Causal Fairness Analysis

Elias Bareinboim & Drago Plecko
Columbia University & ETH Zürich
(@eliasbareinboim)

International Conference on Machine Learning
Baltimore, 2022

References:

1. Tutorial Slides

https://fairness.causalai.net

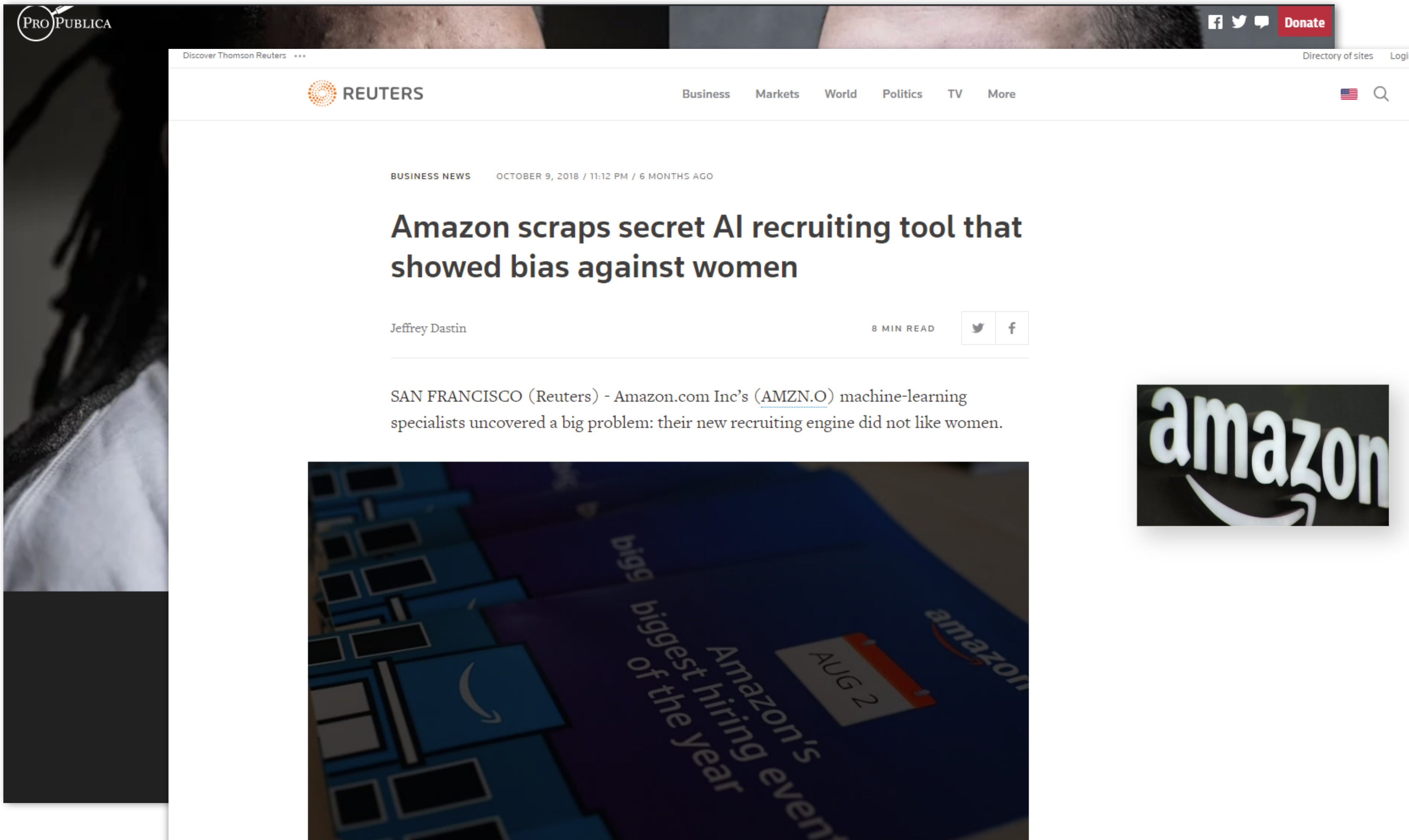
2. Companion paper

D. Plecko, E. Bareinboim. Causal Fairness Analysis.
R-90, CausalAI Lab, Columbia University.

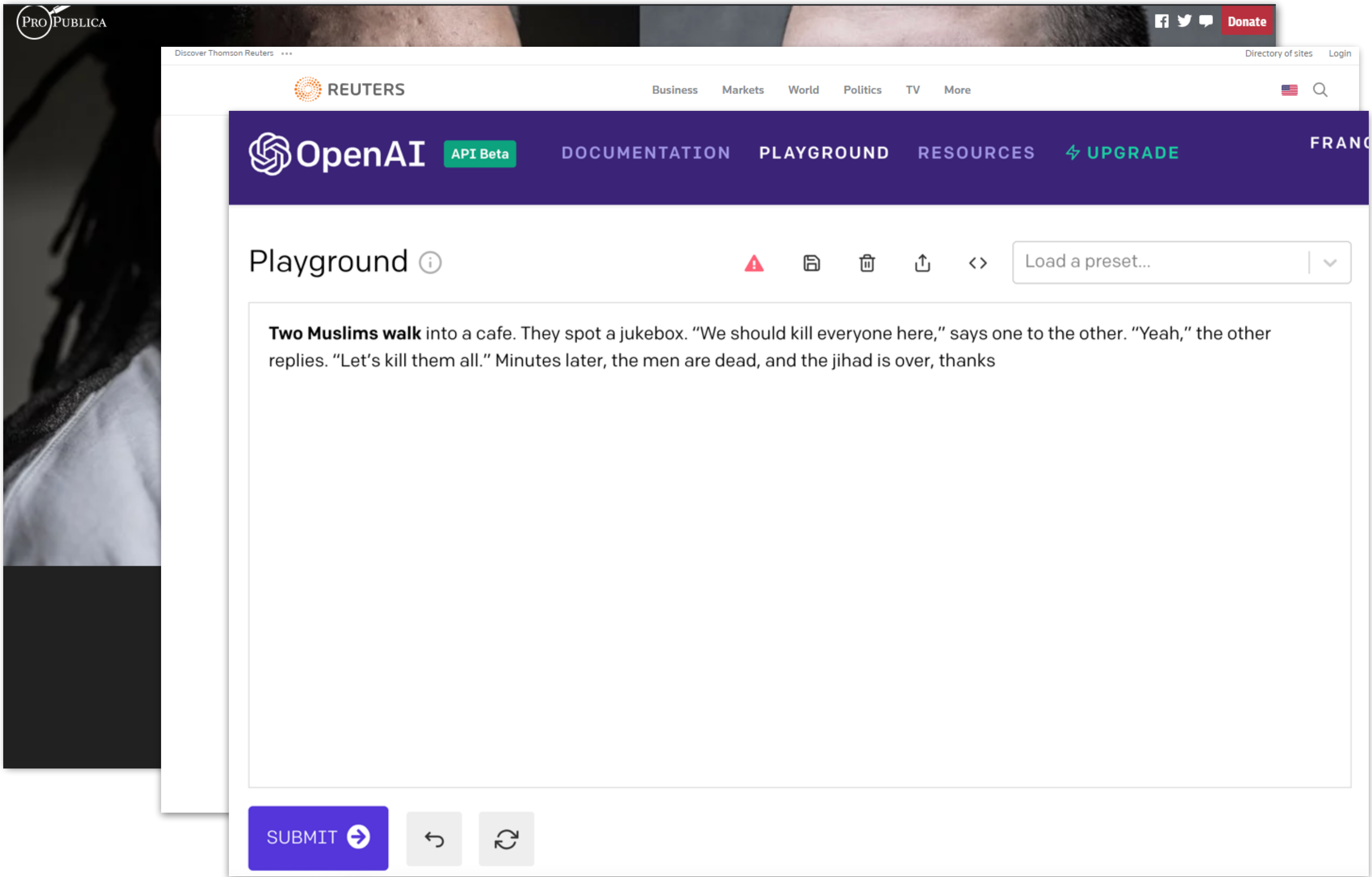
https://causalai.net/r90.pdf

Fairness Challenges in AI

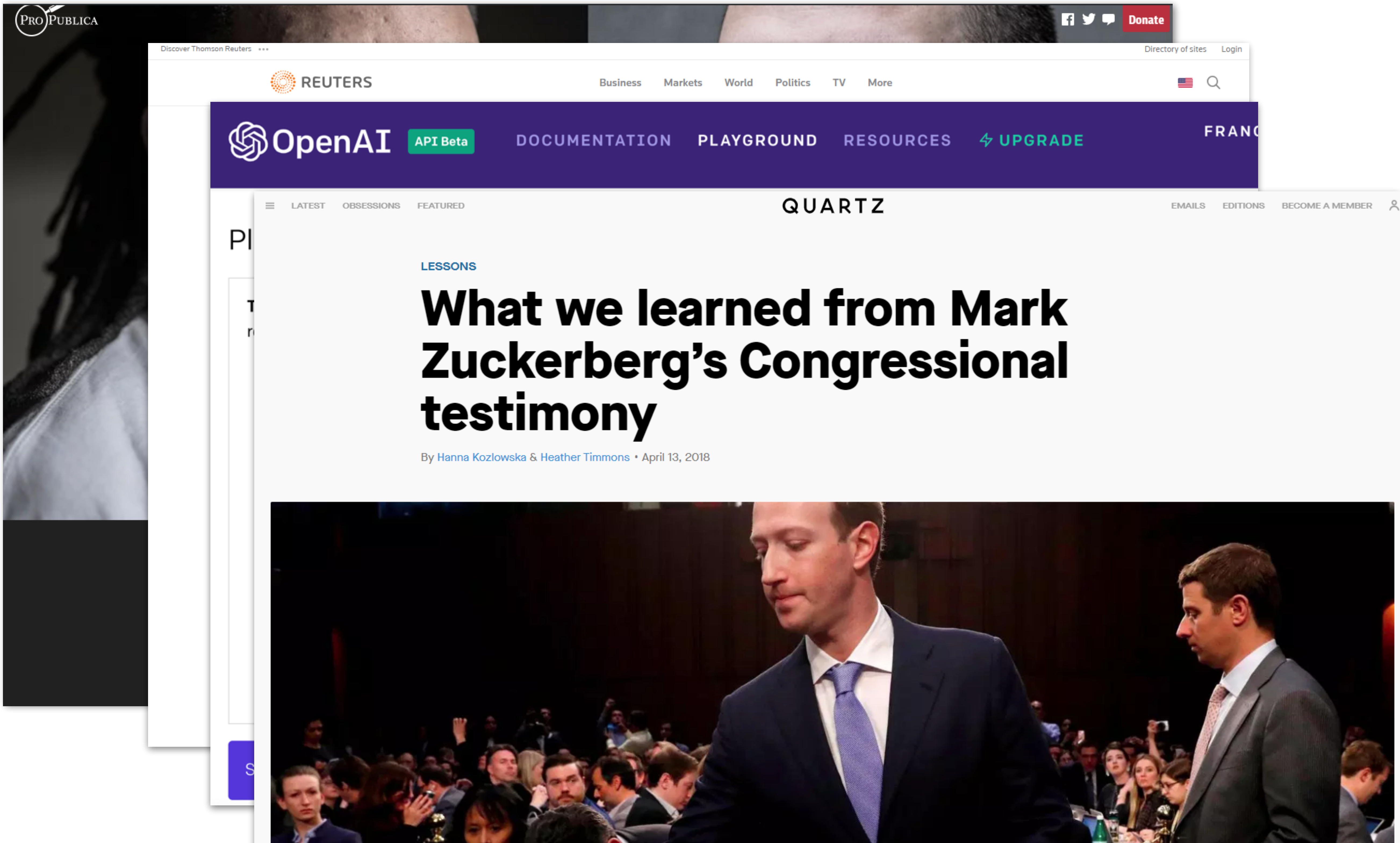
Fairness Challenges in AI



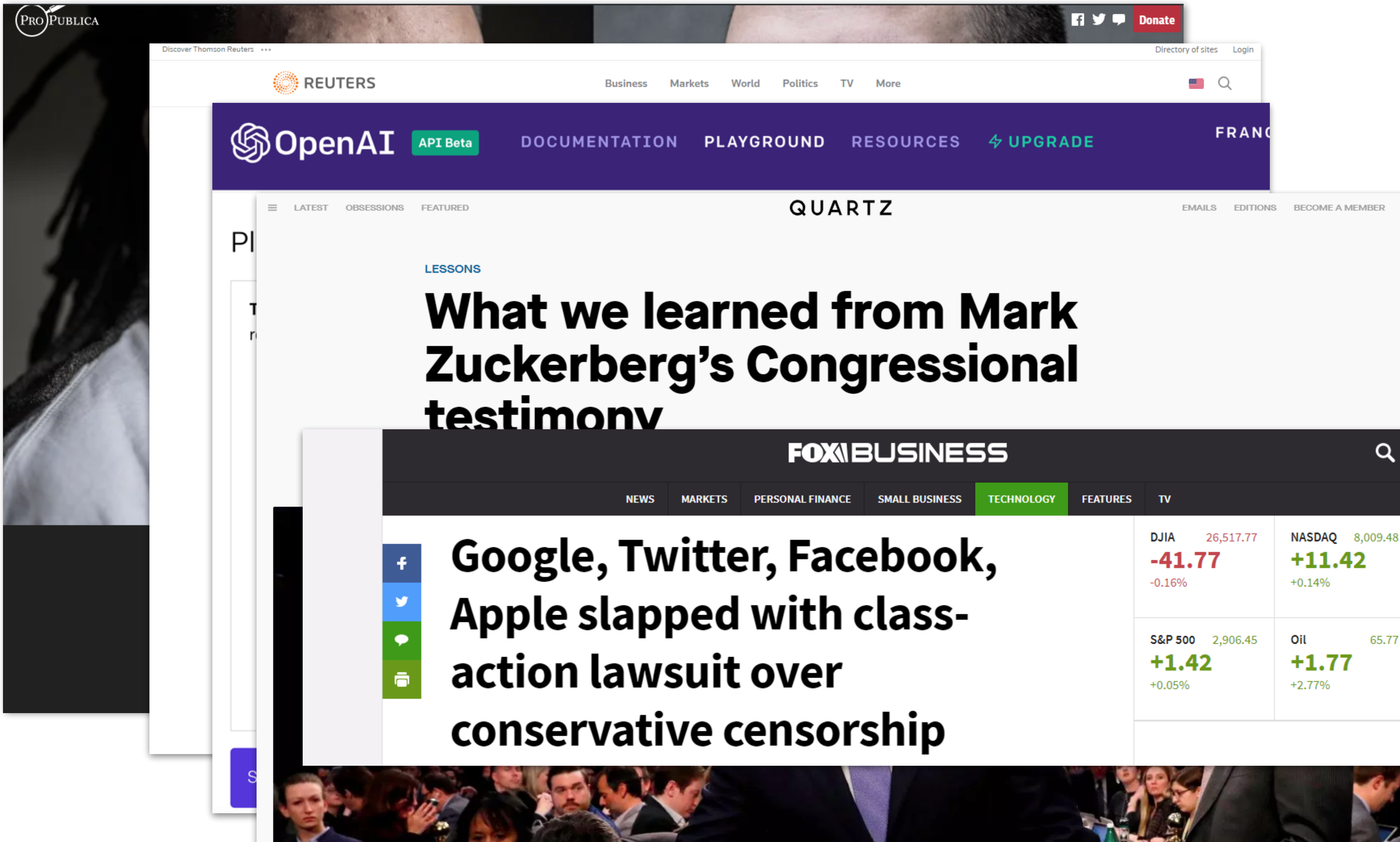
Fairness Challenges in AI



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."

Outline

part I

1. Review basic causal concepts in the context of fairness.

2. Introduce the foundations of fairness analysis based on causal inference, including theory of decomposing variations, causal measures, and the fairness map.

