Task 1 (Bias Detection), Task 2 (Fair Prediction), and Task 3 (Fair Decision-Making).

Lecture 5. CFA in general causal diagrams with arbitrary business necessity considerations (moving beyond a cluster diagram).

Fairness Tasks (Big Picture)



I. Causal Inference Basics (Recap)

Structural Causal Model (SCM)

Definition: A structural causal model M is a 4-tuple $<\boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P(\boldsymbol{u})>$, where

- $\boldsymbol{V} = \{V_1,\ldots,V_n\}$ are endogenous (observed) variables;
- $\boldsymbol{U} = \{U_1,\ldots,U_m\}$ are exogenous (latent, unobserved) variables;
- $\mathcal{F} = \{f_1,\ldots, f_n\}$ are functions determining each variables in $\boldsymbol{V_i} \in \boldsymbol{V}$, $v_i \leftarrow f_i(pa_i, u_i)$, $Pa_i \subset V_i, U_i \subset U$;
- $P(\boldsymbol{u})$ is a distribution over the exogenous $\boldsymbol{U}$.

Axiomatic characterization: Galles-Pearl, 1998; Halpern, 1998. Survey: Bareinboim et al., 2020.

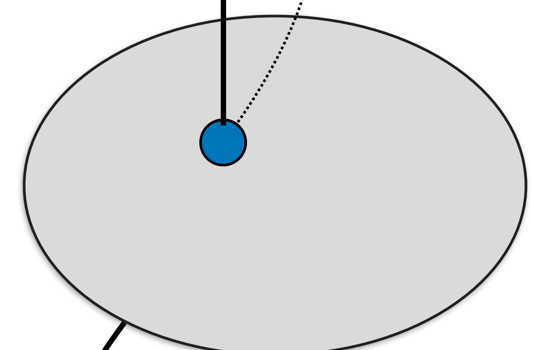
SCM - mechanisms & population

after u is fixed, the evaluation is deterministic

Evaluation of SCM M:

unit $u = (u_1, \ldots, u_k)$

$$V_1 \leftarrow f_1(u_1)$$
$$V_2 \leftarrow f_2(v_1, u_2)$$
$$\vdots$$
$$V_k \leftarrow f_k(v_1, \ldots, v_{k_1}, u_k)$$



space of units $\mathscr{U}$

distribution over units $P(u)$

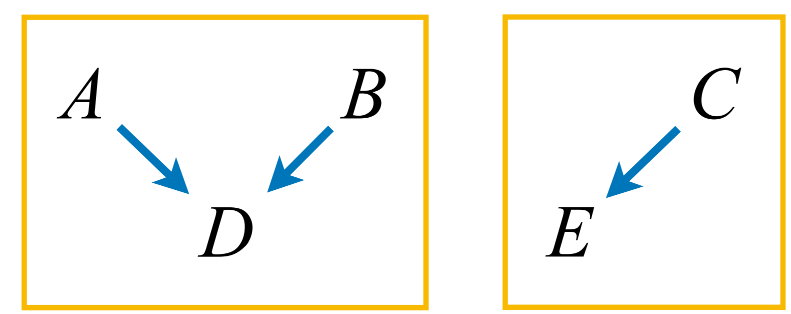
Mechanisms $\mathscr{F}$ + Distribution $P(u)$ = M

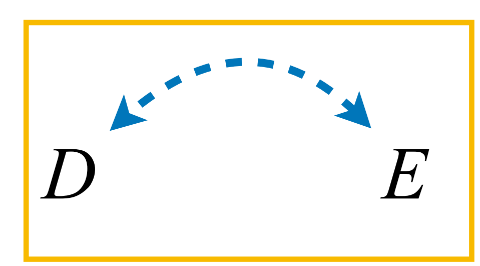
SCM *M* → Causal Diagram *G*

- Every SCM M induces a causal diagram G.
- Represented as a directed acyclic graph (DAG), where:
  - Each $V_i \in \boldsymbol{V}$ is a node,
  - There is an edge $V_i \longrightarrow V_j$ if $V_i \in Pa_j$, and
  - There is a bidirected edge $V_i \longleftrightarrow V_j$ if $U_i \cap U_j \neq \varnothing$.

$\boldsymbol{V} = \{A, B, C, D\}$
$\boldsymbol{U} = \{U\}$

$D \leftarrow f_d\,(A,B,U)$
$E \leftarrow f_e\,(C,U)$





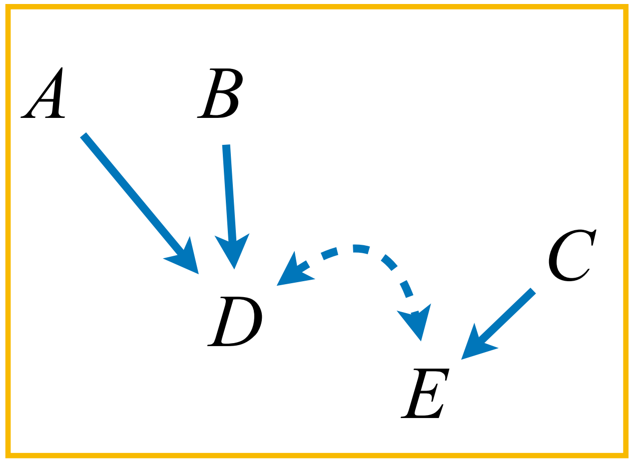
SCM *M* → Causal Diagram *G*

- Every SCM M induces a causal diagram G.
- Represented as a directed acyclic graph (DAG), where:
  - Each $V_i \in \boldsymbol{V}$ is a node,
  - There is an edge $V_i \longrightarrow V_j$ if $V_i \in Pa_j$, and
  - There is a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ if $U_i \cap U_j \neq \varnothing$.

$\boldsymbol{V} = \{A, B, C, D\}$
$\boldsymbol{U} = \{U\}$

$D \leftarrow f_d\,(A,B,U)$
$E \leftarrow f_e\,(C,U)$

G



Counterfactuals' Semantics

- Definition (Potential Response): Let $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$. The potential response of $\boldsymbol{Y}$ to action $do(\boldsymbol{X} = \boldsymbol{x})$, denoted by $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$, is the solution for $\boldsymbol{Y}$ of the set of equations in the model $M_{\boldsymbol{x}}$, where the equations of $\boldsymbol{X}$ are replaced with $\boldsymbol{x}$ (i.e. $Y_x(u) = \boldsymbol{Y}_{M\boldsymbol{x}}(\boldsymbol{u})$).

- Definition (Counterfactual): Let $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$. The counterfactual sentence "the value $\boldsymbol{Y}$ would have obtained, had $\boldsymbol{X}$ been $\boldsymbol{x}$ for unit $\boldsymbol{U}=\boldsymbol{u}$" is interpreted as the potential response $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$.

Observational & Counterfactual Distributions

- For counterfactual quantities, their distribution can be defined via the SCM $\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle$, which induces a family of joint distributions over counterfactual events $Y_x, \ldots, Z_w$ for any $Y, Z, \ldots, X, W \subseteq V$:

$$P^{\mathcal{M}}(y_x, \ldots, z_w) = \sum_u \mathbb{1}\Big(Y_x(u) = y, \ldots, Z_w(u) = z \Big) P(u) \,.$$

- A special case of this, when the subscripts $x, \ldots, z$ are empty, gives the so-called observational distribution. In that case, we simply consider a set of variables $Y \subseteq V$ and the observational distribution is defined by:

$$P^{\mathcal{M}}(y) = \sum_u \mathbb{1}\Big(Y(u) = y \Big) P(u) \,.$$

Fairness Examples & Standard Fairness Model

Example 1 (Berkeley admission). Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts & humanities). For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

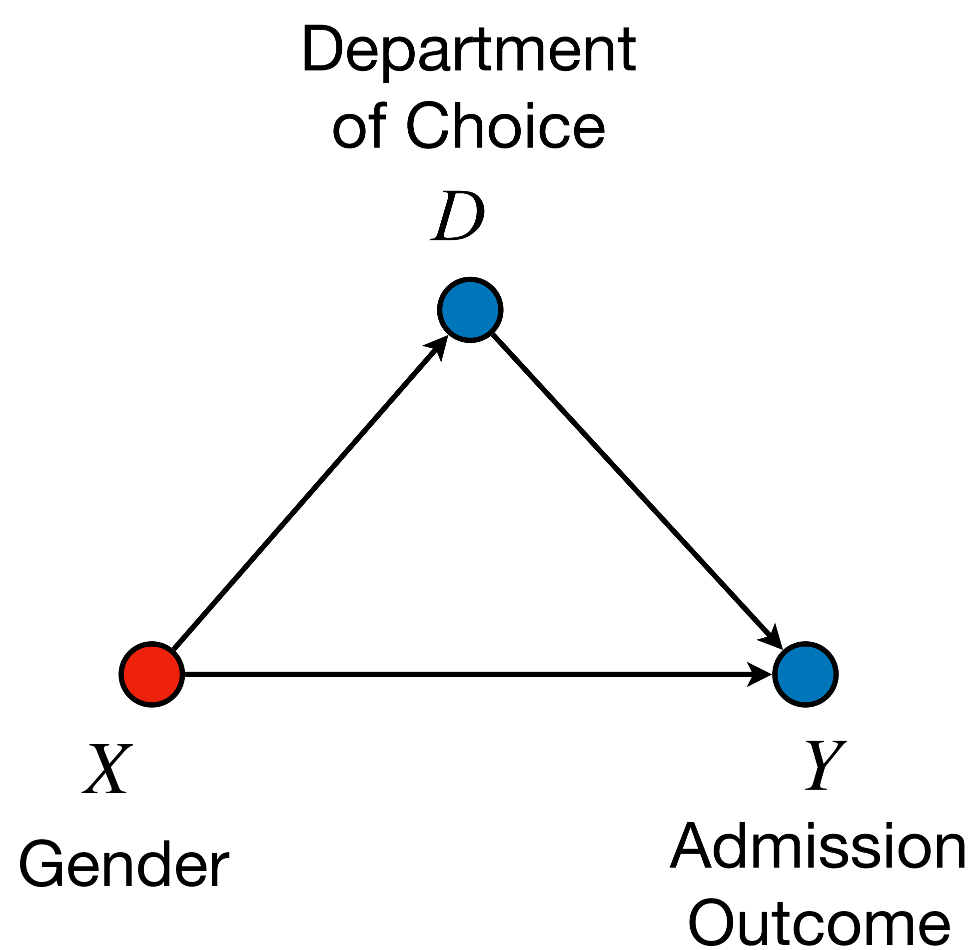
SCM M*

$X \leftarrow$ Bernoulli(0.5)

$D \leftarrow$ Bernoulli$(0.5 + \lambda X)$

$Y \leftarrow$ Bernoulli$(0.1 + \alpha X + \beta D)$

(Truth-Unobserved)



* Bickel, P., Eugene H, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." Science 187.4175 (1975): 398-404.

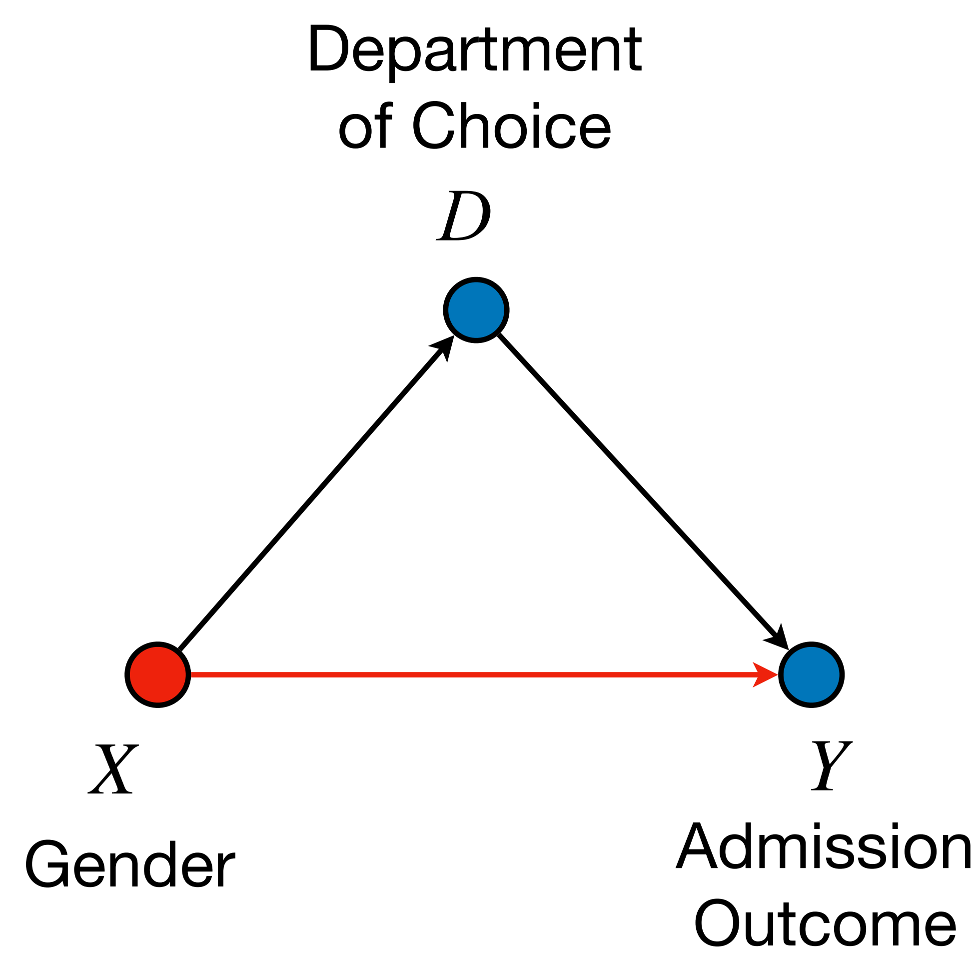
Example 1 (Berkeley admission). Students apply for university admission (Y), and choose specific departments to which they wish to join ($D = 0$ for sciences, $D = 1$ for arts & humanities).
For the purpose of discrimination monitoring, gender is also recorded ($X = 0$ for male, $X = 1$ for female).

- Data analysis reveals that

$$\mathrm{TV}_{x_0,x_1}(Y) = E\left[Y \mid x_1\right] - E\left[Y \mid x_0\right] < 0$$

- A female applicant is predicted to have a lower probability of admission compared to a male applicant.

  Q: *Is this enough to conclude that female students at Berkeley were discriminated during admission?*

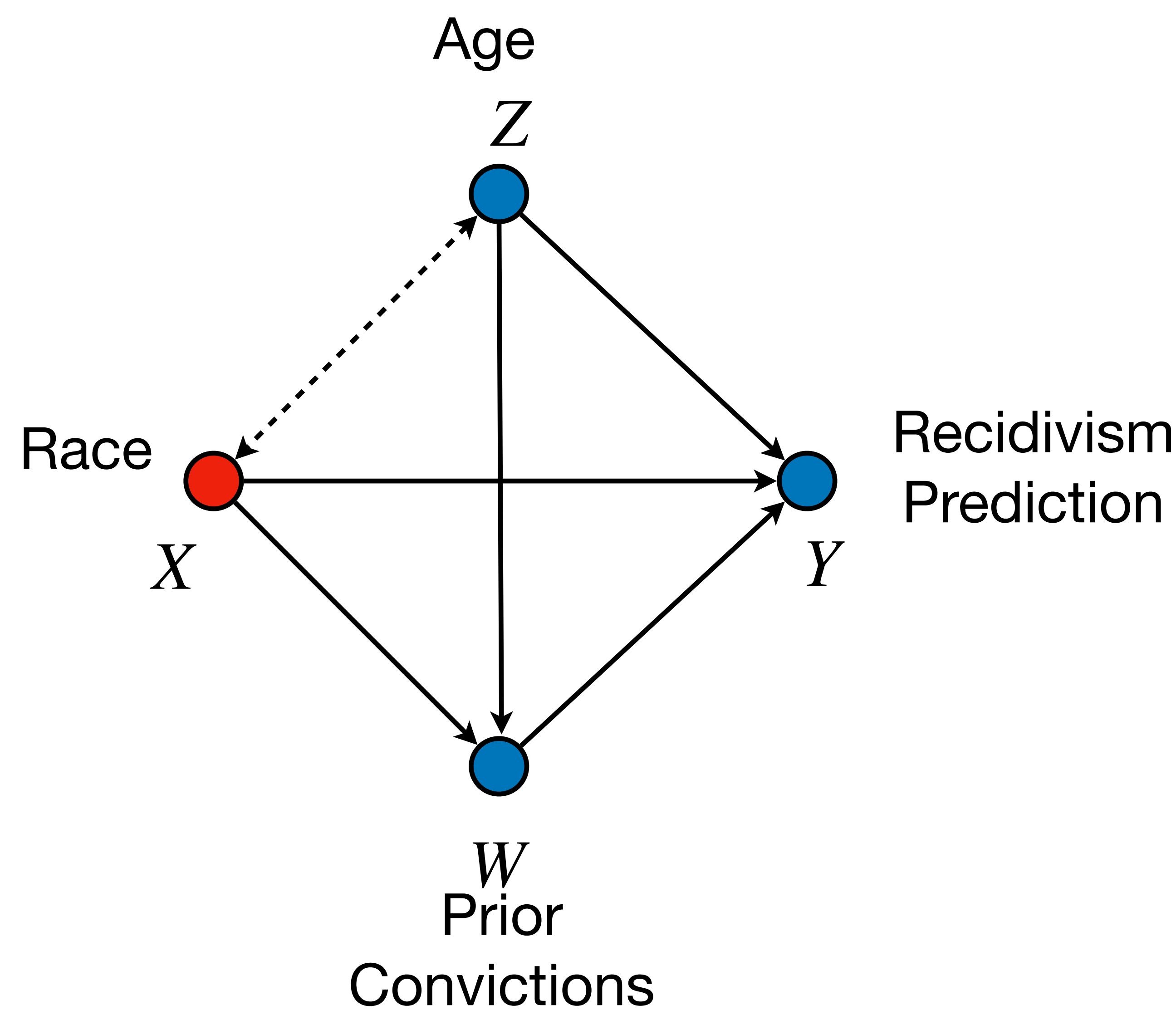


Example 2 (COMPAS prediction). Northpointe are trying to predict whether a person will recidivate after being released (Y). Variable Z represents the age, W represents prior convictions, and X represents race ($X = 0$ for White-Caucasian, $X = 1$ for Non-White).

SCM M*

$X \leftarrow$ Bernoulli$(0.5 + \lambda U)$
$Z \leftarrow \mathcal{N}(40 + \mu U, \sigma^2)$
$W \leftarrow$ Poisson$(0.5 + \alpha X + \beta Z)$
$Y \leftarrow$ Bernoulli$(0.1 + \delta X + \eta W + \phi Z)$

(Truth-Unobserved)