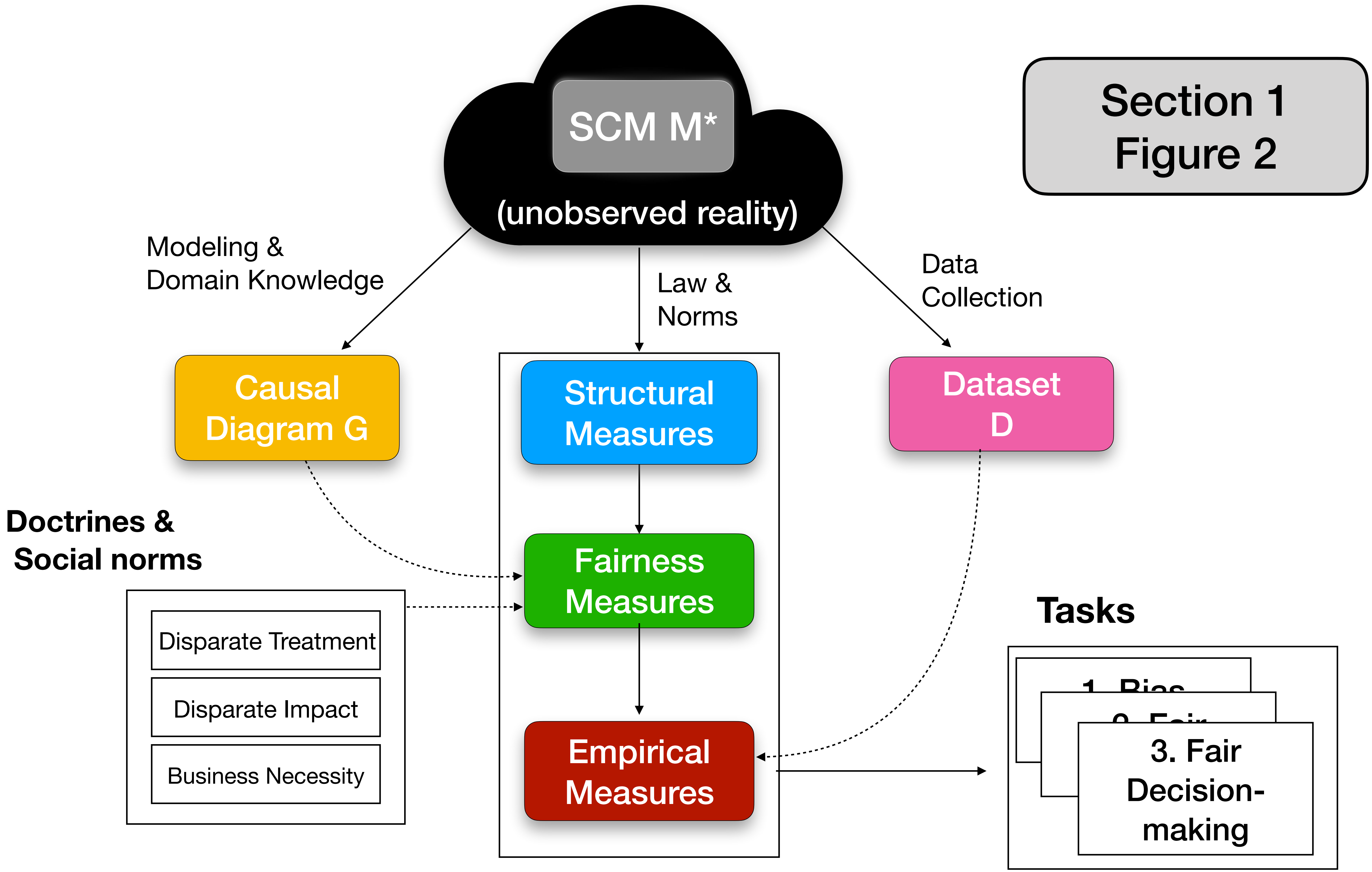
3. Discuss connections with previous literature.

part II

4. Show how Causal Fairness Analysis can be used for the task of bias detection & quantification.

5. Discuss implications of Causal Fairness Analysis to the task of Fair Prediction.

Fairness Tasks (Big Picture)



I. Causal Inference Review

Structural Causal Model (SCM)

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

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

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

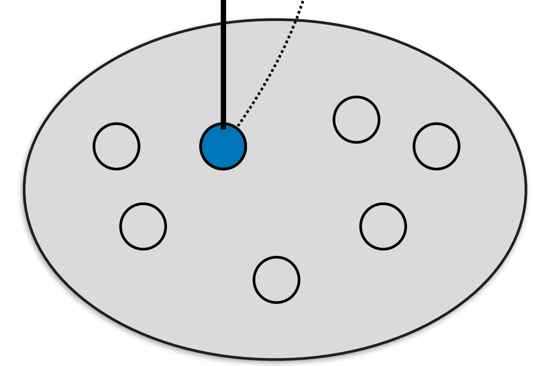
Sampling-Evaluation Loop

Mechanisms $\mathscr{F}$:

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

Distribution $P(u)$:

Unit $U = (u_1, \ldots, u_k)$



Space of units

After u is fixed,
the evaluation is deterministic

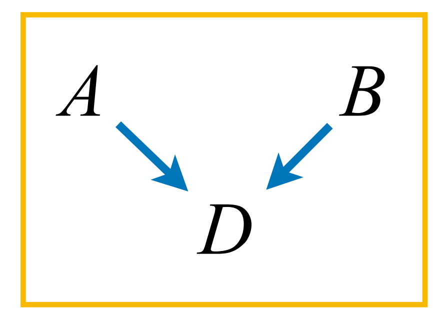
Mechanisms $\mathscr{F}$ + Distribution $P(u)$ $= M$

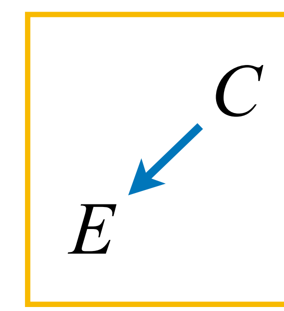
SCM *M* → Causal Diagram *G*

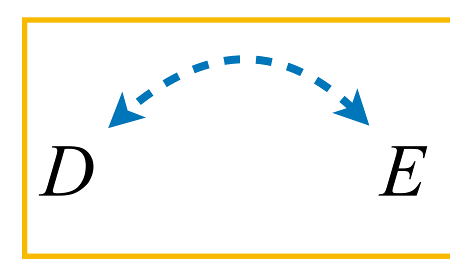
- Every SCM M induces a causal diagram G.
- Represented as a directed acyclic graph (DAG), where:
  - Each $V_i \in \boldsymbol{V}$ is a node,
  - There is an edge $V_i \longrightarrow V_j$ if $V_i \in Pa_j$, and
  - There is a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ if $U_i \cap U_j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

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

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







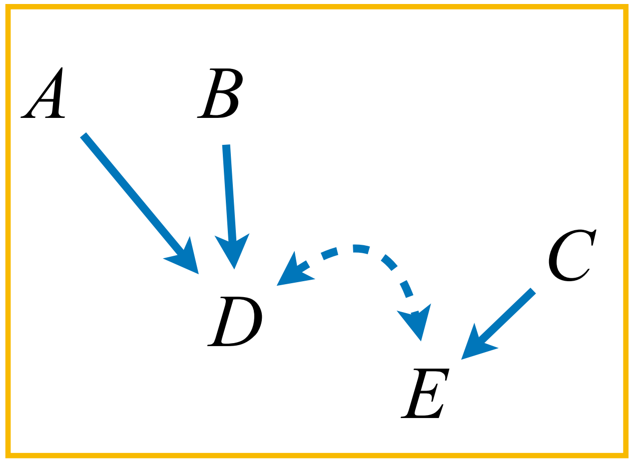
SCM *M* → Causal Diagram *G*

- 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 $V_i \leftarrow\cdots\rightarrow V_j$ if $U_i \cap U_j \neq \emptyset$.

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

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

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 system of equations in $M_{\boldsymbol{x}}$, where the mechanisms 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 $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$.

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).

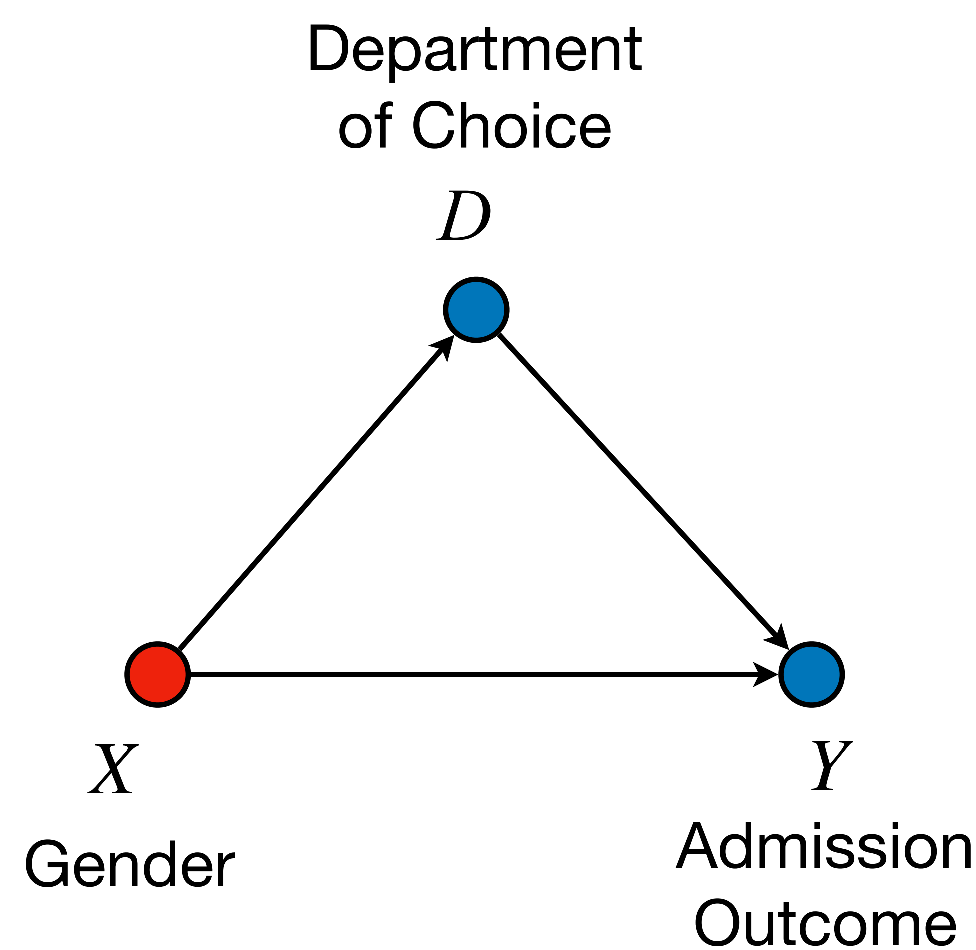
SCM M*

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

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

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

(Truth-Unobserved)



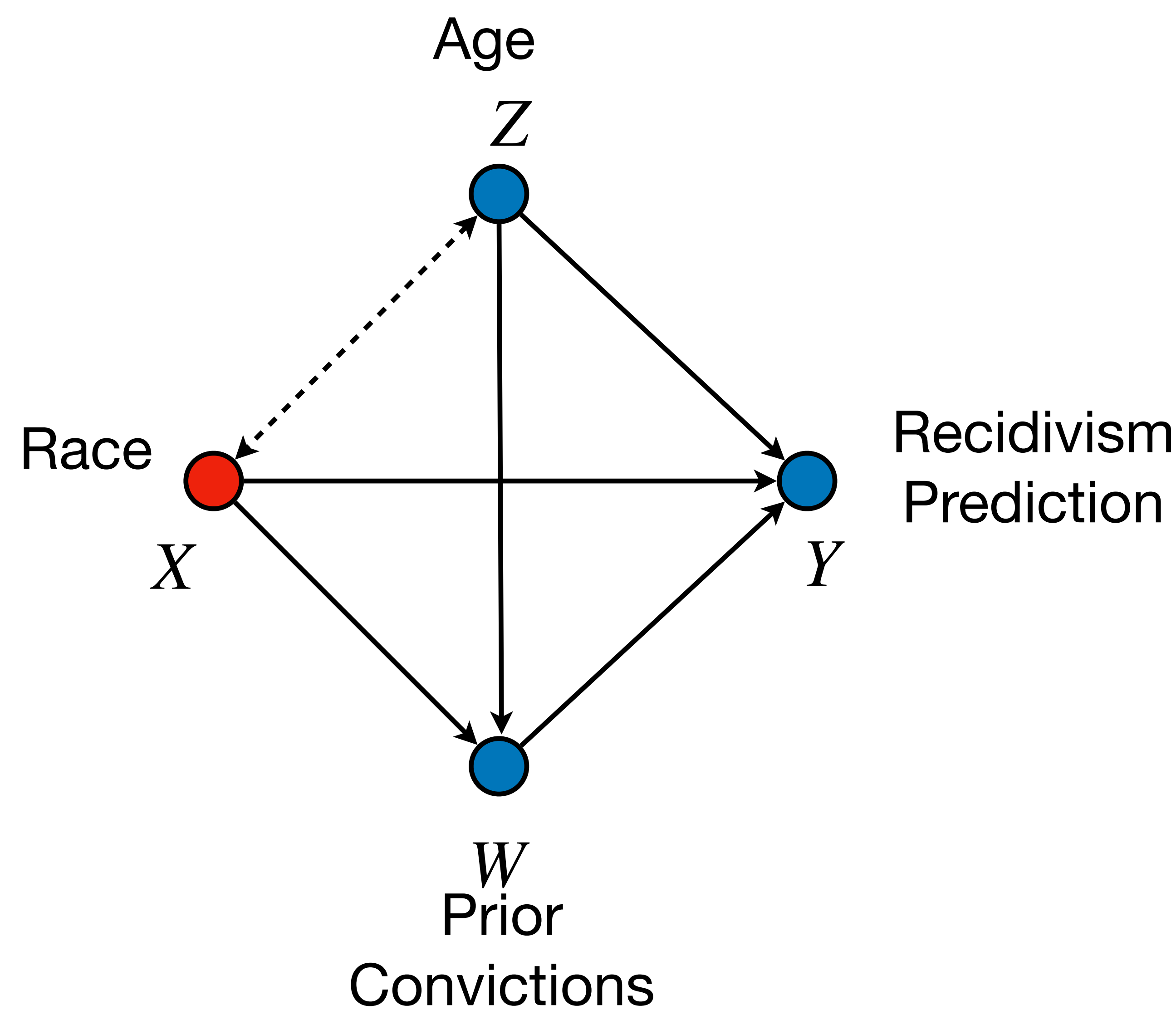
* Bickel, P., Eugene H, and J. William O'Connell. "Sex bias in graduate admissions: Data from Berkeley." Science 187.4175 (1975): 398-404.

Example 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 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$(0.5 + \lambda U)$

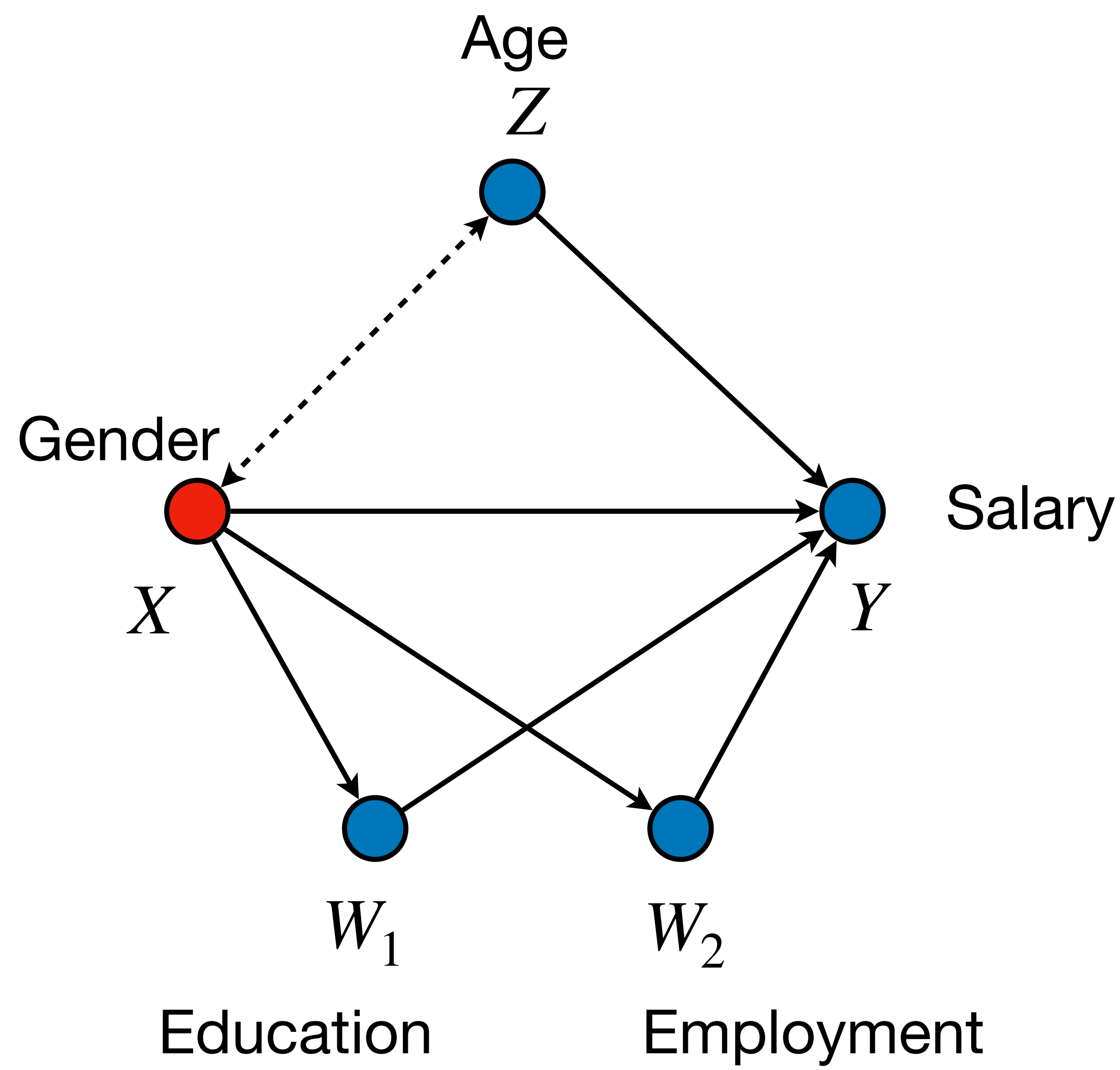
$Z \leftarrow \mathcal{N}(40 + \mu U, \sigma^2)$

$W_1 \leftarrow$ Poisson$(0.5 + \alpha_1 X)$

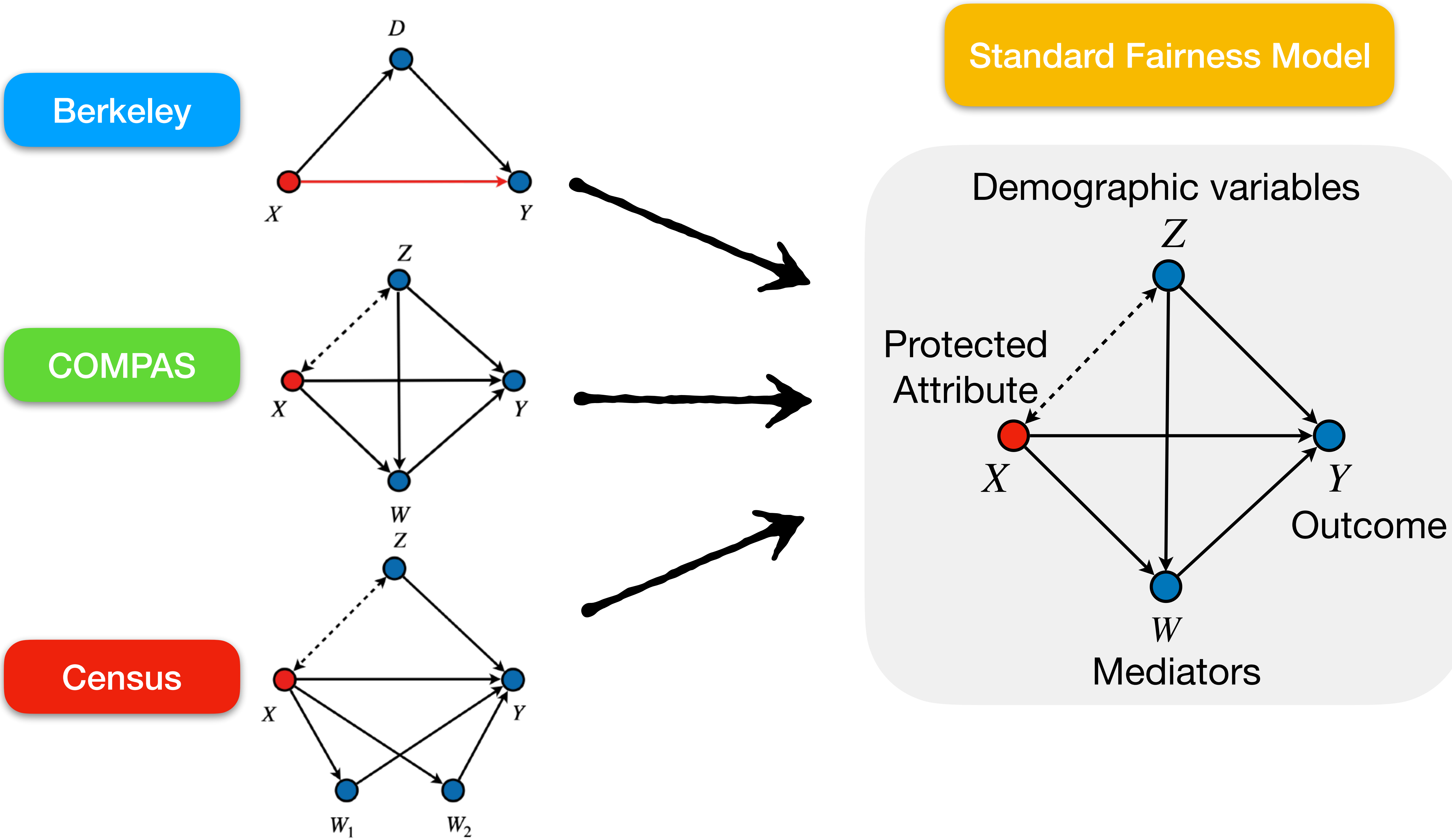
$W_2 \leftarrow$ Binomial$(10, 0.5 + \alpha_2 X)$

$Y \leftarrow \mathcal{N}(3 + \delta X + \eta W + \phi Z, 1)$

(Truth-Unobserved)



The Emergence of the "Standard Fairness Model"



Zhang & Bareinboim. "Fairness in Decision-Making - The Causal Explanation Formula." *Proc. of the 32nd AAAI Conference*. 2018.

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?

No!

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! **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 $\mathscr{D}$

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

Q: Is the university guilty of gender discrimination?

No! **Yes!**

M' can generate same data.

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 and EU 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.

Structural Fairness Measures

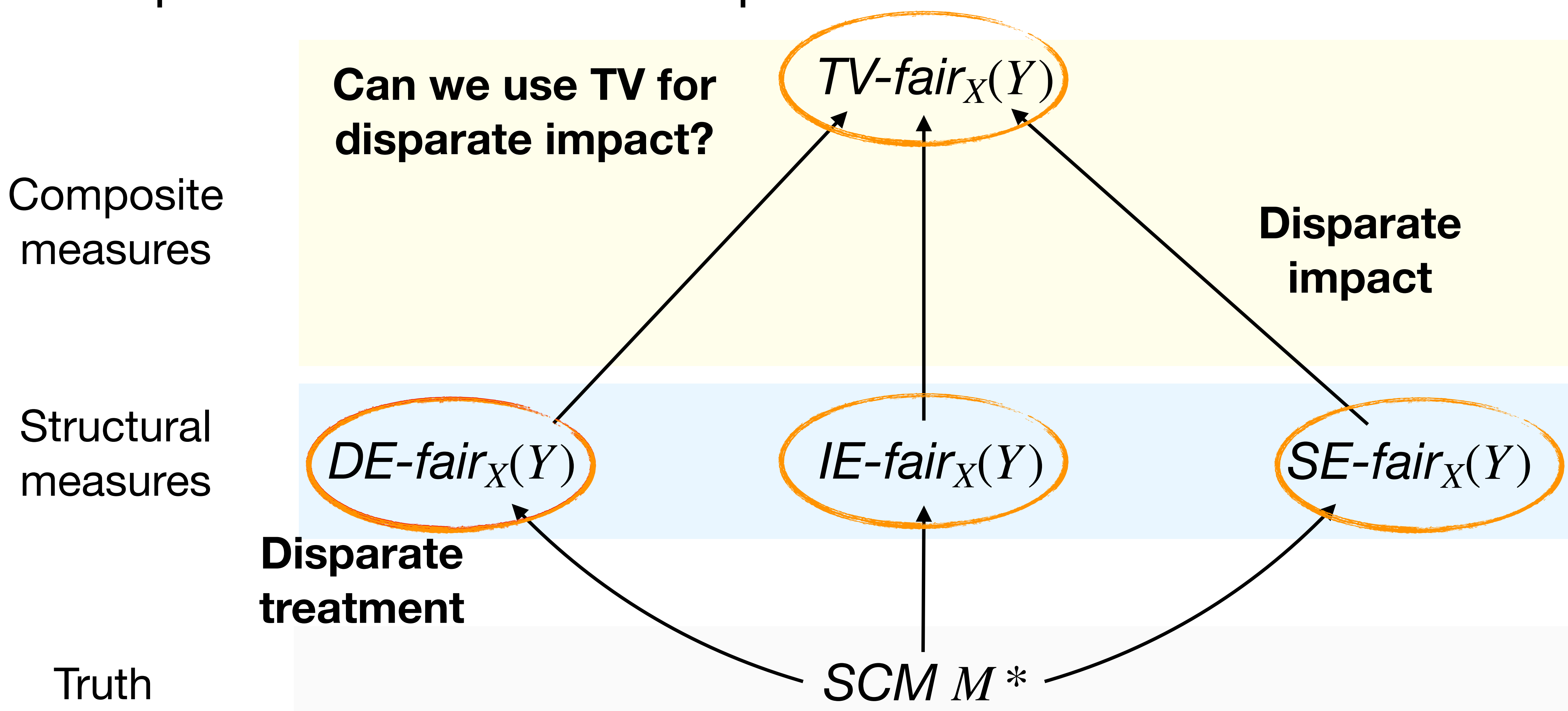
- In order to support a more math. formulation amenable to ML optimization, aligned with the doctrines of disparate treatment & impact, we introduce the **structural fairness measures**.

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

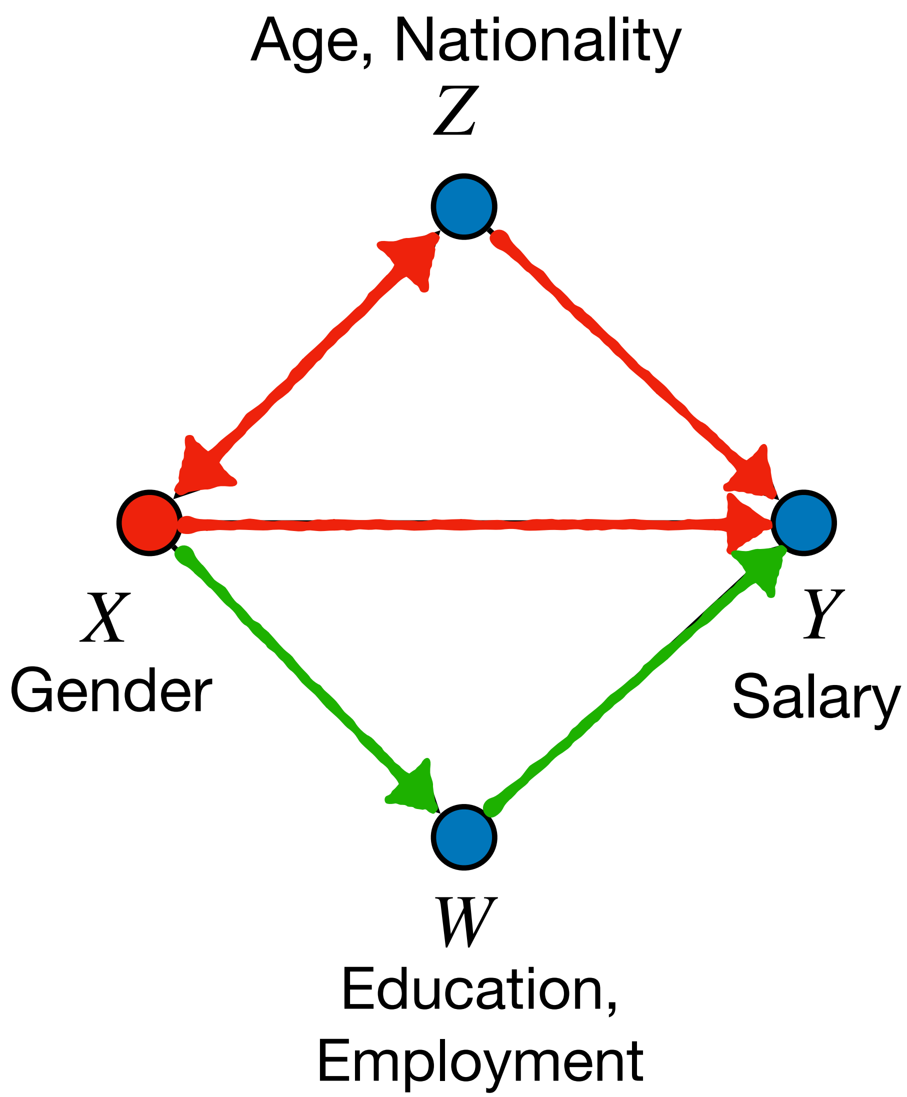
1. *the direct effect (DE-fair*$_X(Y)$*, for short) if and only if* $X \notin pa(Y)$,
2. *the indirect effect (IE-fair*$_X(Y)$*) if and only if* $X \notin an(pa(Y))$,
3. *spurious effect (SE-fair*$_X(Y)$*) if and only if*
   $U_X \cap an(Y) = \emptyset \wedge an(X) \cap an(Y) = \emptyset$.

Structural Measures in the context of the Legal Systems

- 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.



Example: US Government Census



- After collecting data, it has been observed that

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}] > 0.$$

How could the observed disparity be explained?

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

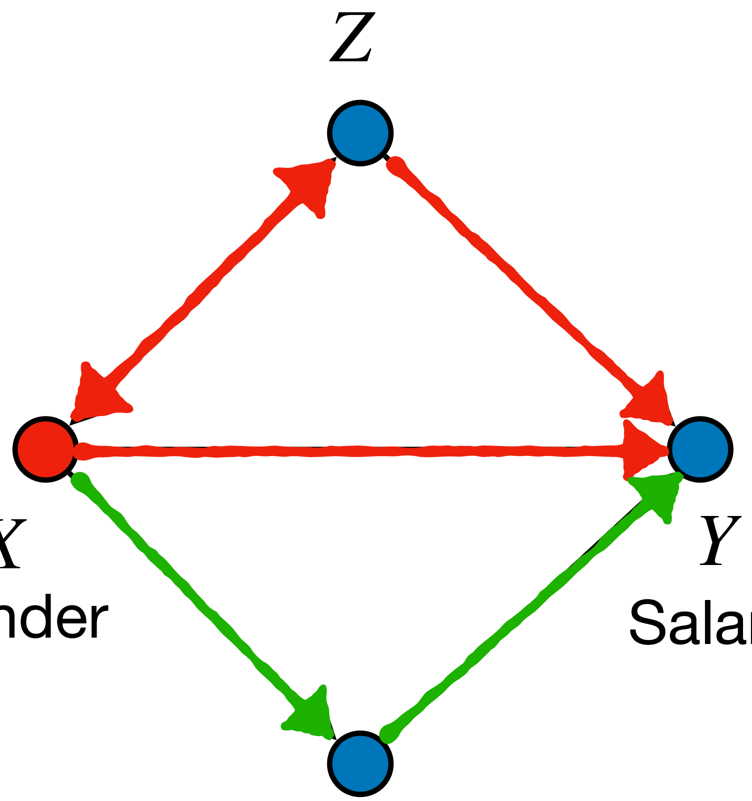
(2) Decisions were based on education or employment: $X \to W \to Y$.

(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



- After collecting data, it has been observed that

$$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}] > 0.$$

How could the observed disparity be explained?

(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: $X \leftrightarrow Z \to Y$.

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.

How to disentangle these variations within TV?

The Attribution Problem

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

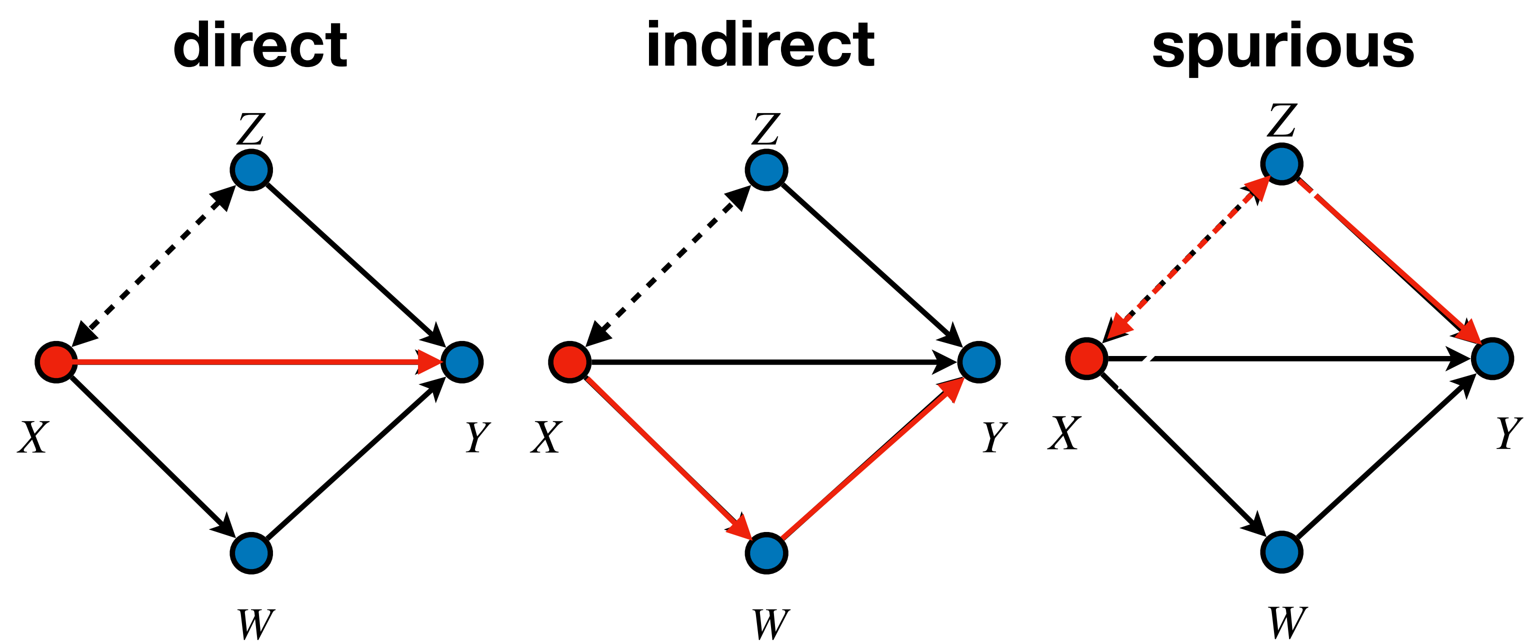
$\text{TV} = E[Y \mid \text{male}] - E[Y \mid \text{female}]$

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!

Admiss

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

Definition. *Let Ω be a*

criterion Q and measure

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

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

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

(i) μ' is admissible

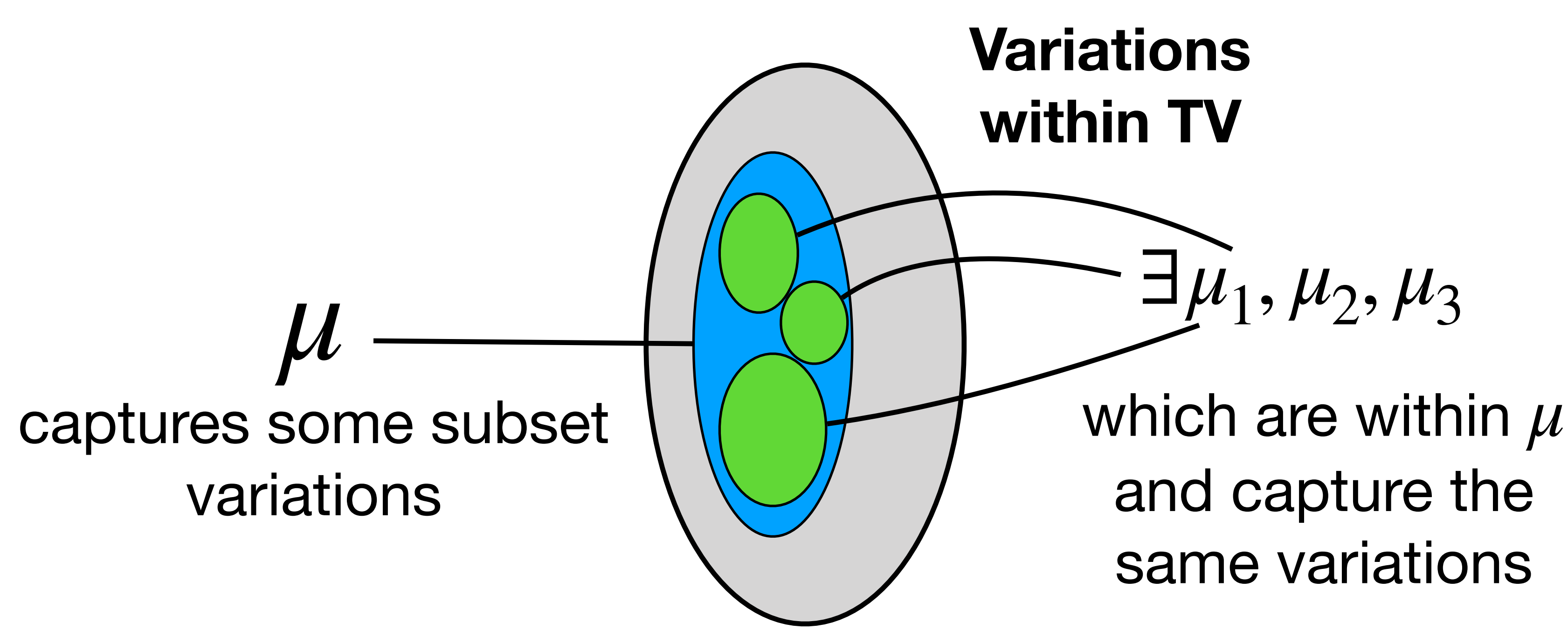
(ii) $\mu'(\mathscr{M}) = 0 \implies \mu(\mathscr{M}) = 0.$