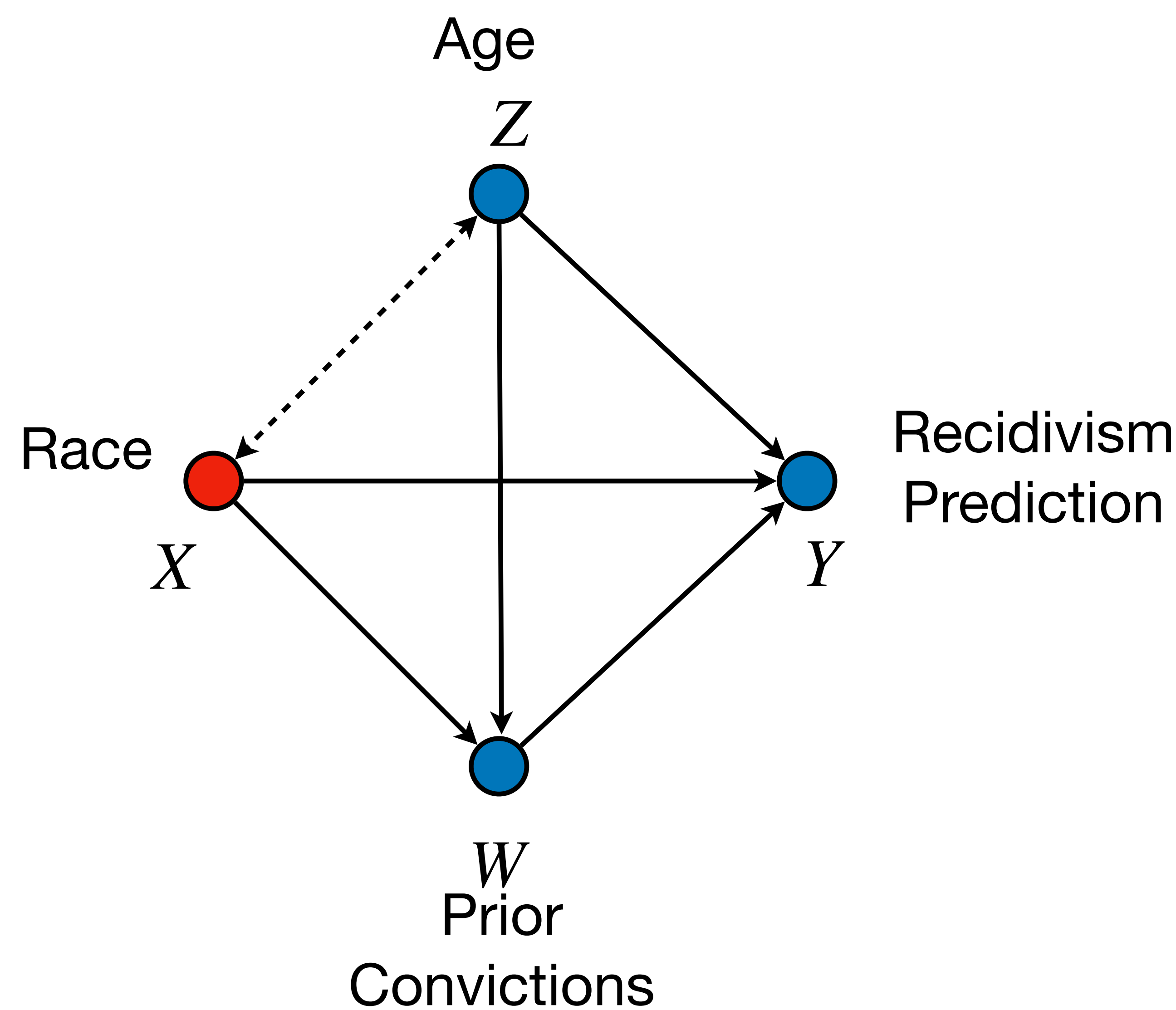
Example 2 (COMPAS prediction). Northpointe are trying to predict whether a person will recidivate after being released (Y). Variable Z represents the age, variable W represents prior convictions, and variable X represents race, ($X = 0$ for White-Caucasian, $X = 1$ for Non-White).

- Data analysis reveals that

$$\mathrm{TV}_{x_0,x_1}(Y) = E\left[Y \mid x_1\right] - E\left[Y \mid x_0\right] > 0$$

- The probability of being classified as high-risk to recidivate is higher in the Non-White group compared to the White-Caucasian group.

- Q: *Can we conclude that Northpointe's software has discriminated against the minority group?*

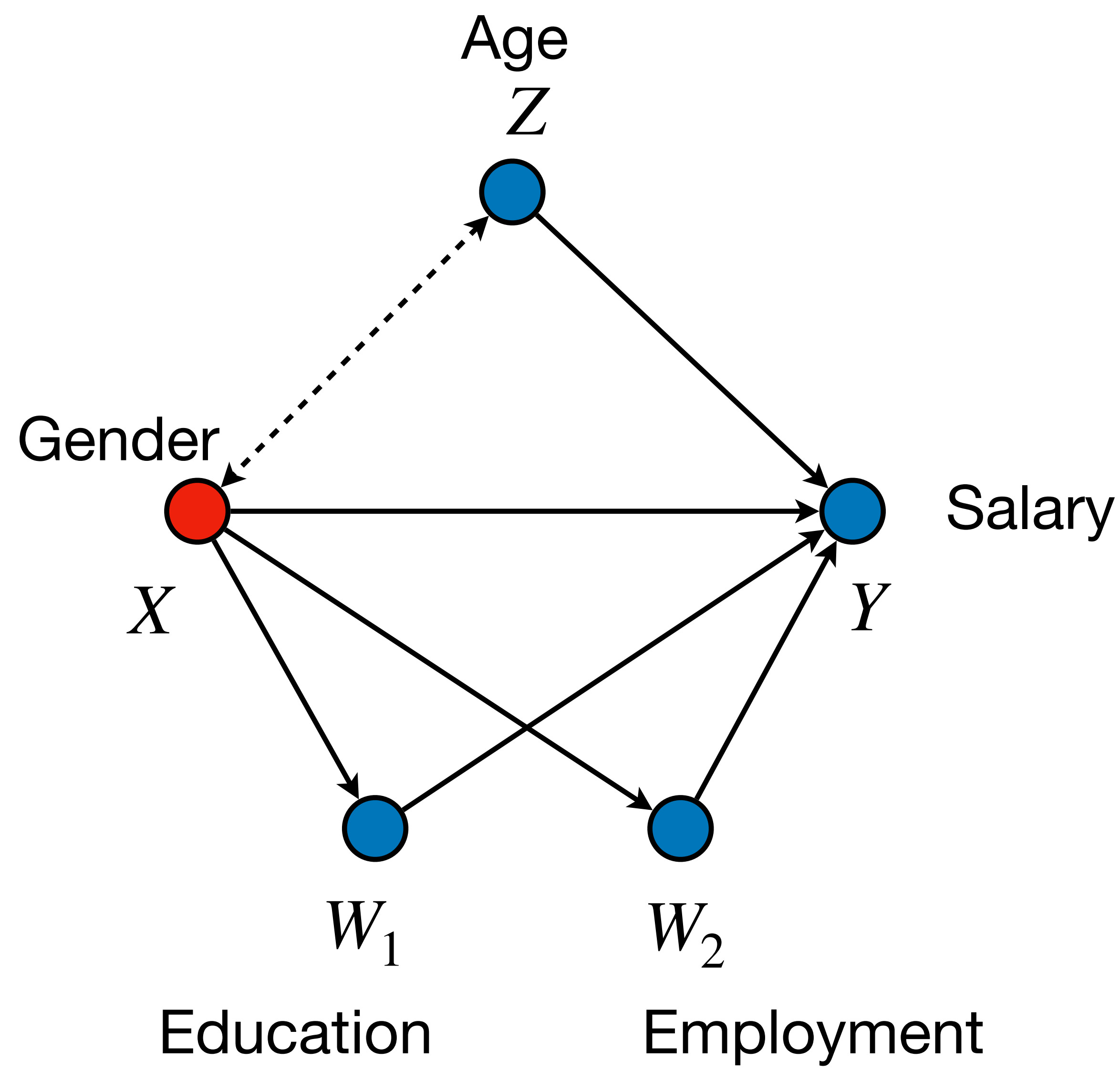


Example 3 (Government Census). The US census data records a person's yearly salary (Y, in tens of thousands of \$). The census also records age (Z), gender ($X = 0$ for male, $X = 1$ for female), education level (W_2) and employment status (W_2).

SCM M*

$X \leftarrow$ Bernoulli$(0.5 + \lambda U)$
$Z \leftarrow \mathcal{N}(40 + \mu U, \sigma^2)$
$W_1 \leftarrow$ Poisson$(0.5 + \alpha_1 X)$
$W_2 \leftarrow$ Binomial$(10{,}0.5 + \alpha_2 X)$
$Y \leftarrow \mathcal{N}(3 + \delta X + \eta W + \phi Z{,}1)$

(Truth-Unobserved)



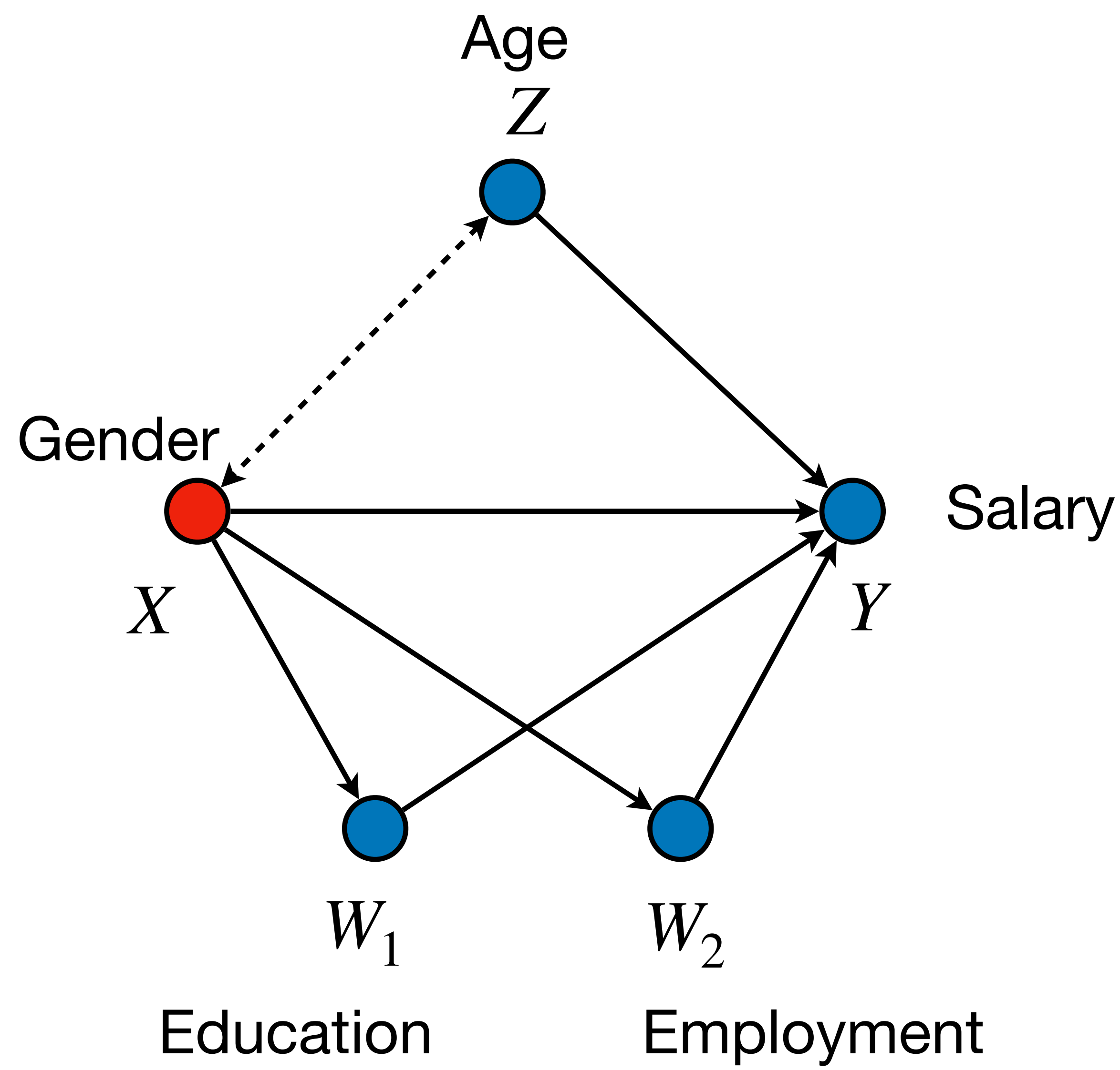
Example 3 (UCI Adult). The US census data records whether a person earns more than \$50,000/year (Y). The census also records age (Z), gender ($X = 0$ for male, $X = 1$ for female), education level (W_1) and employment status (W_2 with 10 job types).