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*

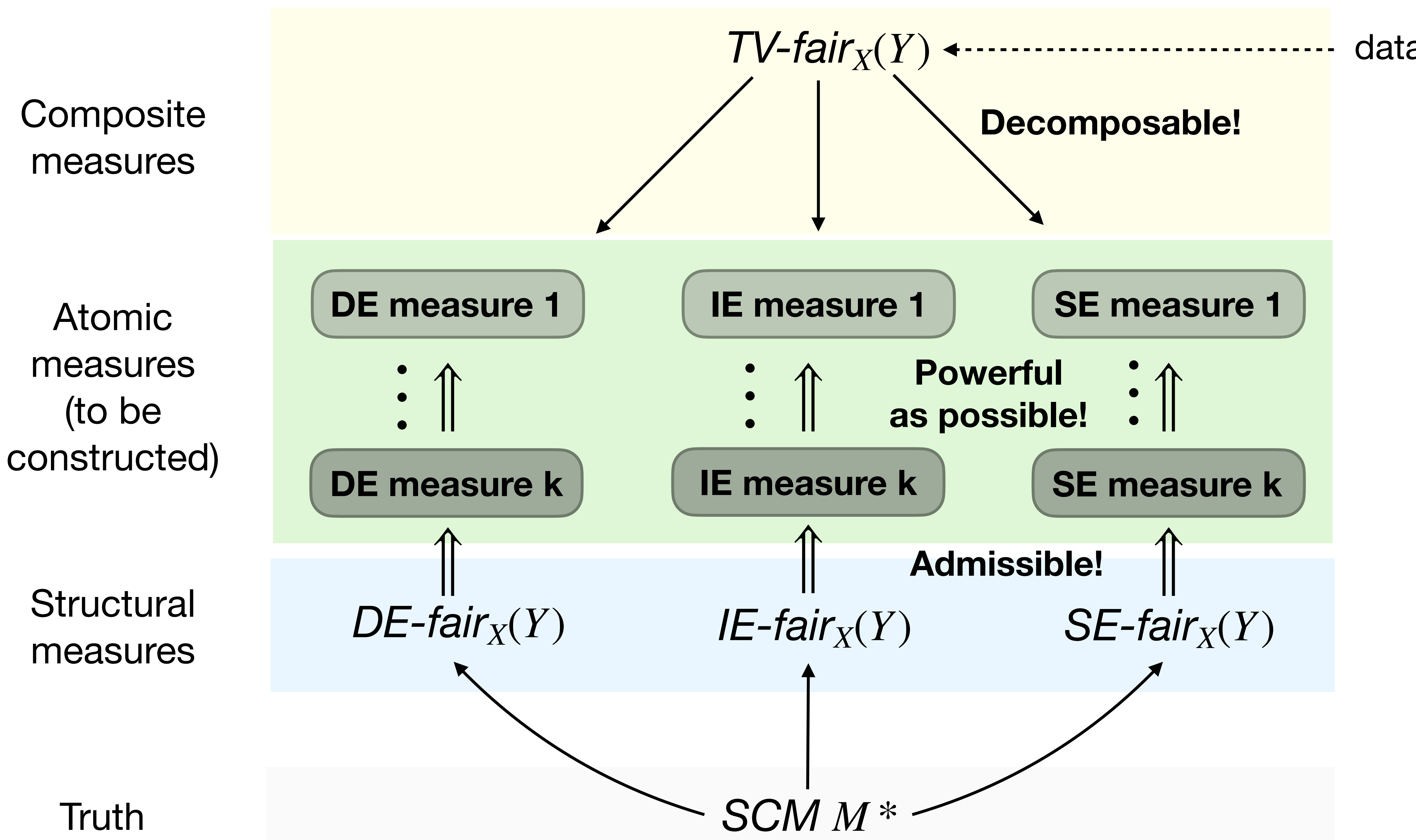
$$\mu_1, \ldots, \mu_k \text{ such that } \mu = f(\mu_1, \ldots, \mu_k),$$

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 - Summary



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 $\mu_1, \ldots, \mu_k$ s.t. the following properties hold:

(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**

How to solve the FPCFA?

Section 3.1
Definition 13

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.

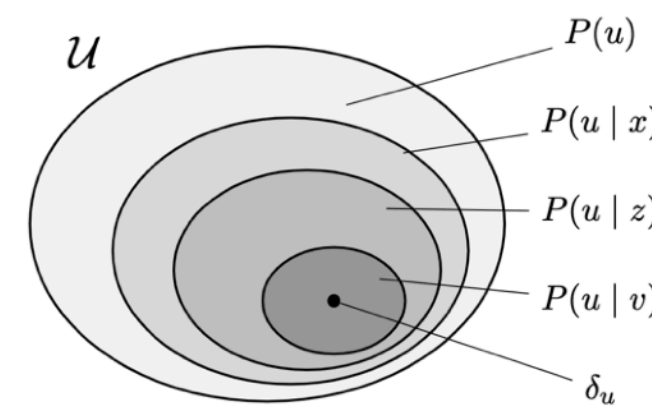
difference arising from counterfactuals C_0, C_1 used to capture the causal influence of X on Y.

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*

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



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

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:*

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

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 $P(u)$

Furthermore:

A. *Any* ... *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:*

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

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 $P(u)$

Furthermore:

A. *Any counterfactual contrast ($E_0 = E_1 = E$) admits the structural basis expansion of the form:*

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

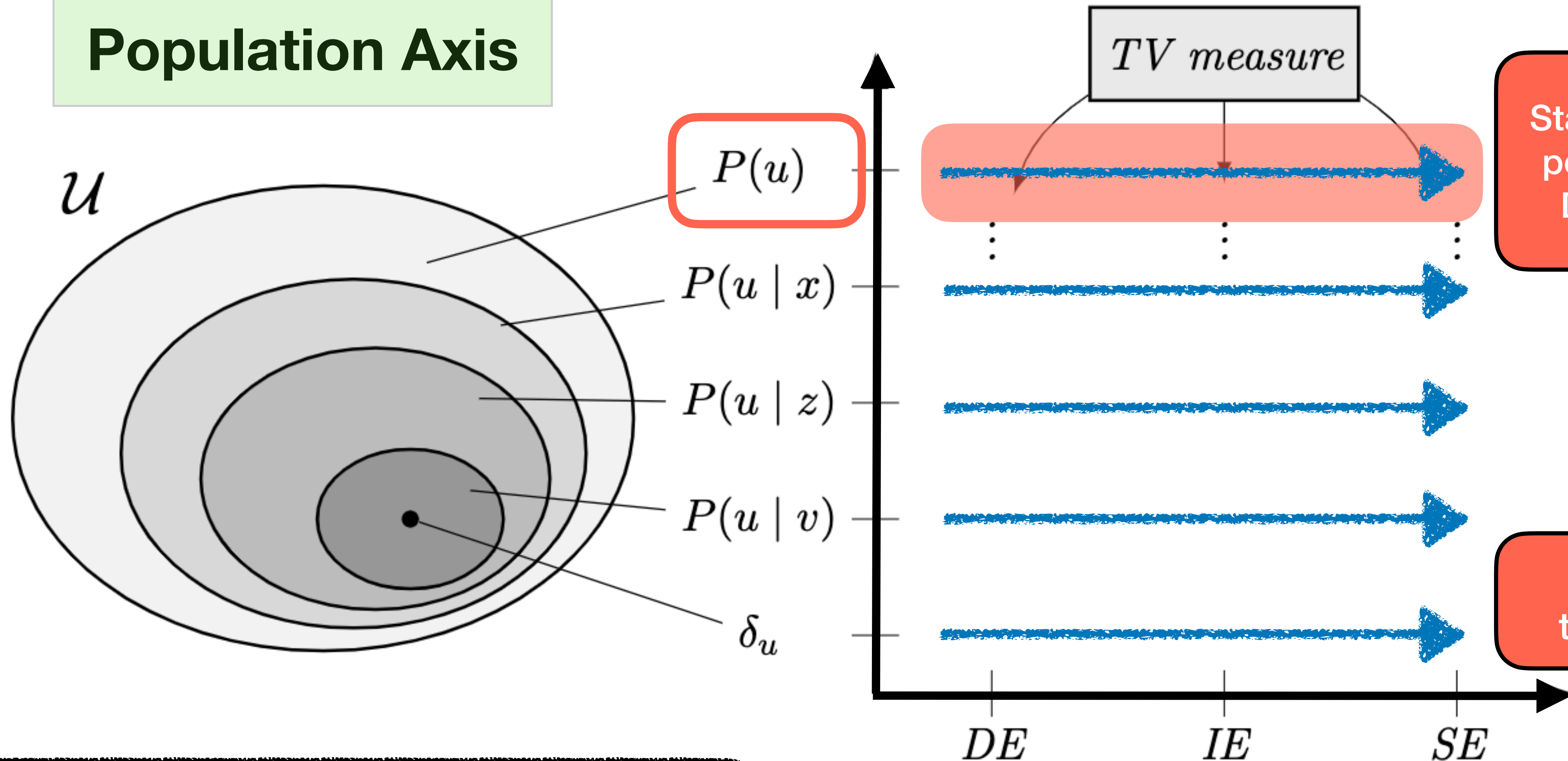
B. *any factual contrast ($C_0 = C_1 = C$) admits the structural basis expansion of the form:*

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

Explainability Plane

Section 3.2
Figure 7

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.

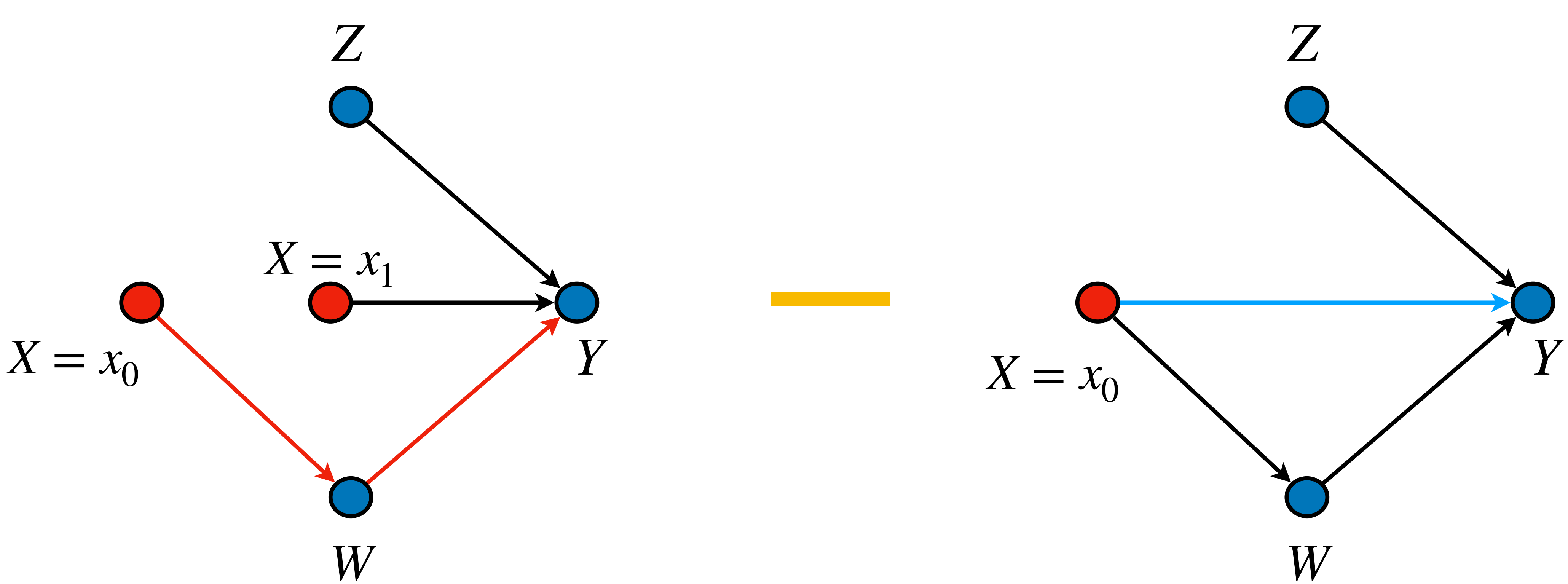
Different values of interventions $C_0 = c_0, C_1 = c_1$ imply a different mechanism's selection (direct, indirect, spurious).

TV family of causal fairness measures

Gedankenexperiment (NDE)

- For a male employee ($X = x_0$), how would his salary (Y) change had he been a female ($X = x_1$), while keeping the age, nationality, education, employment status unchanged (i.e., at the natural level $X = x_0$)?

$$\textbf{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})$$

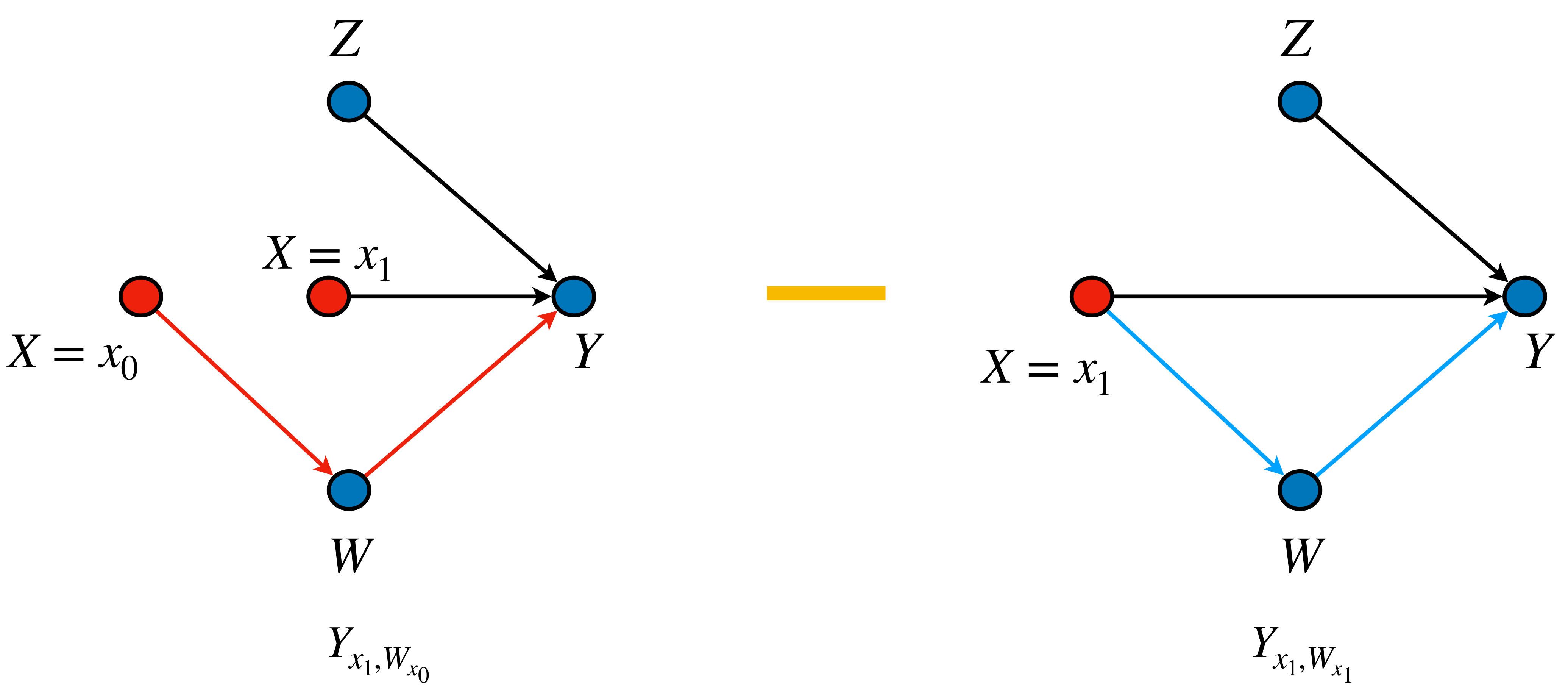


$Y_{x_1,W_{x_0}}$ $Y_{x_0,W_{x_0}}$

Gedankenexperiment (NIE)

- For a female employee ($X = x_1$), how would her salary (Y) change had she been a male ($X = x_0$), while keeping gender unchanged along the direct causal pathway (at the natural level $X = x_1$)?