- Data analysis reveals that

$$\text{TV}_{x_0,x_1}(Y) = E\left[Y \mid x_1\right] - E\left[Y \mid x_0\right] < 0$$

- A female employee is predicted to have a lower chance of high income compared to a male employee.

- Q: *Is this enough to conclude that female are systematically discriminated in various companies in the US?*

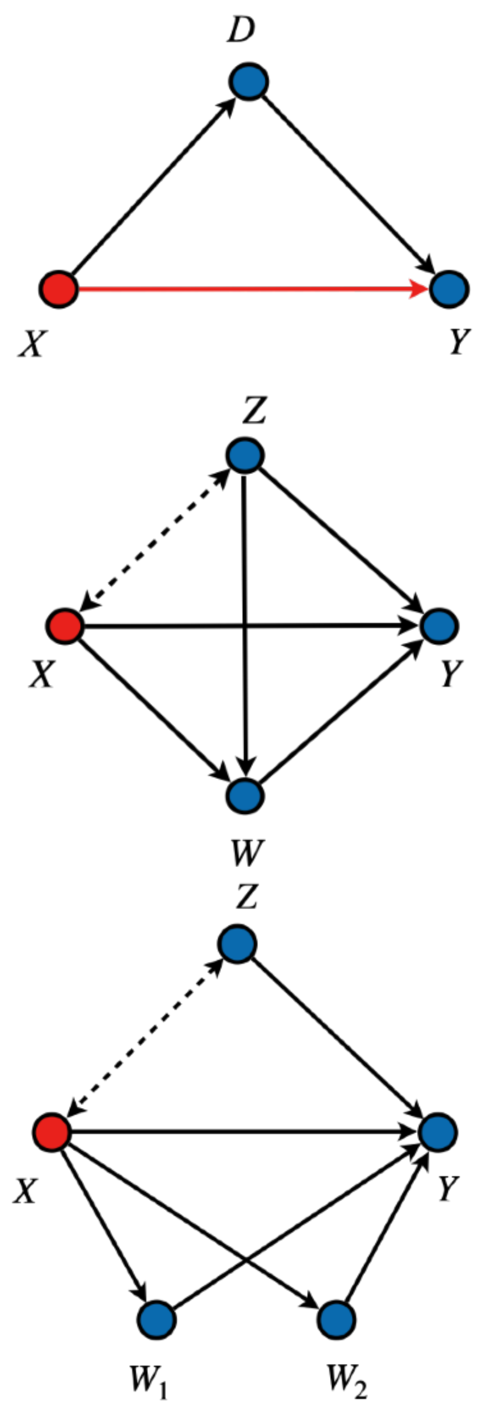


The Emergence of the "Standard Fairness Model"

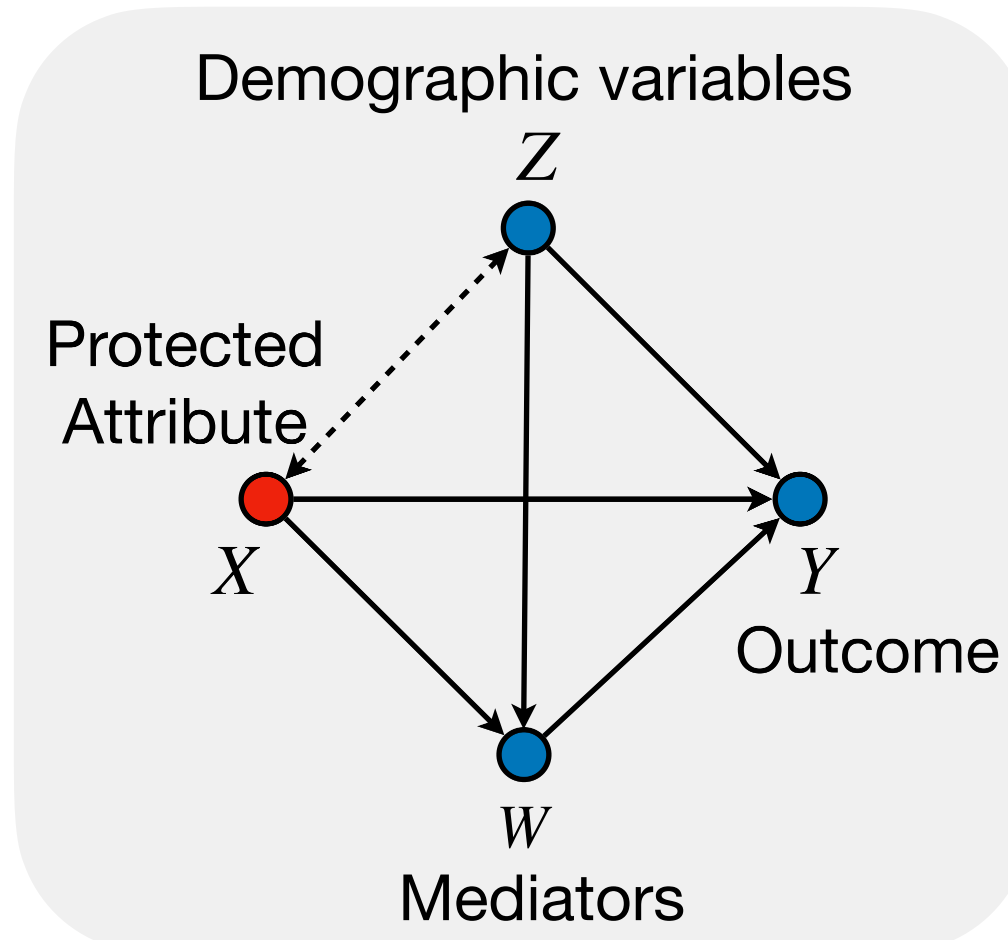
Berkeley

COMPAS

Census



Standard Fairness Model



The Fundamental Problem of Causal Fairness Analysis (FPCFA)

(How to explain observed disparities found in the data in terms of the unobservable causal mechanisms?)

The Fundamental Problem of Causal Fairness Analysis

observed

Female applicants are 14% less likely of being accepted to the university than their male counterparts!

Data $\mathscr{D}$

$TV_{x0,\,x1} = 14\%$

Q: Is the university guilty of gender discrimination?

unobserved

SCM M* (truth):

$X \leftarrow \text{Bernoulli}(0.5)$

$D \leftarrow \text{Bernoulli}(0.5 + 0.2X)$

$Y \leftarrow \text{Bernoulli}(0.1 + 0 * X + 0.3D)$

Active Mechanisms

Direct	Indirect	Spurious
✗	✓	

The Fundamental Problem of Causal Fairness Analysis

observed

Female applicants are 14% less likely of being accepted to the university than their male counterparts!

Data $\mathscr{D}$

$TV_{x_0, x_1} = 14\%$

Q: Is the university guilty of gender discrimination?

No!

unobserved

SCM M* (truth):

$X \leftarrow$ Bernoulli(0.5)
$D \leftarrow$ Bernoulli$(0.5 + 0.2X)$
$Y \leftarrow$ Bernoulli$(0.1 + 0 * X + 0.3D)$

Active Mechanisms

Direct	Indirect	Spurious
✗	✓	—

The Fundamental Problem of Causal Fairness Analysis

observed

Female applicants are 14% less likely of being accepted to the university than their male counterparts!

Data $\mathscr{D}$

$TV_{x0,\,x1}$ = 14%

Q: Is the university guilty of gender discrimination?

No!

M' can generate same data.

unobserved

SCM M* (truth):

$X \leftarrow$ Bernoulli(0.5)

$D \leftarrow$ Bernoulli$(0.5 + 0.2X)$

$Y \leftarrow$ Bernoulli$(0.1 + 0 * X + 0.3D)$

SCM M' (hypothesized):

$X \leftarrow$ Bernoulli(0.5)

$D \leftarrow$ Bernoulli$(0.5 + 0.2X)$

$Y \leftarrow$ Bernoulli$(0.1 + 0.3X + 0 * D)$

Active Mechanisms

Direct	Indirect	Spurious
✗	✓	—

The Fundamental Problem of Causal Fairness Analysis

observed

Female applicants are 14% less likely of being accepted to the university than their male counterparts!

Data $\mathscr{D}$

$TV_{x0, x1} = 14\%$

Q: Is the university guilty of gender discrimination?

No! **Yes!**

unobserved

SCM M* (truth):

$X \leftarrow \text{Bernoulli}(0.5)$

$D \leftarrow \text{Bernoulli}(0.5 + 0.2X)$

$Y \leftarrow \text{Bernoulli}(0.1 + 0 * X + 0.3D)$

Active Mechanisms

Direct	Indirect	Spurious
✗	✓	—

SCM M' (hypothesized):

$X \leftarrow \text{Bernoulli}(0.5)$

$D \leftarrow \text{Bernoulli}(0.5 + 0.2X)$

$Y \leftarrow \text{Bernoulli}(0.1 + 0.3X + 0 * D)$

Active Mechanisms

Direct	Indirect	Spurious
✓	✗	—

The Fundamental Problem of Causal Fairness Analysis

observed

Female applicants are 14% less likely of being accepted to the university than their male counterparts!

Data $\mathscr{D}$

$TV_{x_0, x_1} = 14\%$

Q: Is the university guilty of gender discrimination?

No! **Yes!**

Don't know!

unobserved

SCM M* (truth):

$X \leftarrow \text{Bernoulli}(0.5)$
$D \leftarrow \text{Bernoulli}(0.5 + 0.2X)$
$Y \leftarrow \text{Bernoulli}(0.1 + 0 * X + 0.3D)$

Active Mechanisms

Direct	Indirect	Spurious
✗	✓	—

SCM M' (hypothesized):

$X \leftarrow \text{Bernoulli}(0.5)$
$D \leftarrow \text{Bernoulli}(0.5 + 0.2X)$
$Y \leftarrow \text{Bernoulli}(0.1 + 0.3X + 0 * D)$

Active Mechanisms

Direct	Indirect	Spurious
✓	✗	—

Legal Doctrines: Disparate Treatment & Impact

- The most common legal doctrines found in the US, EU, and throughout the world are known as <u>disparate treatment</u> and <u>disparate impact</u>.
- <u>Disparate treatment</u> is focused on how changes induced by the treatment, or the protected attribute X, affects the outcome Y. In words, how the decision-making criteria changes with X. In CI, this is represented by the notion known as “direct effect.”
- <u>Disparate impact</u> is related to how outcome Y behaves, and trying to understand disparities regardless of the treatment.
    - There are exceptions, & other central notions in legal settings include what is known as “business necessity” (see also “red lining”).
- In general, most of the legal discussions revolve around showing specific causal links, depending on what is permitted or forbidden following society’s standards and expectations.

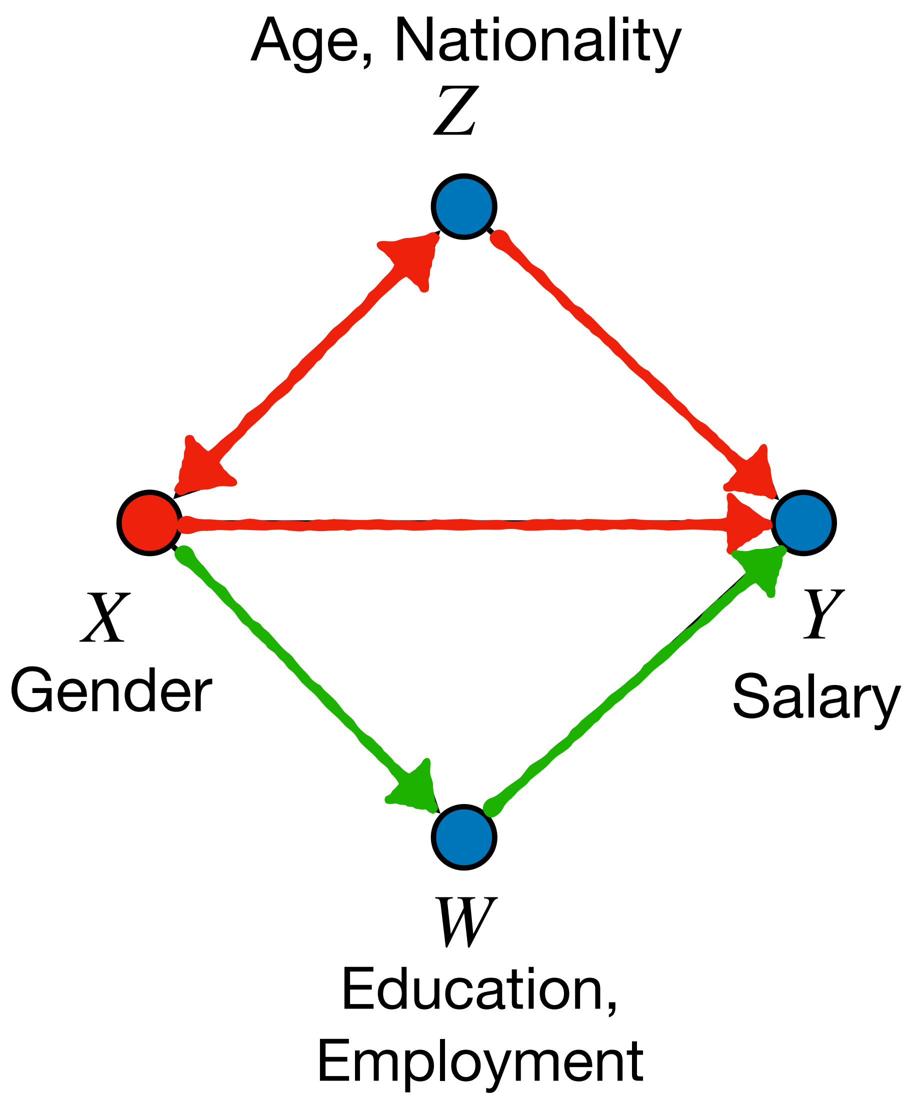
Legal Doctrines of Fairness

Disparate Treatment → Prohibits the use of the protected attribute in the decision process. This is often written as "a similarly situated person who is not a member of the protected class would not have suffered the same fate".

Disparate Impact → Disparate impact occurs when a facially neutral practice has an adverse impact on members of the protected group (the doctrine focuses on outcome fairness). Under this doctrine most commonly fall the cases in which discrimination is unintended or implicit (e.g., redlining).

Business Necessity → Business necessity allows the usage of certain variables that are correlated with the outcome, due to their relevance to the business itself (e.g., PhD degrees in hightech companies).

Example: US Government Census



- The observed disparity in

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$$

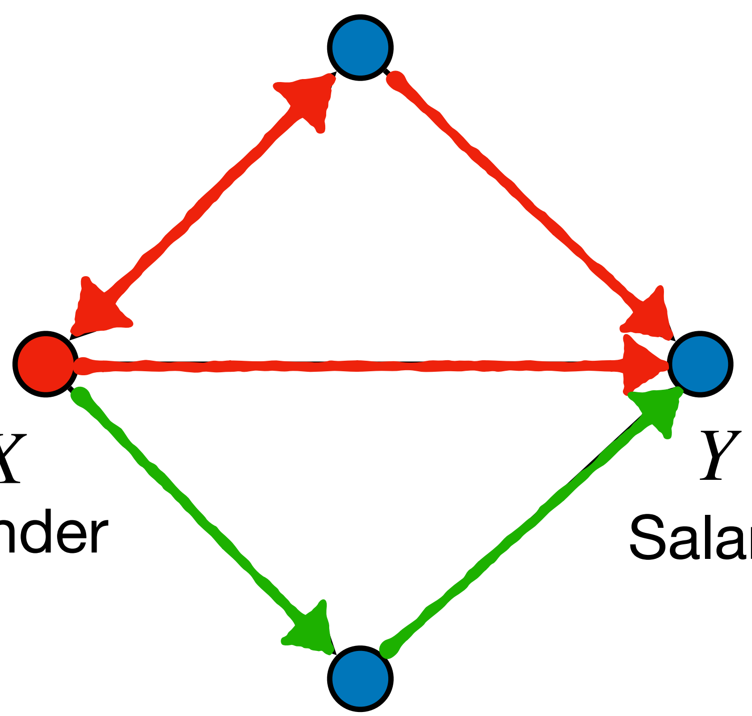
could be explained in different ways, i.e.,

(1) The salary decision is based on employee' gender: $X \to Y$.

(2) Decisions were based on education or employment: $X \to W \to Y$.

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \to Y$.

Example: US Government Census



- The observed disparity in

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$$

could be explained in different ways, i.e.,

(1) The salary decision is based on employee' gender: $X \to Y$.

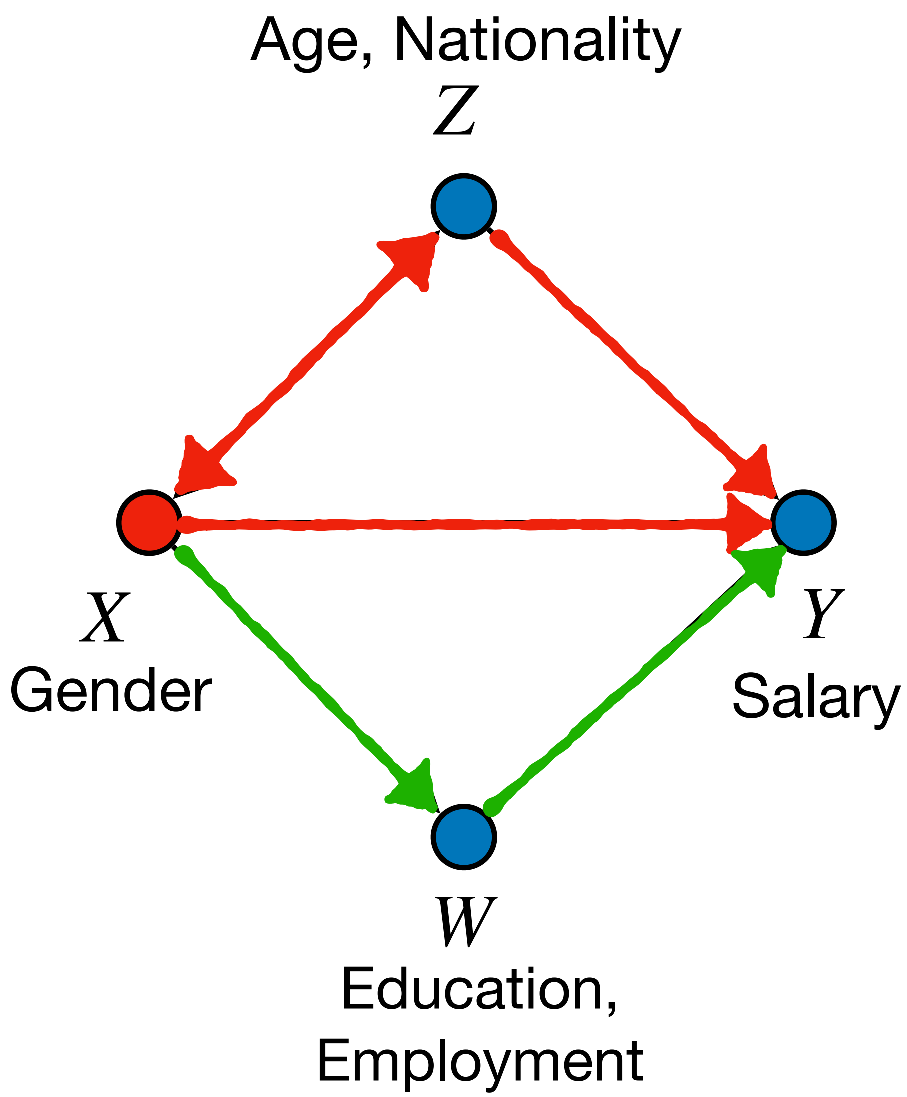
(2) Decisions were based on education or employment: $X \to W \to Y$.