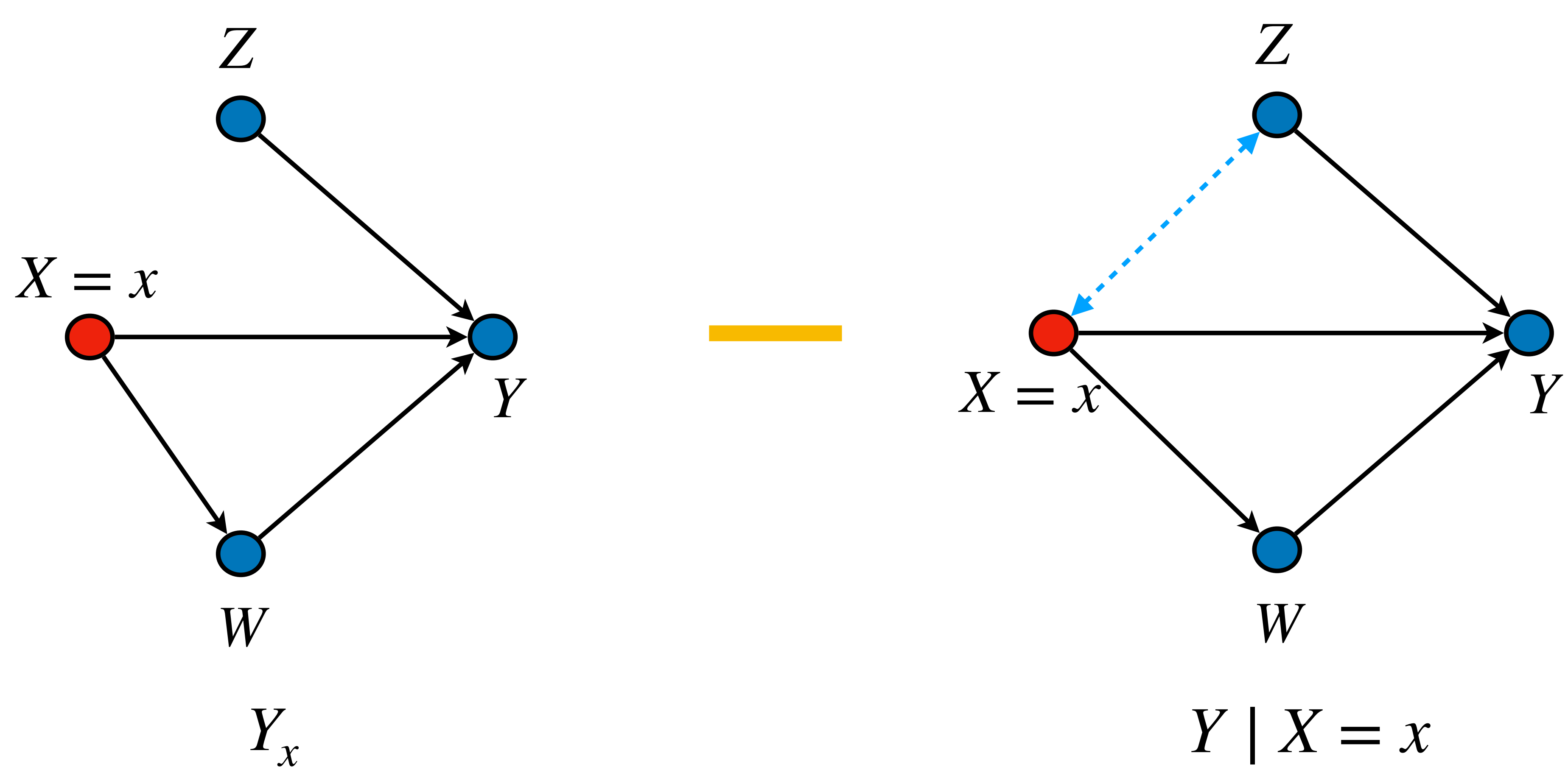
$$\textbf{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1,W_{x_1}})$$



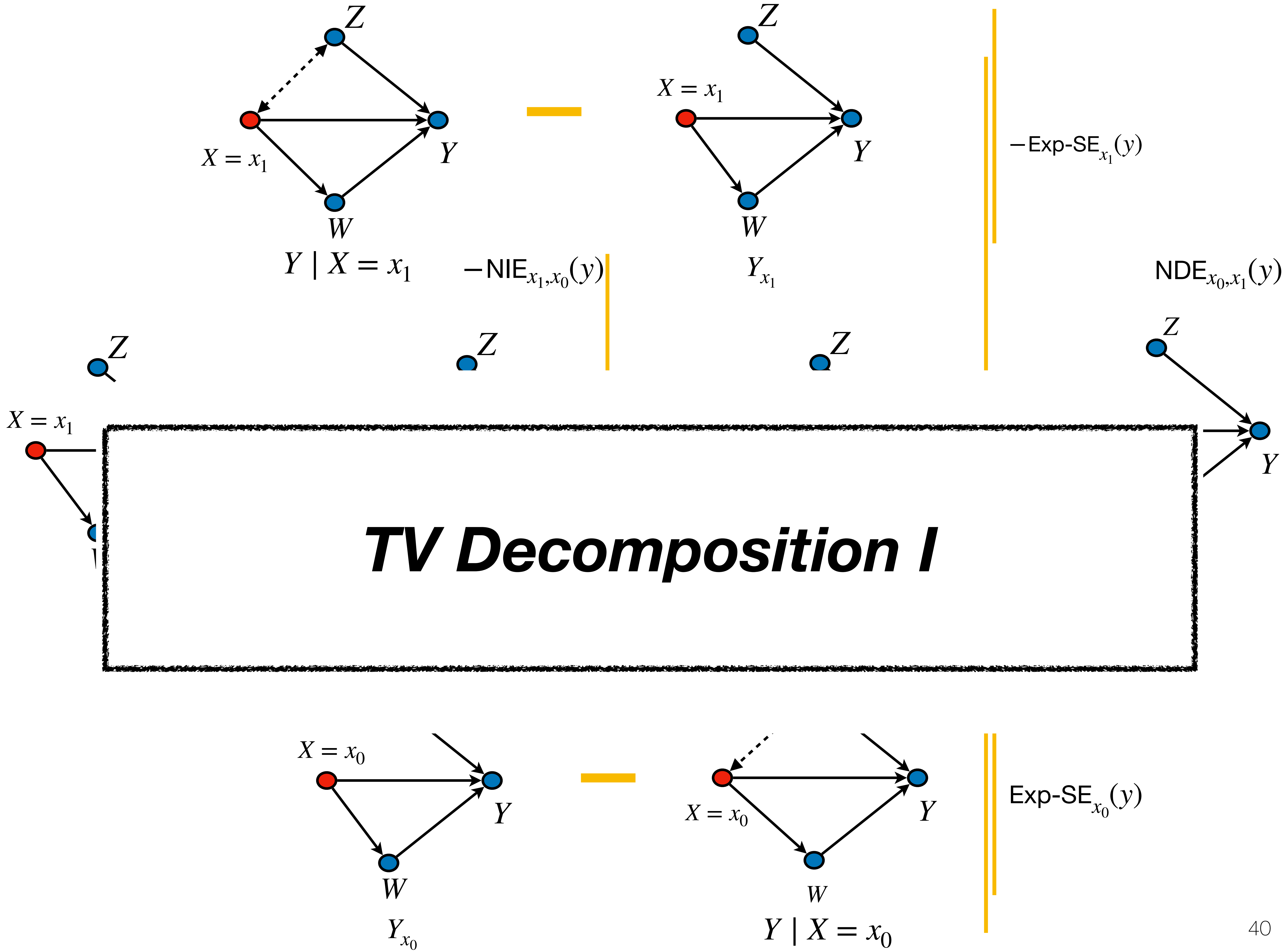
Gedankenexperiment (Exp-SE)

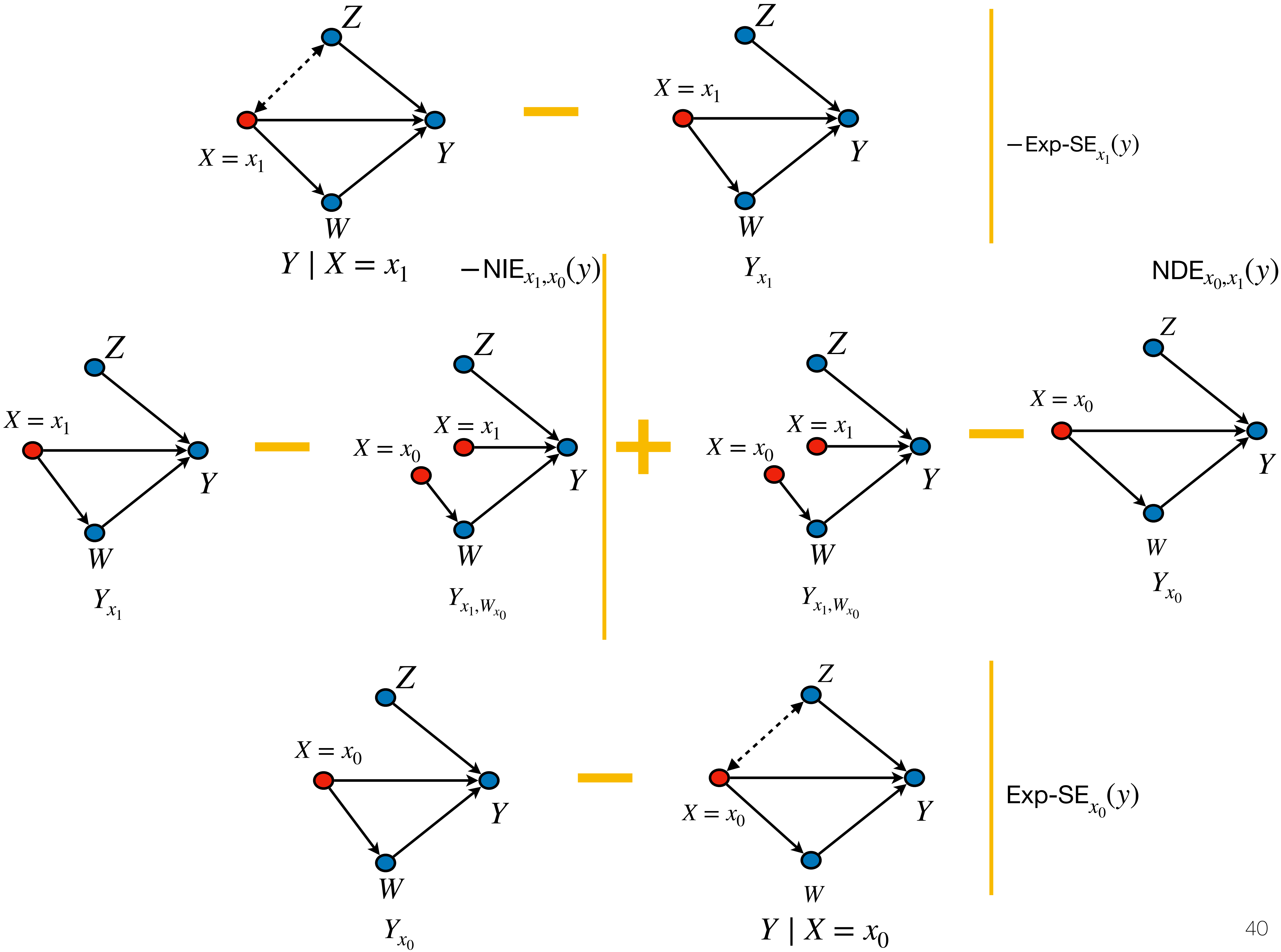
- How would an individuals salary (Y) change if their gender is set to male (or female) by intervention, compared to observing their salary as male (female)?

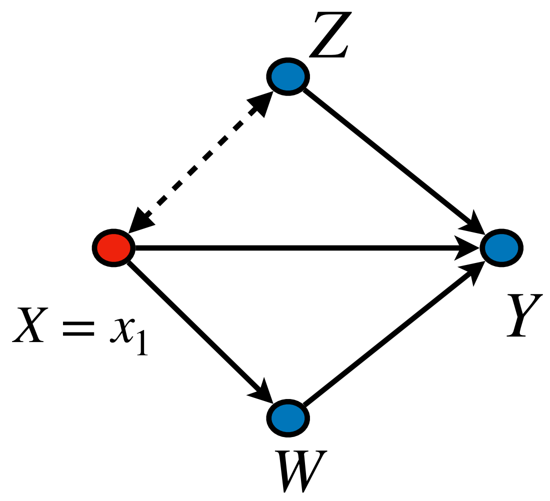
$$\textbf{Exp-SE}_x(y) = P(y_x) - P(y \mid x)$$



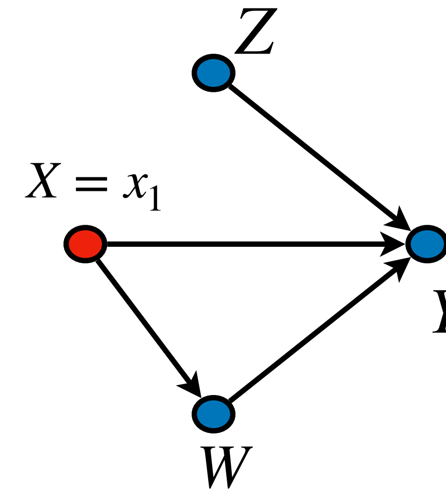
TV Decomposition I




$Y \mid X = x_1$
Y_{x_1}
$-\text{Exp-SE}_{x_1}(y)$
$-\text{NIE}_{x_1,x_0}(y)$
Y_{x_1}
$Y_{x_1,W_{x_0}}$
$Y_{x_1,W_{x_0}}$
Y_{x_0}
$\text{NDE}_{x_0,x_1}(y)$
Y_{x_0}
$Y \mid X = x_0$
$\text{Exp-SE}_{x_0}(y)$



$Y \mid X = x_1$

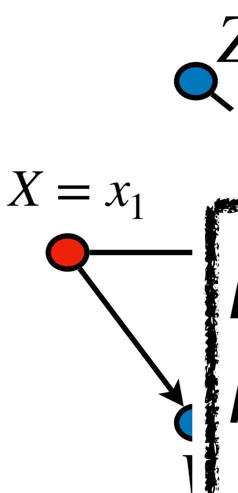


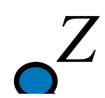
Y_{x_1}

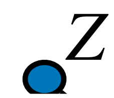
$-$Exp-SE$_{x_1}(y)$

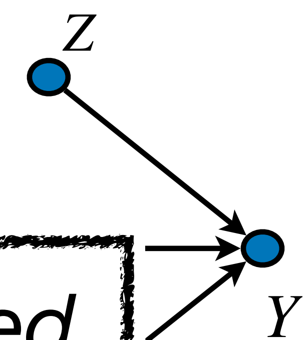
$-$NIE$_{x_1,x_0}(y)$

NDE$_{x_0,x_1}(y)$



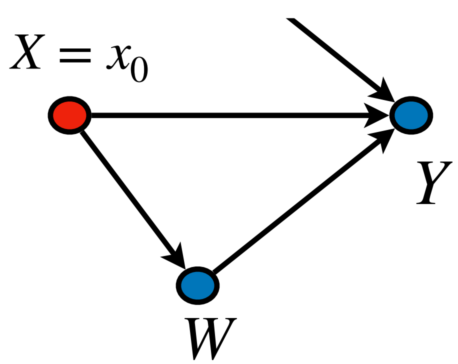




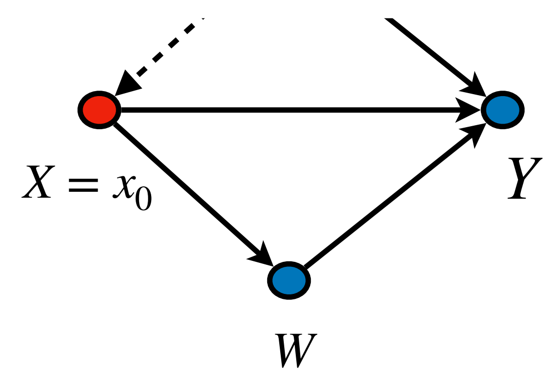


Lemma. *The total variation measure can be decomposed into its direct, indirect, and spurious variations:*

$$TV_{x_0,x_1}(y) = NDE_{x_0,x_1}(y) - NIE_{x_1,x_0}(y) - (\text{Exp-SE}_{x_1}(y) - \text{Exp-SE}_{x_0}(y))\,.$$



Y_{x_0}



$Y \mid X = x_0$

Exp-SE$_{x_0}(y)$

Relation to Structural Fairness

Corollary. *The criteria based on NDE, NIE, and Exp-SE measures are admissible with respect to structural direct, indirect, and spurious fairness. Formally, these facts are written as:*

$$\begin{aligned} \text{S-DE} &\implies \text{NDE-fair} \\ \text{S-IE} &\implies \text{NIE-fair} \\ \text{S-SE} &\implies \text{Exp-SE-fair} \end{aligned}$$

admissibility w.r.t. structural

In practice, for example, by computing the NDE, we can test for the presence of structural direct effect.

Testing Structural Fairness in Practice

- Our previous corollary shows that

$$\text{S-DE} \implies \text{NDE-fair}\,.$$

- By taking this statement's contrapositive, we can see that

$$\text{NDE}_{x_0,x_1}(y) \neq 0 \implies \neg\text{S-DE}\,.$$

- Therefore, in practice, one may use the following hypothesis testing procedure for testing structural direct effect,

$$H_0 : \text{NDE}_{x_0,x_1}(y) = 0.$$

A similar approach can be used for the NIE and Exp-SE since

$$\text{S-IE} \implies \text{NIE-fair}$$

$$\text{S-SE} \implies \text{Exp-SE-fair}$$

This will be used to connect with the disparate treatment and impact doctrines later on.