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \to Y$.

(1) suggests a typical case of disparate treatment.

(1+2+3) & the implied TV's disparity suggest a disparate impact case.

Example: US Government Census



- The observed disparity in

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$$

could be explained in different ways, i.e.,

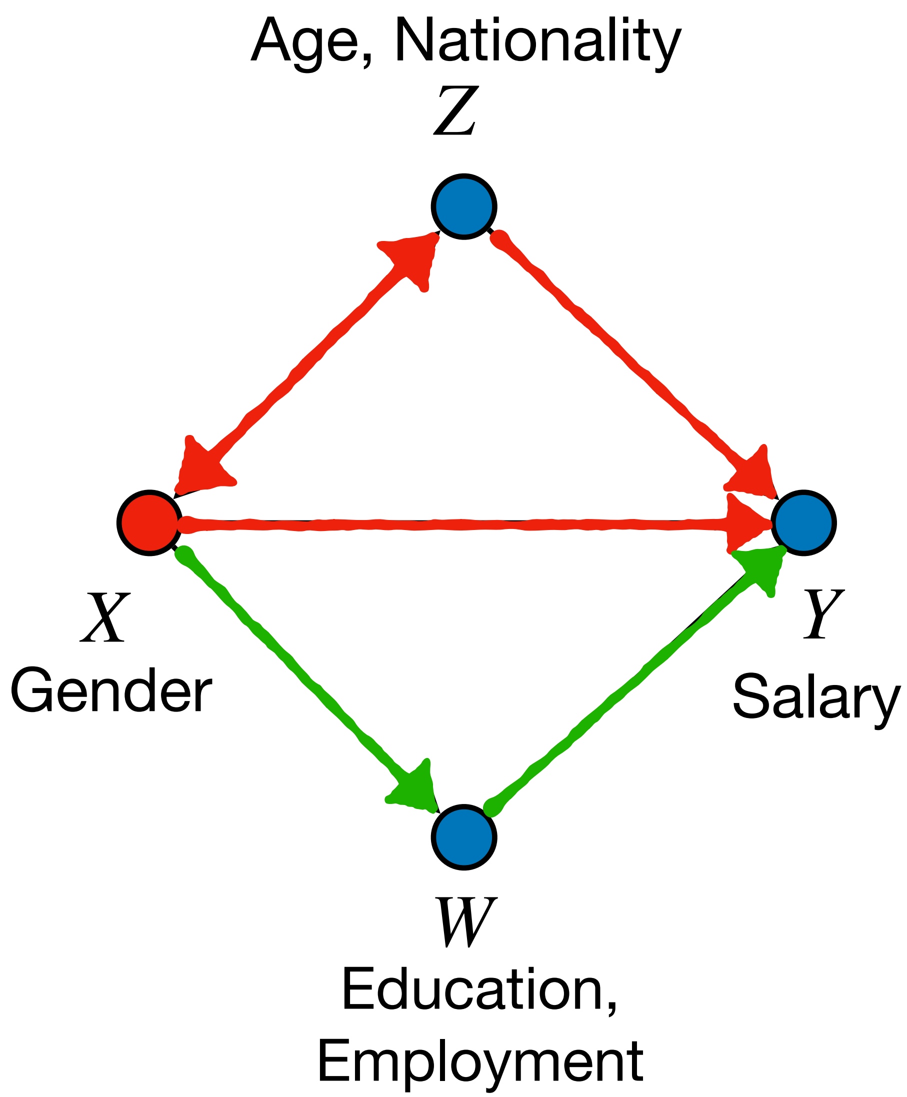
(1) The salary decision is based on employee' gender: $X \rightarrow Y$.

(2) Decisions were based on education or employment: $X \rightarrow W \rightarrow Y$.

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \rightarrow Y$.

After a legal argument, the jury may be okay with *Y*'s variations due to education, but not okay with the variations due to gender or age.

Example: US Government Census



- The observed disparity in

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$$

could be explained in different ways, i.e.,

(1) The salary decision is based on employee' gender: $X \rightarrow Y$.

(2) Decisions were based on education or employment: $X \rightarrow W \rightarrow Y$.

(3) Age or nationality are used to infer the person's gender: $X \leftrightarrow Z \rightarrow Y$.

After a legal argument, the jury may be okay with *Y*'s variations due to education, but not okay with the variations due to gender or age.

$\boxed{\text{How to disentangle these variations within TV?}}$

The Attribution Problem

On the one hand, we consider the observed statistical disparity:

TV = $E[Y \mid \text{male}] - E[Y \mid \text{female}]$

Need a framework/measures that allow for the decomposition of the variations within TV

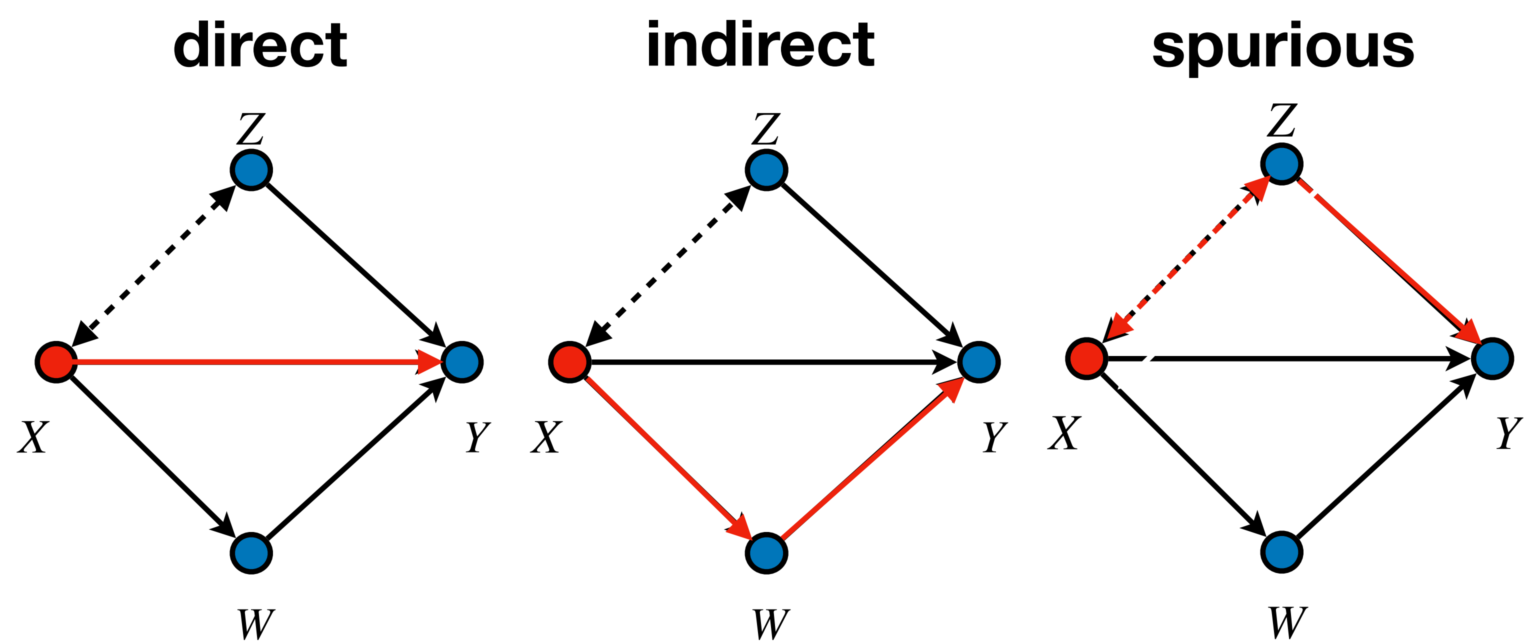
On the other, we need to "ground" (or attribute) the variations to different legal doctrines"

Disparate Treatment

Disparate Impact

Business Necessity

But, we know that TV contains



variations

$\Longrightarrow$ This entanglement makes the attribution problem challenging!

Structural Fairness Measures

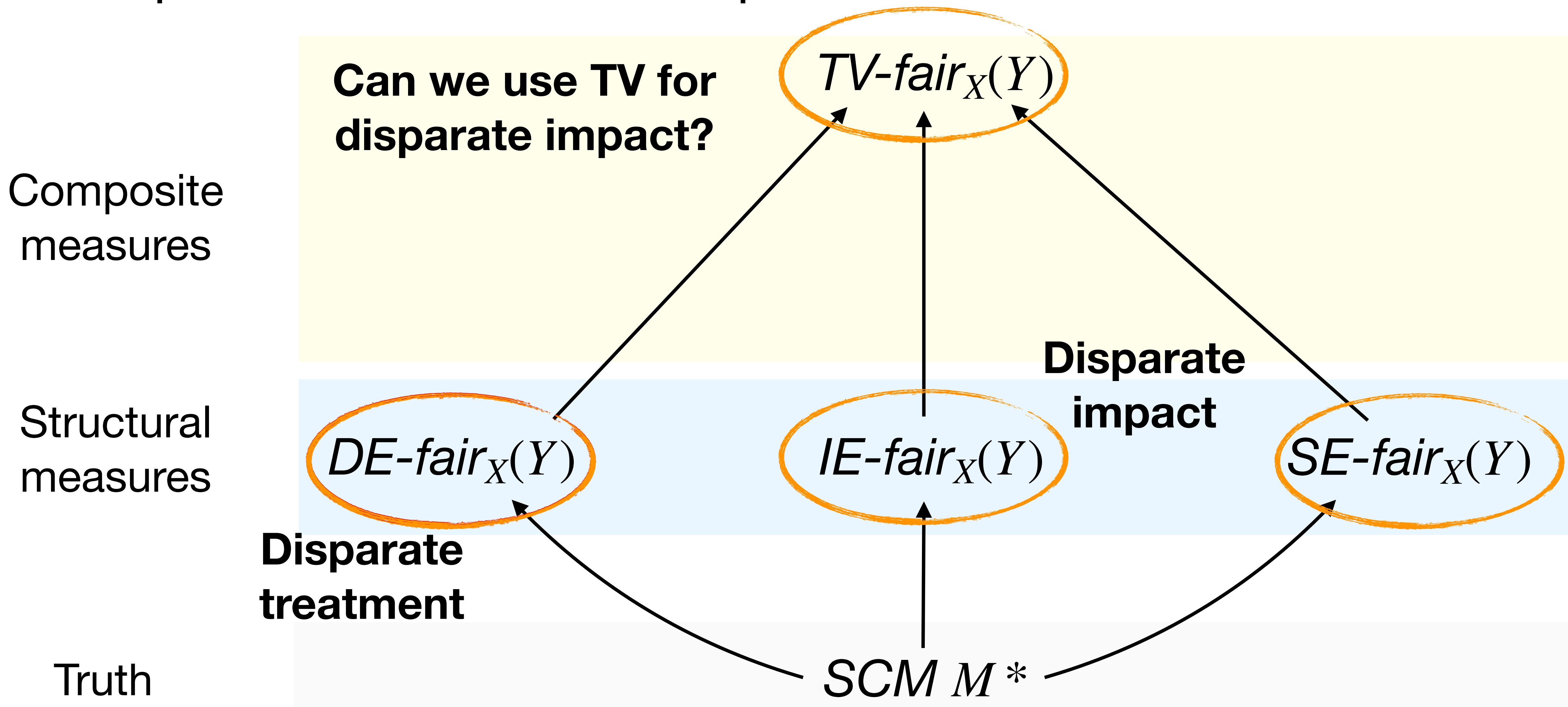
- In order to underpin a more formal discussion amenable to ML, and motivated by the doctrines of disparate treatment & impact, we introduce the **structural fairness measures**. These will represent building blocks of more refined notions.

A. ***Definition.*** *Let pa*(V_i) *and an*(V_i) *be the parents and ancestors of* V_i *in the causal diagram* $\mathcal{G}$*, respectively.*
For an SCM M*,* Y *is fair w.r.t.* X *in terms of:*

1. *the direct effect (DE-fair*$_X(Y)$*, for short) if and only if* $X \notin$ *pa*(Y)*,*
2. *the indirect effect (IE-fair*$_X(Y)$*) if and only if* $X \notin$ *an*$($*pa*$(Y))$*,*
3. *spurious effect (SE-fair*$_X(Y)$*) if and only if* $U_X \cap an_{G_{\underline{X}}}(Y) = \emptyset \wedge an(X) \cap an_{G_{\underline{X}}}(Y) = \emptyset$.

Structural Fairness Measures

- The structural measures represent idealized conditions in which discrimination can be thought about and articulated.
- If we go back to the legal doctrines, we can start connecting disparate treatment and impact with the structural measures.



Admissibility & Power

Definition. *Let Ω be a class of SCMs on which a structural criterion Q and measures μ and μ' are defined.*

- *The measure μ is said to be admissible w.r.t Q if*

$$\forall \mathcal{M} \in \Omega : Q(\mathcal{M}) = 0 \implies \mu(\mathcal{M}) = 0.$$

- *The measure μ' is said to be more powerful than μ if*

(i) μ' *is admissible*

(ii) $\mu'(\mathcal{M}) = 0 \implies \mu(\mathcal{M}) = 0.$

Admissibility & Power

Definition. *Let Ω be a class of SCMs on which a structural criterion Q and measures μ and μ' are defined.*

- *The measure μ is said to be admissible w.r.t Q if*

Note: Power and Admissibility are the analogues of necessity and sufficiency for the corresponding fairness measures.

- *The me*

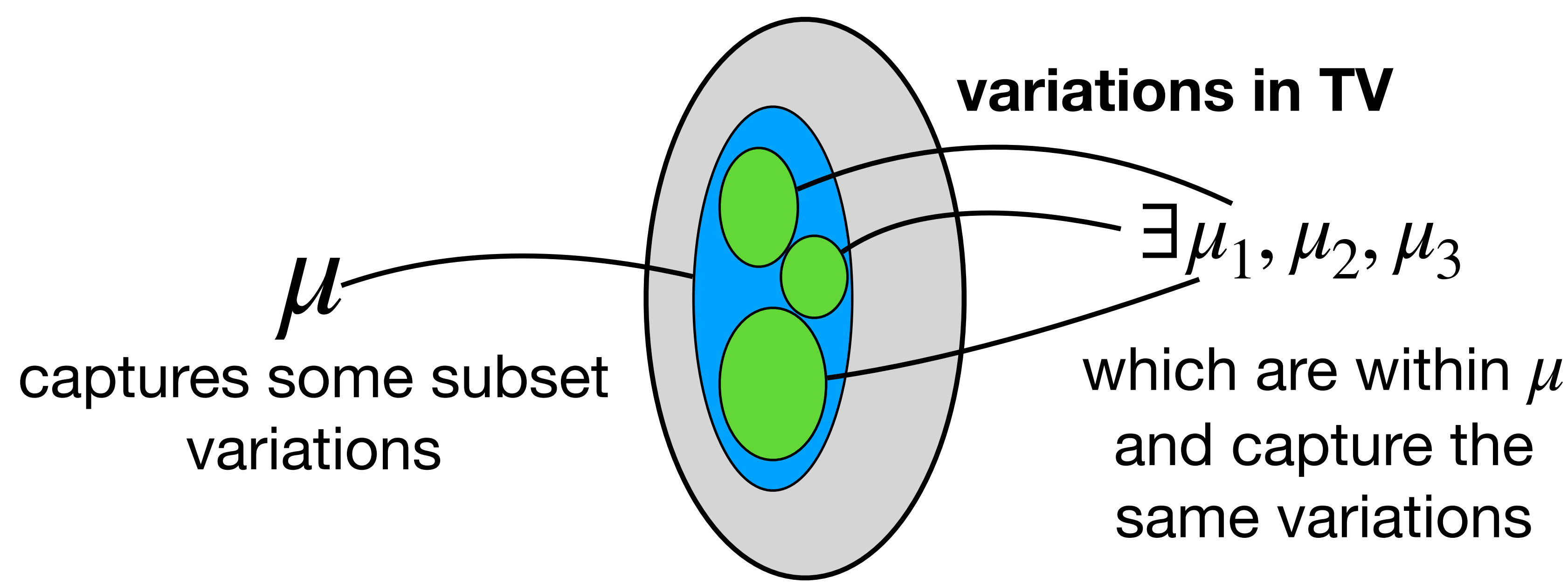
(ii) $\mu'(\mathcal{M}) = 0 \implies \mu(\mathcal{M}) = 0.$

Decomposability

Definition. *Let Ω be a class of SCMs and μ be a measure defined over it. μ is said to be Ω-decomposable if there exist measures*

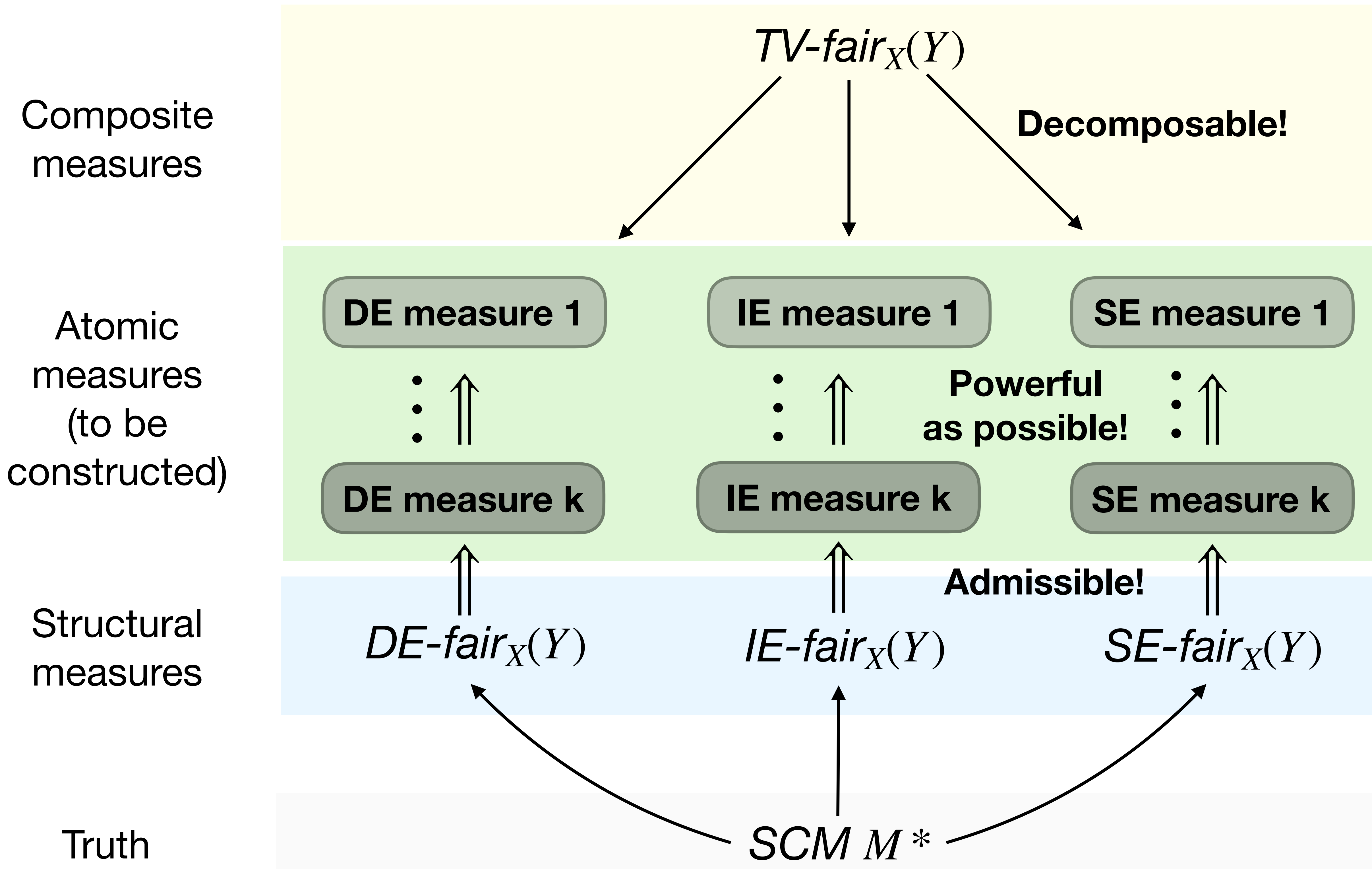
$$\mu_1, \ldots, \mu_k \text{ such that } \mu = f(\mu_1, \ldots, \mu_k),$$

and where f is a non-trivial function vanishing at the origin, i.e., $f(0,\ldots,0) = 0$.