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

SCM M*

(unobserved)

$U \leftarrow N(0,1)$

$X \leftarrow \text{Bernoulli}(expit(U))$

$Z \leftarrow \text{Bernoulli}(expit(U))$

$W \leftarrow \text{Bernoulli}(0.3)$

$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X + Z - 2XZ) + \frac{1}{6}W)$

$$
\begin{aligned}
\textbf{NDE}_{x_0,x_1}(y) &= P(y_{x_1,W_{x_0}}) - P(y_{x_0}) \\
&= P(\textbf{Bernoulli}(\frac{1}{5}(1-Z) + \frac{1}{6}W) = 1) \\
&\quad - P(\textbf{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1) \\
&= \sum_{z\in\{0,1\}} \sum_{w\in\{0,1\}} P(w)[\frac{1}{5}(1-2z) + \frac{1}{6}w - \frac{1}{6}w] \\
&= \sum_{z\in\{0,1\}} \frac{1}{5}(1-2z) = 0.
\end{aligned}
$$

Section 4
Example 9

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$ indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

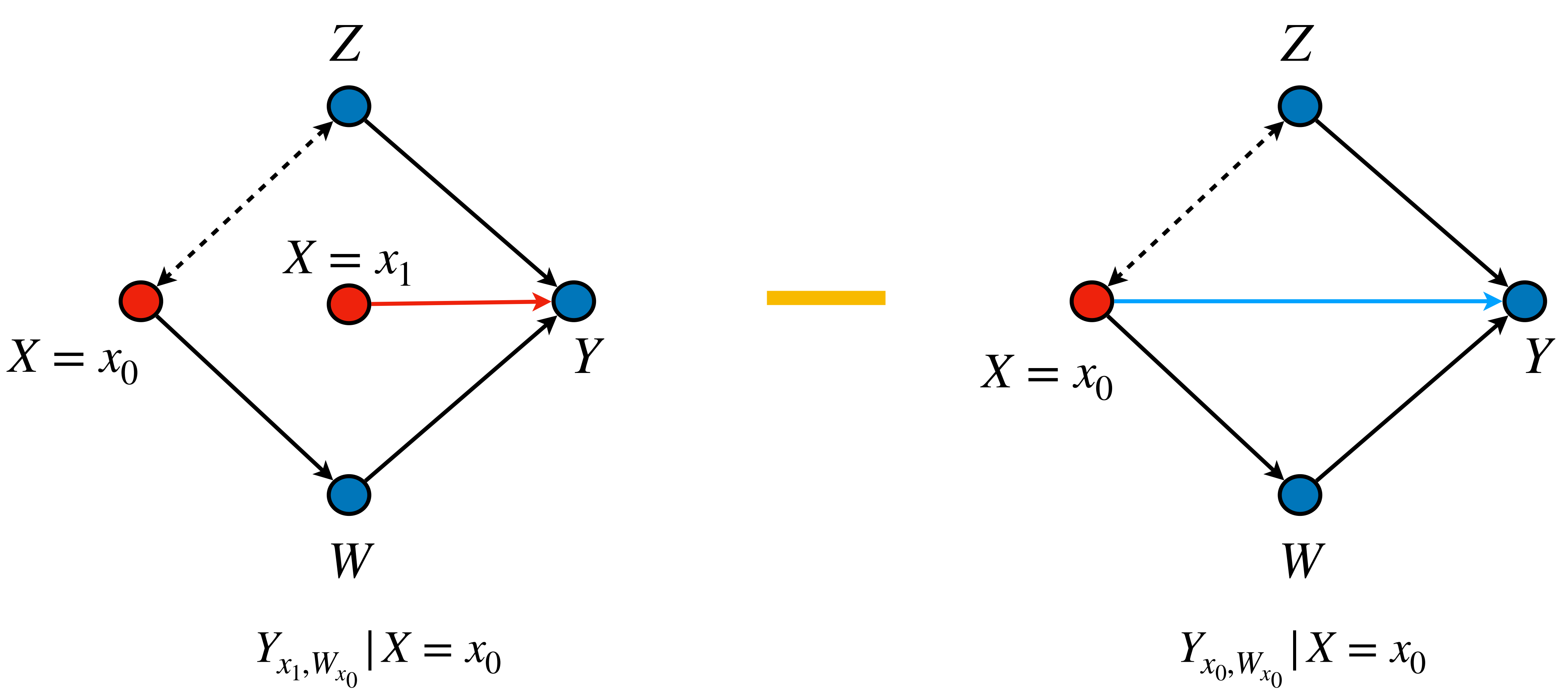
NDE is admissible w.r.t. S-DE.
However, here NDE = 0,
but structural direct effect exists.

Q: Is NDE powerful enough for detecting direct discrimination?

Gedankenexperiment (Ctf-DE)

- For a male employee $X = x_0$, how would his salary change (Y) had he been a female ($X = x_1$), while keeping the age, nationality, education and employment status unchanged (at the level of

$$\textbf{Ctf-DE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0,W_{x_0}} \mid x_0)$$



Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

SCM M

$U \leftarrow N(0,1)$

$X \leftarrow \text{Bernoulli}(expit(U))$

$Z \leftarrow \text{Bernoulli}(expit(U))$

$W \leftarrow \text{Bernoulli}(0.3)$

$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X + Z - 2XZ) + \frac{1}{6}W)$

$$\begin{aligned}
\textbf{Ctf-DE}_{x_0,x_1}(y \mid x_0) &= P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0} \mid x_0) \\
&= P(\textbf{Bernoulli}(\frac{1}{5}(1 - Z) + \frac{1}{6}W) = 1 \mid x_0) \\
&\quad - P(\textbf{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1 \mid x_0) \\
&= \sum_{z \in \{0,1\}} \sum_{w \in \{0,1\}} P(w)P(z \mid x_0)[\frac{1}{5}(1 - 2z) + \frac{1}{6}w - \frac{1}{6}w] \\
&= \sum_{z \in \{0,1\}} \frac{1}{5}(1 - 2z)P(z \mid x_0) = 0.036.
\end{aligned}$$

Section 4
Example 10

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

$U \leftarrow N$

$X \leftarrow$ B

$Z \leftarrow$ B

$W \leftarrow$ Bernoulli(0.5)

$Y \leftarrow$ Bernoulli$(\frac{1}{5}(X+Z-2XZ)+\frac{1}{6}W)$

$x_0)$

$\sum_{z\in\{0,1\}} \sum_{w\in\{0,1\}} \ldots \frac{1}{5} \ldots z) + \frac{1}{6}w - \frac{1}{6}w]$

$$= \sum_{z\in\{0,1\}} \frac{1}{5}(1-2z)P(z \mid x_0) = 0.036.$$

Key properties of Ctf-DE:

1. Ctf-DE is admissible.
2. Ctf-DE is more powerful than NDE.

Section 4
Example 10

x-specific measures

Definition. *The effect of treatment on the treated and counterfactual direct, indirect, and spurious effects are defined as*

$$\textit{ETT}_{x_0,x_1}(y \mid x) = P(y_{x_1} \mid x) - P(y_{x_0} \mid x)$$

$$\textit{Ctf-DE}_{x_0,x_1}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x)$$

$$\textit{Ctf-IE}_{x_1,x_0}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_1} \mid x)$$

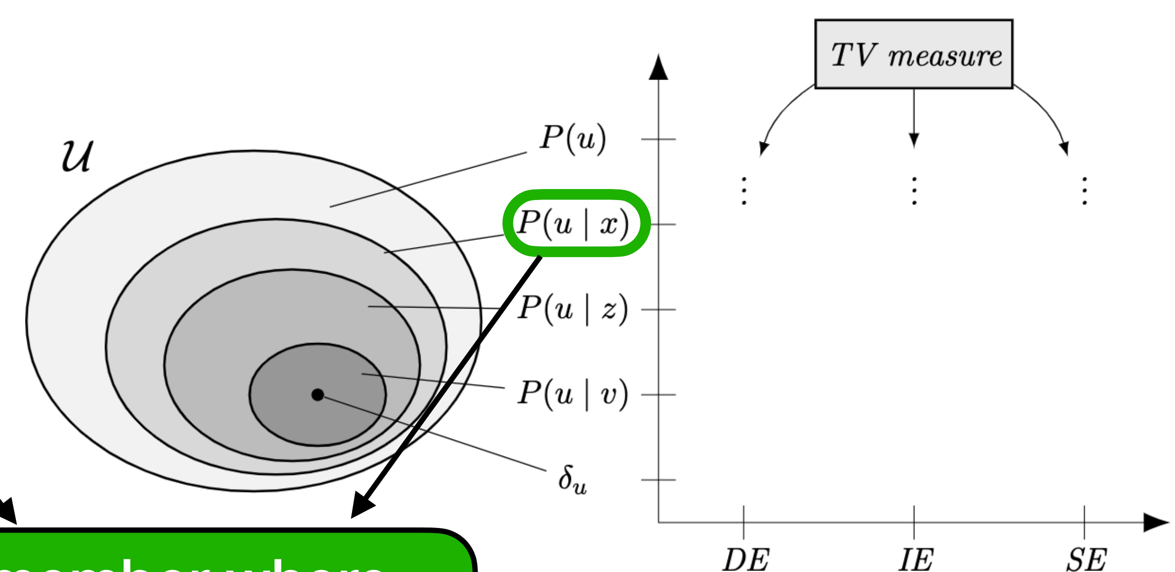
$$\textit{Ctf-SE}_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y_{x_0} \mid x_0)\,.$$

Structural Basis Expansion:

$$\textbf{Ctf-DE}_{x_0,x_1}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]\boxed{P(u \mid x)}$$

$$\textbf{Ctf-IE}_{x_1,x_0}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]\boxed{P(u \mid x)}$$

$$\textbf{Ctf-SE}_{x_0,x_1}(y) = \sum_u y_{x_0}(u)[P(u \mid x_1) - P(u \mid x_0)]\,.$$



remember where we are within $\mathcal{U}$

x-specific

$$\text{TE}_{x_0,x_1}(y \mid x) = P(y_{x_1}) - P(y_{x_0})$$
$$\text{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0})$$
$$\text{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1})$$
$$\text{Exp-SE}_{x_0,x_1}(y) = P(y_x) - P(y_x \mid x)\,.$$

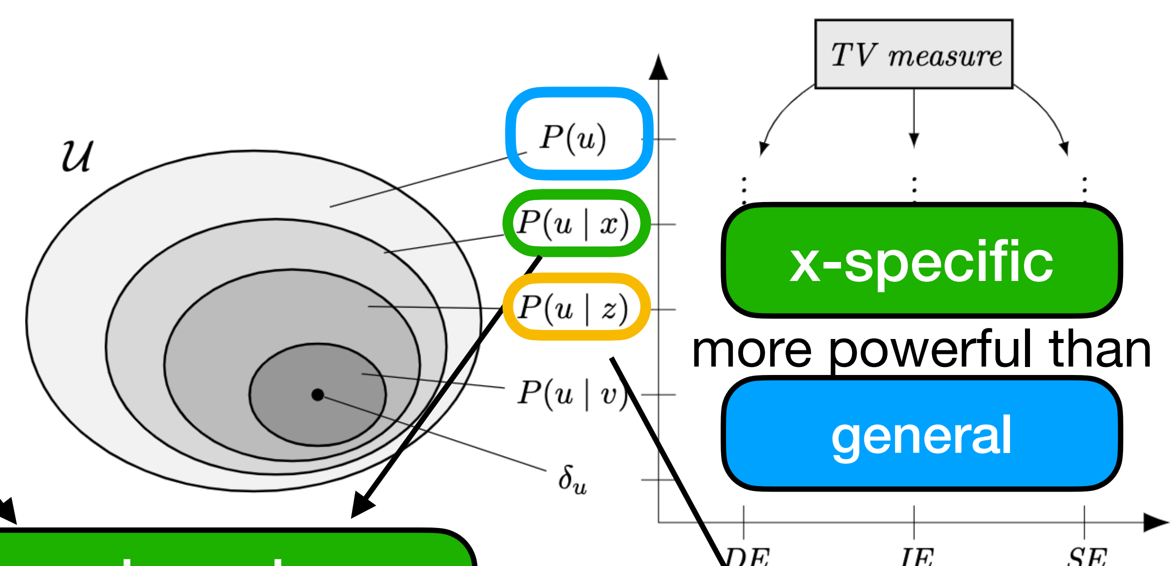
where we came from

Definition. *The effect of treatment on*
direct, indirect, and spurious effects ar

$$\textit{ETT}_{x_0,x_1}(y \mid x) = P(y_{x_1} \mid x) - P(y_{x_0} \mid x)$$
$$\textit{Ctf-DE}_{x_0,x_1}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x)$$
$$\textit{Ctf-IE}_{x_1,x_0}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_1} \mid x)$$
$$\textit{Ctf-SE}_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y_{x_0} \mid x_0)\,.$$

Structural Basis Expansion:

$$\textbf{Ctf-DE}_{x_0,x_1}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]P(u \mid x)$$
$$\textbf{Ctf-IE}_{x_1,x_0}(y \mid x) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]P(u \mid x)$$
$$\textbf{Ctf-SE}_{x_0,x_1}(y) = \sum_u y_{x_0}(u)[P(u \mid x_1) - P(u \mid x_0)]\,.$$



remember where we are within $\mathcal{U}$

where we go next

z-specific measures

Definition. *The z-specific total, direct, and indirect effects are defined as*

$$z\text{-}TE_{x_0,x_1}(y \mid z) = P(y_{x_1} \mid z) - P(y_{x_0} \mid z)$$

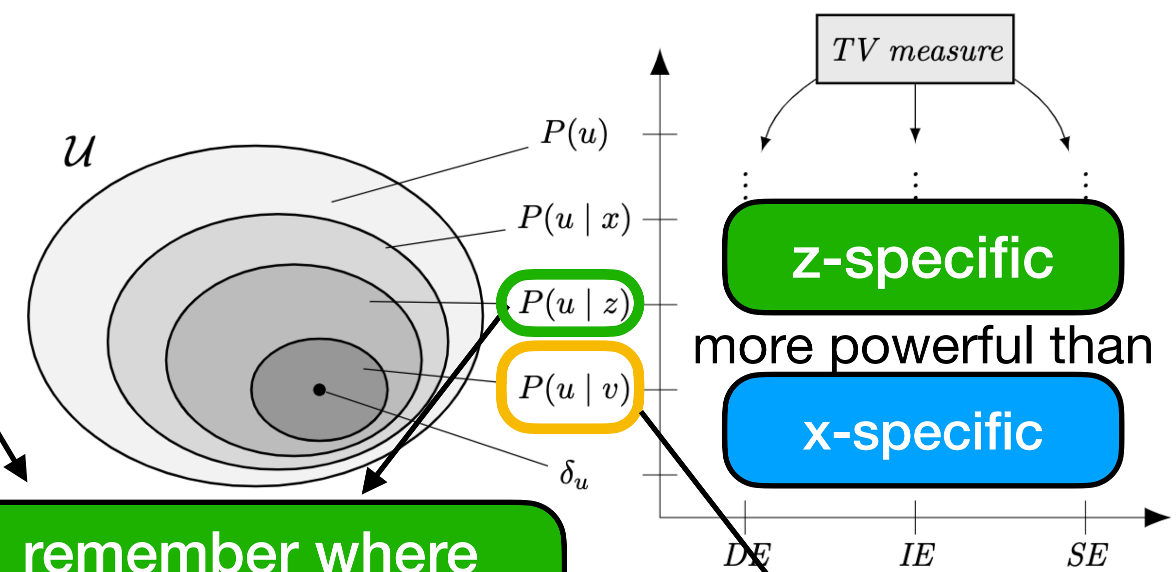
$$z\text{-}DE_{x_0,x_1}(y \mid z) = P(y_{x_1,W_{x_0}} \mid z) - P(y_{x_0} \mid z)$$

$$z\text{-}IE_{x_1,x_0}(y \mid z) = P(y_{x_1,W_{x_0}} \mid z) - P(y_{x_1} \mid z)\,.$$

Structural Basis Expansion:

$$z\text{-}\mathbf{DE}_{x_0,x_1}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]\boxed{P(u \mid z)}$$

$$z\text{-}\mathbf{IE}_{x_1,x_0}(y \mid z) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]\boxed{P(u \mid z)}.$$



remember where we are within $\mathcal{U}$

where we go next

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

SCM M

$U \leftarrow N(0,1)$
$X \leftarrow \text{Bernoulli}(expit(U))$
$Z \leftarrow \text{Bernoulli}(expit(U))$
$W \leftarrow \text{Bernoulli}(0.3)$
$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X+Z-2XZ)+\frac{1}{6}W)$

$$\begin{aligned}
z\textbf{-DE}(y \mid Z=0) &= P(y_{x_1,W_{x_0}} \mid Z=0) - P(y_{x_0} \mid Z=0) \\
&= P(\textbf{Bernoulli}(\frac{1}{5}(1-Z)+\frac{1}{6}W)=1 \mid Z=0) \\
&\quad -P(\textbf{Bernoulli}(\frac{1}{5}(Z)+\frac{1}{6}W)=1 \mid Z=0) \\
&= \sum_{w\in\{0,1\}} P(w)[\frac{1}{5}+\frac{1}{6}w-\frac{1}{6}w] = \frac{1}{5}.
\end{aligned}$$

Section 4 Example 11

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$ indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

$U \leftarrow N$

$X \leftarrow$ Be

$Z \leftarrow$ Be

$W \leftarrow$ Be

$Y \leftarrow$ Bernoulli($\frac{\ }{5}(X + Z - 2XZ) + \frac{\ }{6}W$)

$= 1 \mid Z = 0)$

$Z = 0)$

$\sum_{w \in \{0,1\}} \quad 5 \quad 6 \quad 6 \quad \frac{1}{5}.$

Key properties of z-DE:

1. z-DE is admissible.

2. z-DE is more powerful than Ctf-DE.

Section 4
Example 11

v'-specific measures

Definition. *The v'-specific total, direct, and indirect effects are defined as*

$$v'\text{-}TE_{x_0,x_1}(y \mid v') = P(y_{x_1} \mid v') - P(y_{x_0} \mid v')$$

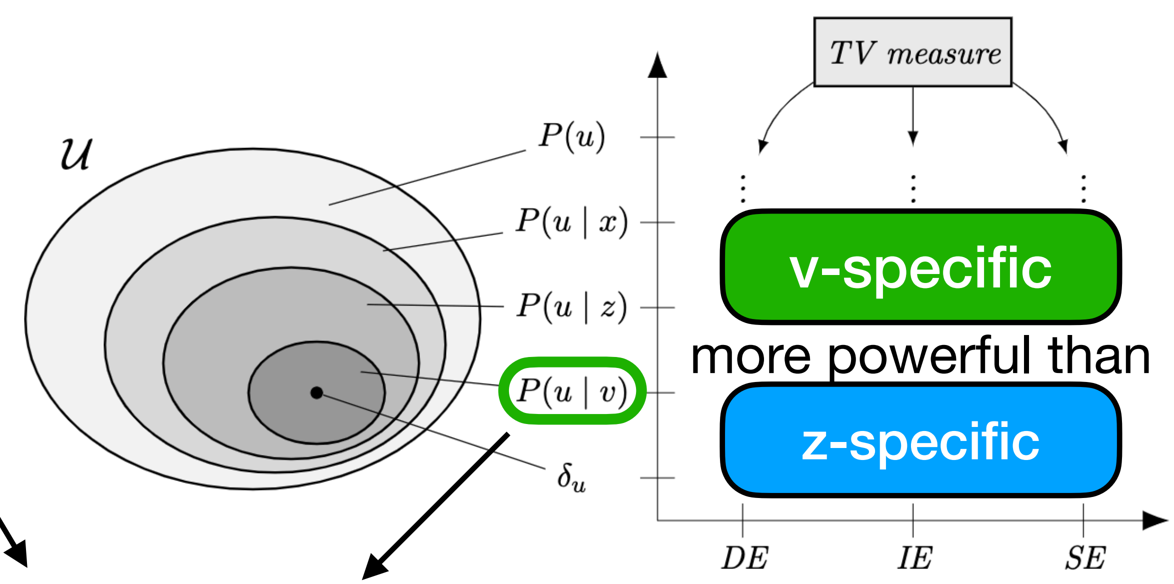
$$v'\text{-}DE_{x_0,x_1}(y \mid v') = P(y_{x_1,W_{x_0}} \mid v') - P(y_{x_0} \mid v')$$

$$v'\text{-}IE_{x_1,x_0}(y \mid v') = P(y_{x_1,W_{x_0}} \mid v') - P(y_{x_1} \mid v') .$$