Note: decomposability can imply lack of admissibility

Admissibility, Power, Decomposability - Motivation



Fundamental Problem of Causal Fairness Analysis (FPCFA)

Definition. Let μ be a fairness measure defined over a space of SCMs Ω. Let $Q_1, \ldots, Q_k$ be a collection of structural fairness criteria. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \ldots, \mu_k$ such that the following properties are satisfied:

(i) μ is *decomposable* w.r.t. $\mu_1, \ldots, \mu_k$ **Decomposability**

(ii) $\mu_1, \ldots, \mu_k$ are *admissible* w.r.t. the structural fairness criteria $Q_1, Q_2, \ldots, Q_k$ **Admissibility**

(iii) $\mu_1, \ldots, \mu_k$ are as *powerful* as possible. **Power**

what is our toolkit for solving FPCFA?

Section 3.1
Definition 3.6

The Anatomy of Contrastive Measures

Definition. *A contrast is any quantity of the form*

$$P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)\,.$$

Section 3.2

where E_0, E_1 *are observed (factual) events and* C_0, C_1 *are counterfactual events to which the outcome* Y *responds.*

A contrast compares the outcome Y of individuals

who coincide with the observed event E_1 versus E_0, in the factual world,

and whose values, possibly counterfactually, were intervened on following C_1 versus C_0.

Contrastive Measures: Factual vs. Counterfactual Basis

Theorem. *Any contrast $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$ can be decomposed into its factual and counterfactual components:*

$$\underbrace{\boxed{P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_1)}}_{\textit{counterfactual contrast}} + \underbrace{P(y_{C_0} \mid E_1) - P(y_{C_0} \mid E_0)}_{\textit{factual contrast}} .$$

We normally think of C_0, C_1, E_0, E_1 as including X.

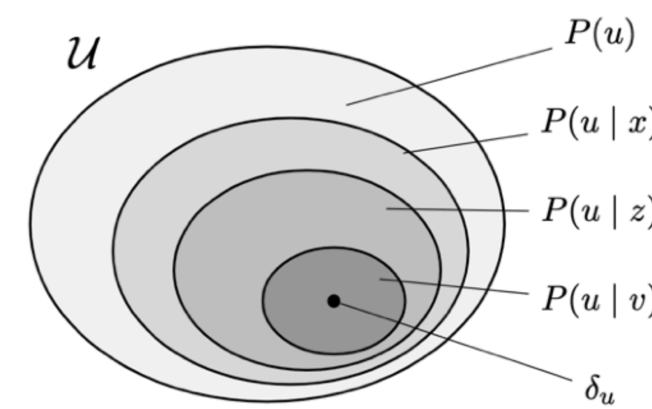
difference arising from counterfactuals C_0, C_1 used to capture the causal influence of X on Y.

difference arising from events E_0, E_1 used to capture non-causal (spurious) influences of X on Y.

Structural Basis Expansion I

Theorem (continued). *Whenever $E_0 = E_1 = e$, any counterfactual contrast* $P(y_{C_1} \mid E = e) - P(y_{C_0} \mid E = e)$ *admits the following structural basis expansion*

$$\sum_u \underbrace{[y_{C_1}(u) - y_{C_0}(u)]}_{\textit{unit-level difference}} \underbrace{P(u \mid E = e)}_{\textit{posterior}}.$$



For a specific unit $U = u$, Y's response to the transition $C_0 \rightarrow C_1$.

Population of units consistent with the factual evidence *E*=e.

Contrastive Measures: Factual vs. Counterfactual Basis

Theorem. *Any contrast $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$ can be decomposed into its factual and counterfactual components:*

$$\underbrace{\boxed{P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_1)}}_{\textit{counterfactual contrast}} + \underbrace{\boxed{P(y_{C_0} \mid E_1) - P(y_{C_0} \mid E_0)}}_{\textit{factual contrast}}.$$

difference arising from counterfactuals C_0, C_1 used to capture causal influences of X on Y.

difference arising from events E_0, E_1 used to capture non-causal influences of X on Y.

Structural Basis Expansion II

Theorem (continued). Whenever $C_0 = C_1 = c$, *any factual contrast* $P(y_c \mid E_1) - P(y_c \mid E_0)$ *admits the following structural basis expansion:*

$$\sum_u \underbrace{y_c(u)}_{\textit{unit outcome}} \underbrace{[P(u \mid E_1) - P(u \mid E_0)]}_{\textit{posterior difference}}.$$

Baseline outcome for a fixed unit $U = u$.

Difference in posteria of how likely unit $U = u$ is selected under events E_0 vs. E_1.

- We will be mostly interested in contrasts w/ $C = x$, so that $X = x$ represents causal pathways.

Theorem (Contrasts & Structural Basis). *Any contrast can be decomposed into its factual and counterfactuals components:*

$$P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0) = P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_1) + P(y_{C_0} \mid E_1) - P(y_{C_0} \mid E_0)\,.$$

mechanisms $\mathscr{F}$

population $P(u)$

Further

A. *Any c ructural basis*

$P(y_C$ $\mid E)$.

unit-level difference *posterior*

Putting it all together…

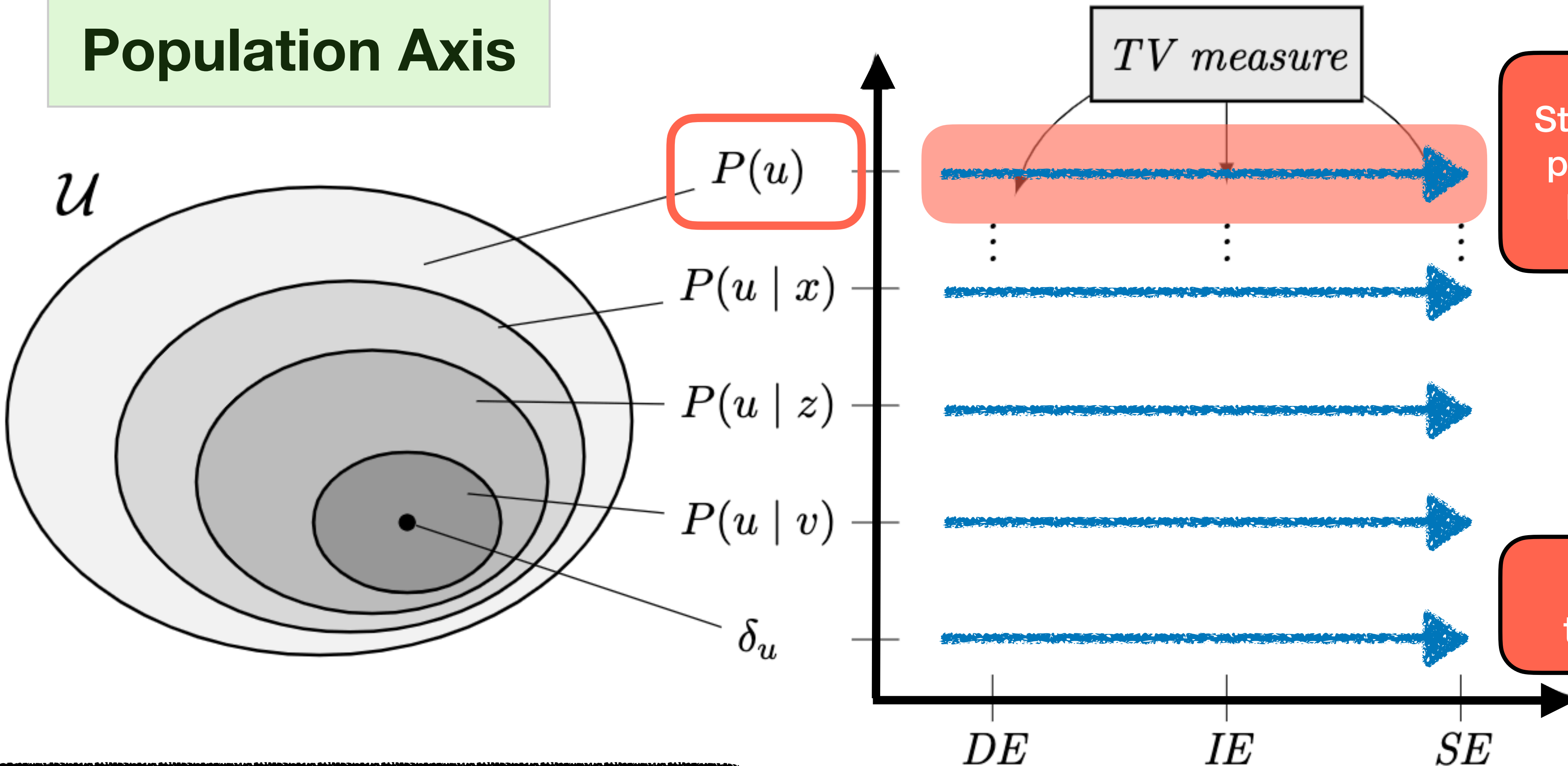
B. *any factual contrast ($C_0 = C_1 = C$) admits the structural basis expansion of the form:*

$$P(y_C \mid E_1) - P(y_C \mid E_0) = \sum_u \underbrace{\boxed{y_C(u)}}_{\text{unit outcome}} \underbrace{\boxed{[P(u \mid E_1) - P(u \mid E_0)]}}_{\text{posterior difference}}\,.$$

Explainability Plane

Section 3.2
Figure 3.2

Population Axis



Start with general population-level DE, IE, and SE

move along
...

And end at the unit level.

Mechanism Axis

Different events $E = e$ imply different units accounted for by the measure.

Different values of interventions $C_0 = c_0, C_1 = c_1$ imply a different mechanism's selection (direct, indirect, spurious).

Explainability Plane