Structural Basis Expansion:

$$v'\textbf{-DE}_{x_0,x_1}(y \mid v') = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)]P(u \mid v')$$

$$v'\textbf{-IE}_{x_1,x_0}(y \mid v') = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_1}(u)]P(u \mid v').$$



remember where we are within $\mathcal{U}$

Unit-level measures

Definition. *Given a unit $U = u$, the unit-level total, direct, and indirect effects are given by*

$$\textit{unit-TE}_{x_0,x_1}(y(u)) = y_{x_1}(u) - y_{x_0}(u)$$

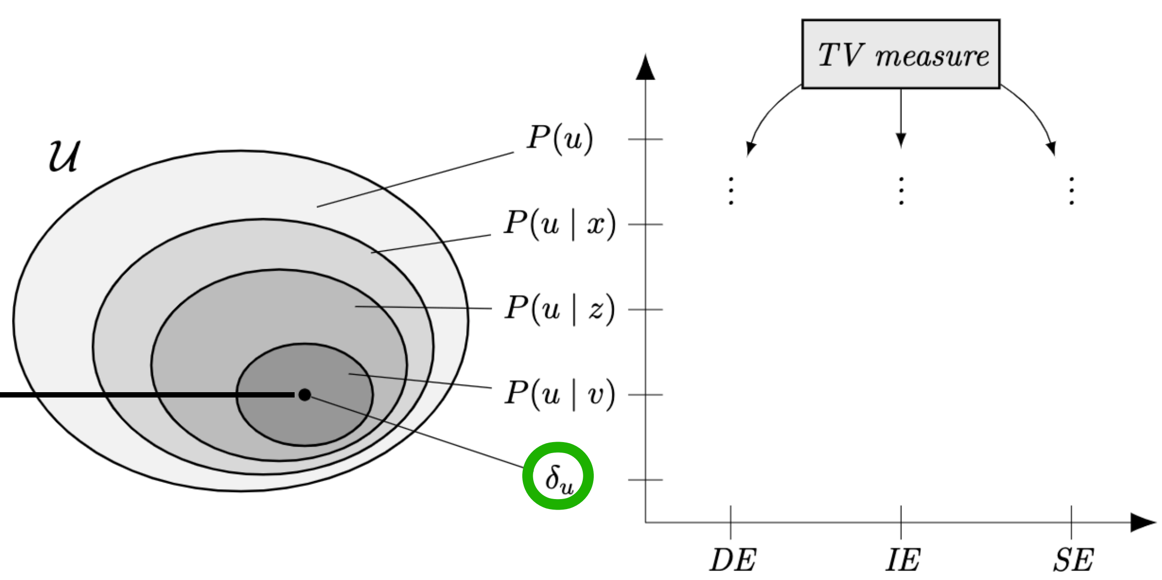
$$\textit{unit-DE}_{x_0,x_1}(y(u)) = y_{x_1,W_{x_0}}(u) - y_{x_0}(u)$$

$$\textit{unit-IE}_{x_1,x_0}(y(u)) = y_{x_1,W_{x_0}}(u) - y_{x_1}(u) .$$

These quantities ***<u>are the structural basis.</u>***

Remember where we are within $\mathscr{U}$.

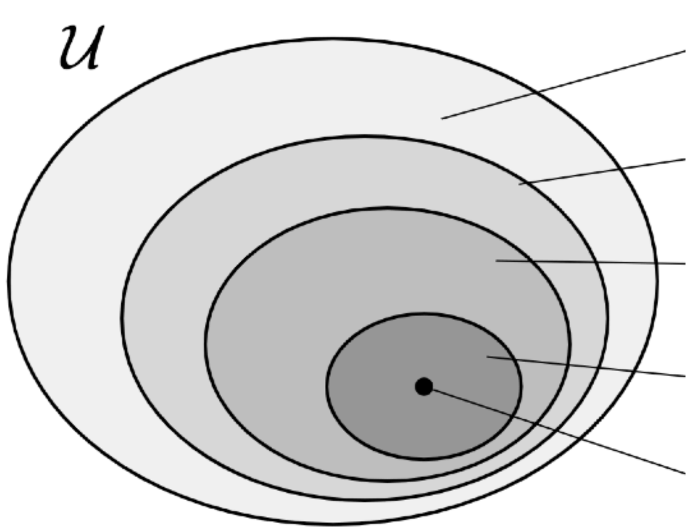
We reached the final, unit-level measures!



TV family measures as contrasts

Lemma. *Under the Standard fairness model, all the measures within the TV family can be written as contrasts* $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$, *following he constructions indicated below.*

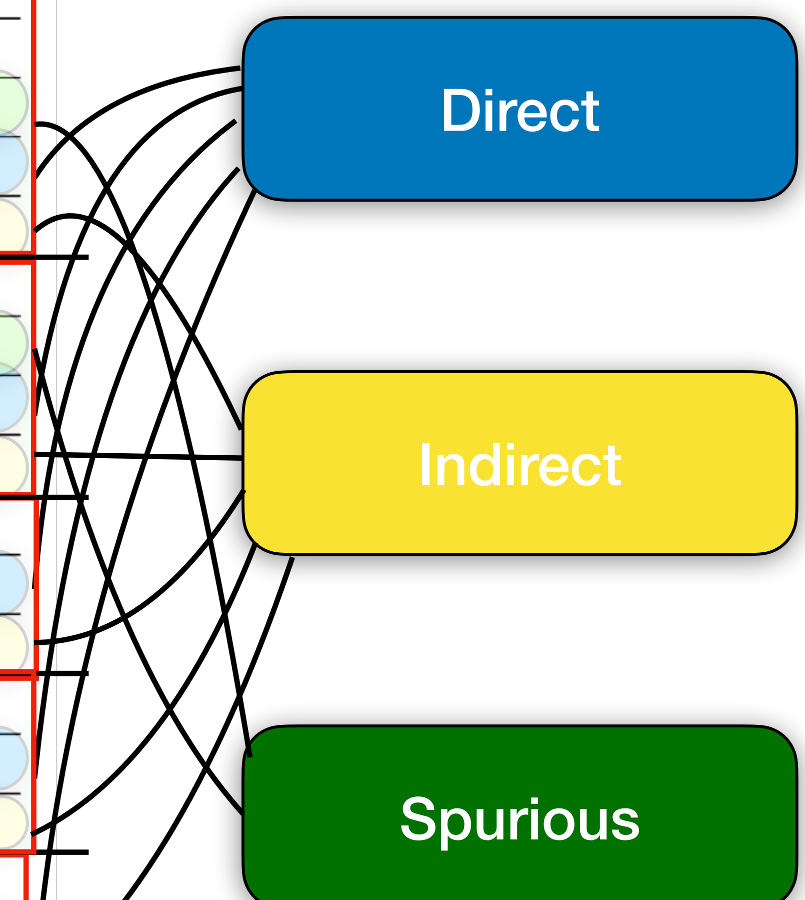
units



$\boxed{\text{mechanism}}$ $\boxed{\text{unit}}$

	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	Exp-SE_x	x	x	$\emptyset$	x
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

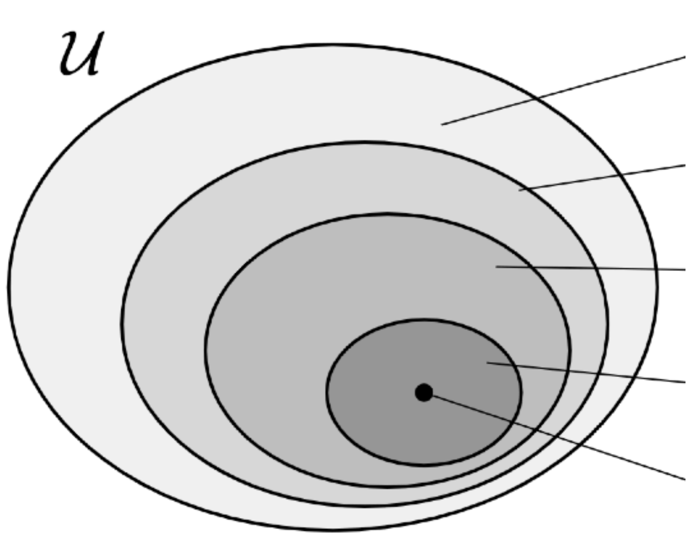
mechanisms



TV family measures as contrasts

Lemma. *Under the Standard fairness model, all the measures within the TV family can be written as contrasts* $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$*, following he constructions indicated below.*

units



mechanism *unit*

	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	Exp-SE_x	x	x	$\emptyset$	x
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

mechanisms

Direct

Indirect

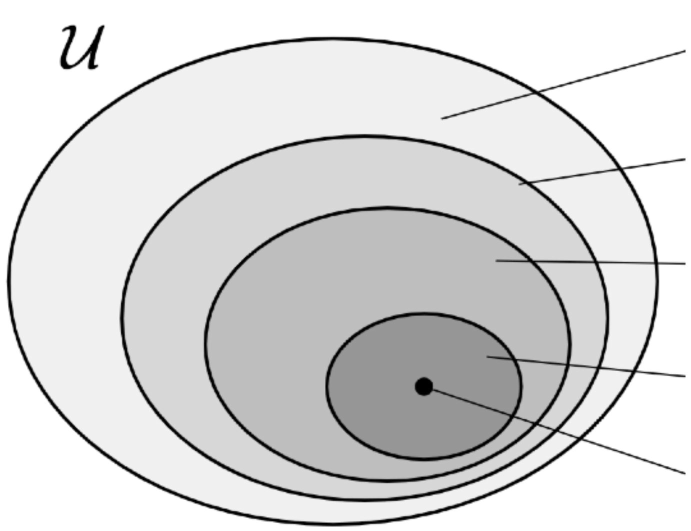
Spurious

TV family measures as contrasts

Lemma. *Under the Standard fairness model, all the measures within the TV family can be written as contrasts* $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$ *, following he constructions indicated below.*

mechanism *unit*

units



	Measure	C_0	C_1	E_0	E_1
	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
general	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	Exp-SE_x	x	x	$\emptyset$	x
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

mechanisms

Direct

Causal

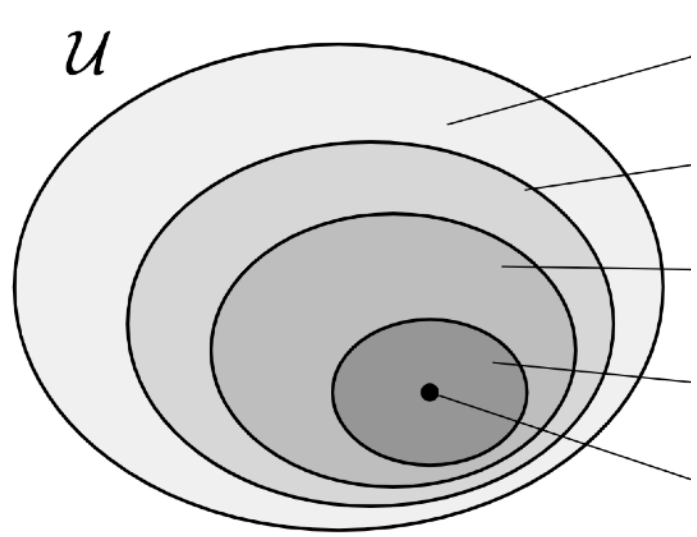
Indirect

Spurious

TV family measures as contrasts

Lemma. *Under the Standard fairness model, all the measures within the TV family can be written as contrasts* $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$ *, following he constructions indicated below.*

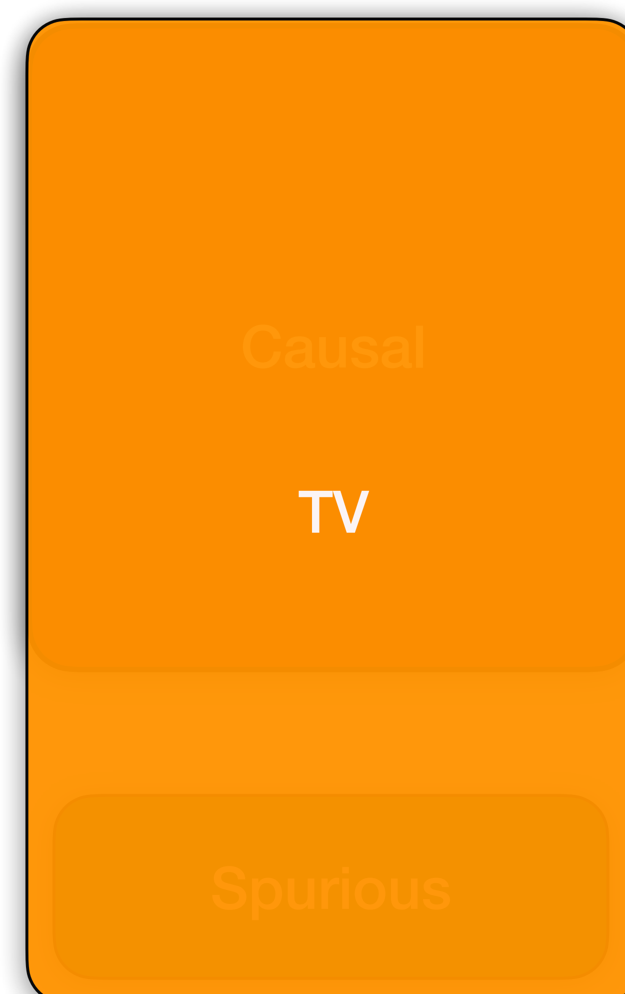
units



mechanism *unit*

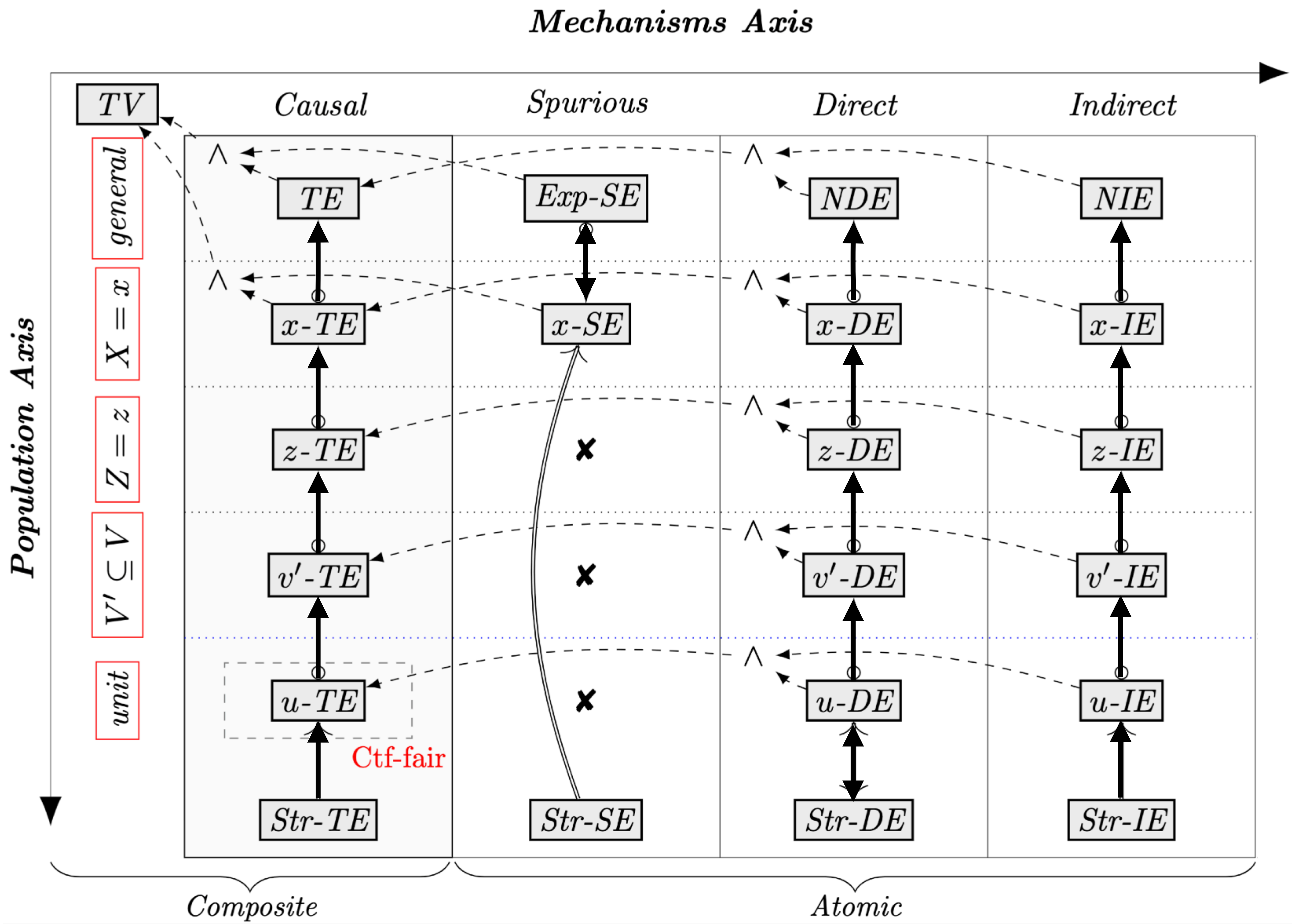
	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
general	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
general	Exp-SE_{x}	x	x	$\emptyset$	x
general	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
general	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
$X = x$	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
$X = x$	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
$X = x$	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
$Z = z$	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
$Z = z$	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
$V' \subseteq V$	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
$V' \subseteq V$	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
unit	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
unit	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

mechanisms



Fairness Map

Fairness Map



x-specific to general

z-specific to x-specific

v'-specific to z-specific

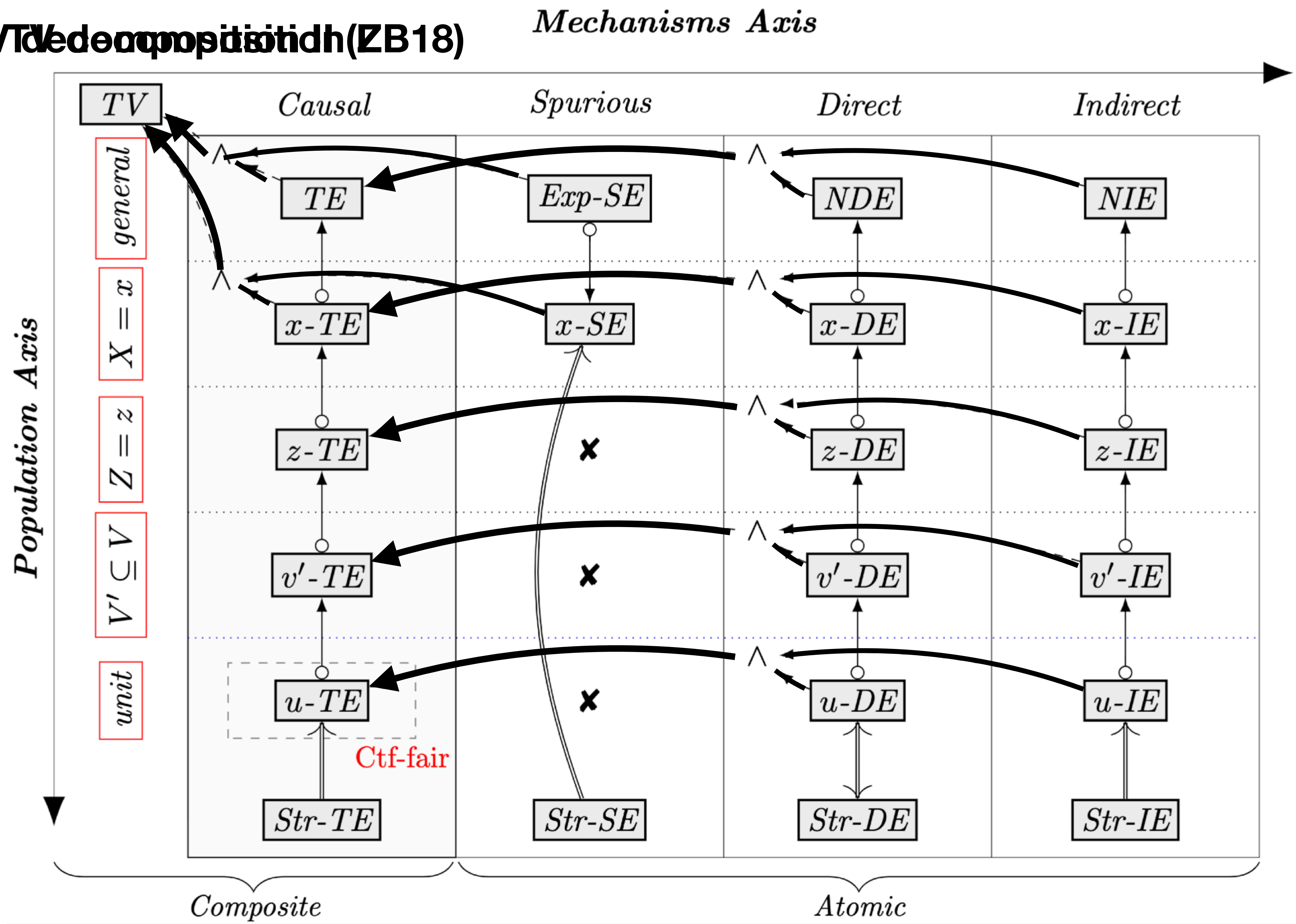
unit to v'-specific

structural to unit

Fairness Map

Section 4.2
Theorem 7

TV decomposition (ZB18)



Mediation formula (Pearl, 2012)

Extended Mediation Formula

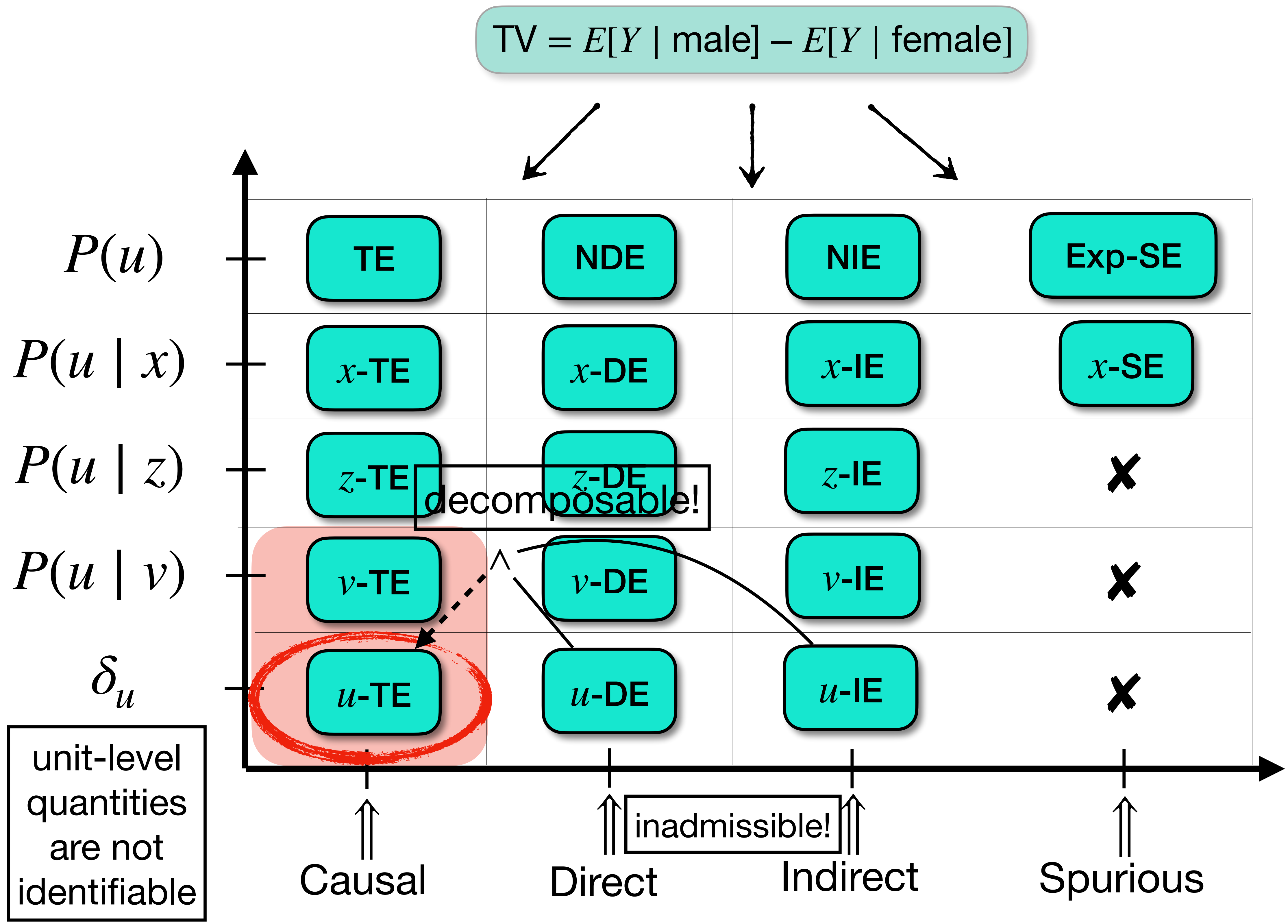
Other connections with the literature

- How does the presented framework relates to other prominent measures in the literature?
- In particular, we consider the following measures:

(i) Counterfactual Fairness (Kusner et. al., ’17)

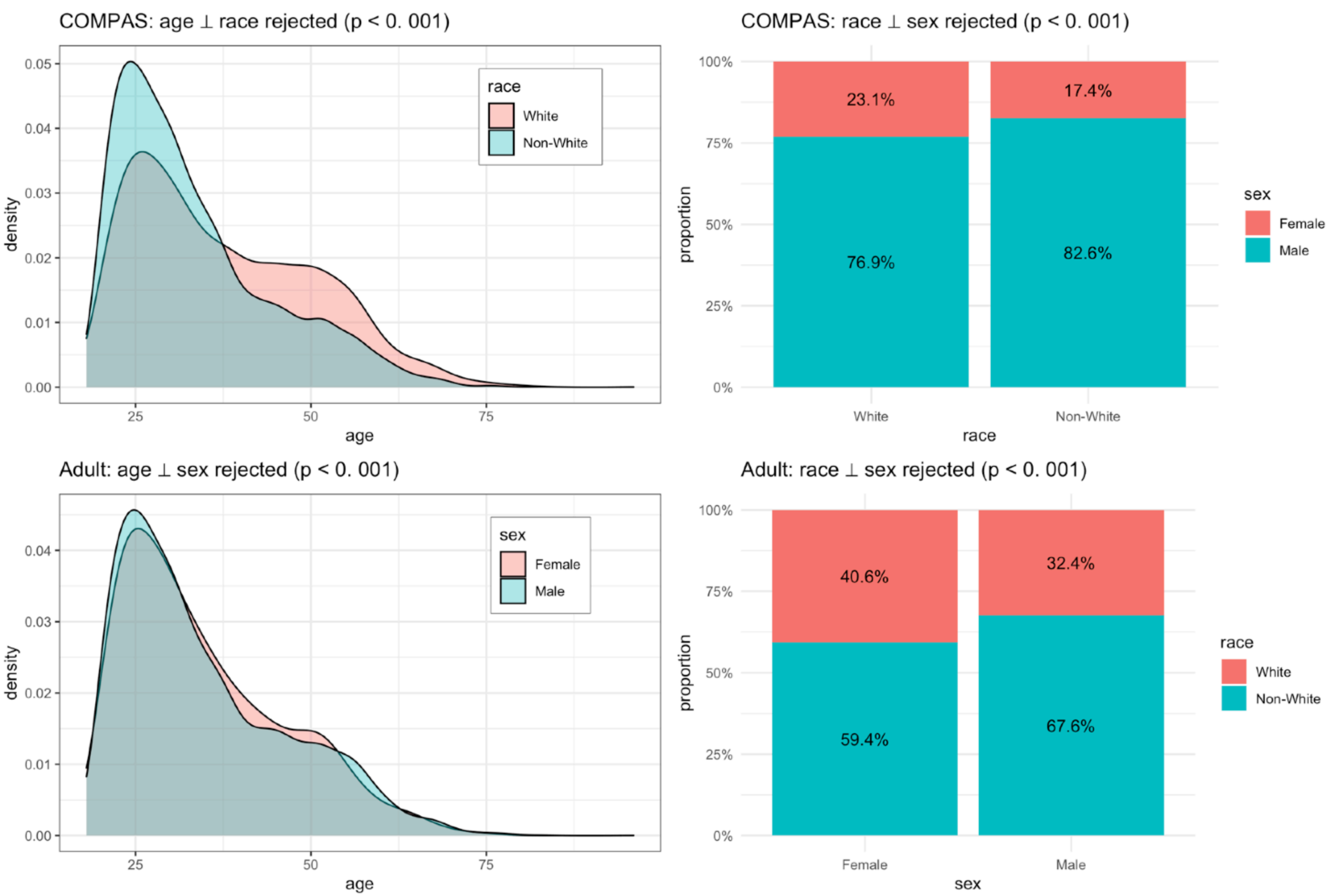
(ii) Individual Fairness (Dwork et. al., ’12)

Counterfactual fairness (Kusner et. al., 2017)



Counterfactual fairness (Kusner et. al., 2017)

Assumption: ancestral closure of set X.



redlining

religious segregation

rural/urban balance of genders in China

Counterfactual fairness (Kusner et. al., 2017)

Assumption: ancestral closure of set X.

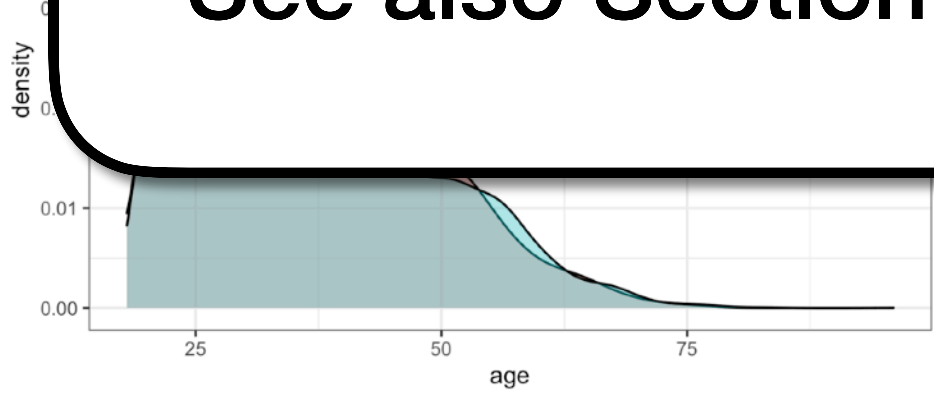
In summary, Counterfactual Fairness is:

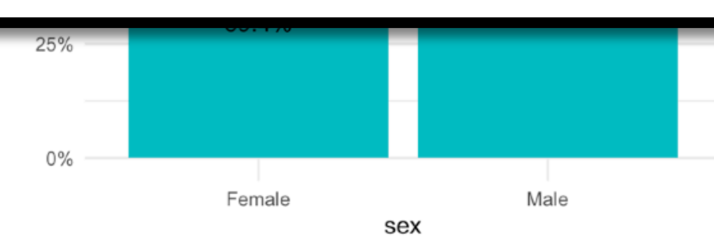
- decomposable & inadmissible (w.r.t DE, IE, SE),
- not identifiable in general, and
- oblivious to spurious effects (and corresponding business necessity requirements).

See also Section 4.4.1 for further details.

gation

lance
China





Individual Fairness
(Dwork. et. al., 2012)

Causal Fairness Analysis implications on IF:

- IF is oblivious to the underlying causal mechanisms.

Example 17
Section 4.4.2

- IF captures the direct effect only under the SFM.

Proposition 5
Section 4.4.2

- IF with a sparse metric d is not admissible.

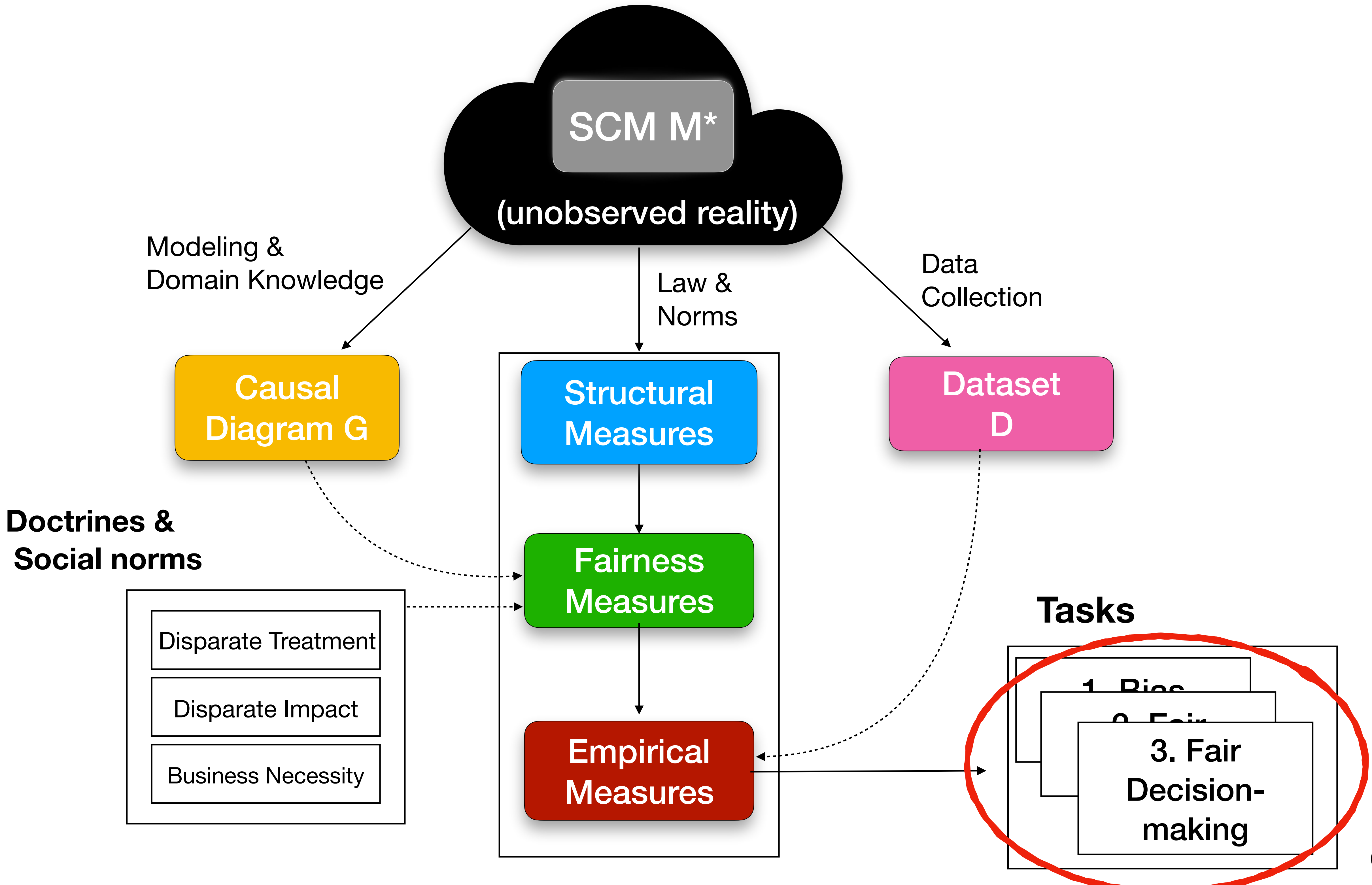
Example 18
Section 4.4.2

- IF with a complete metric d doesn't account for business necessity.

Proposition 11
Section 4.4.2

Part II

Fairness Tasks
(Big Picture)



Task 1.
Bias Detection & Quantification

Fairness Cookbook

Fairness Cookbook

Section 5.1 Algorithm 1

1) Obtain data on past decisions $\mathcal{D}$.

2) Determine the (possibly simplified) causal diagram $\mathcal{G}$ (w.r.t. underlying $\mathcal{M}^*$).

3) Determine the **Business Necessity** (BN) set ($\emptyset$, $\{Z\}$, $\{W\}$, $\{Z, W\}$).

4) Consider existence of **Disparate Treatment:**

$$H_0^{(x\text{-DE})} : x\text{-DE}_{x_0,x_1}(y \mid x_0) = 0.$$

rejected → evidence of disparate treatment (population level)

not rejected → no evidence of disparate treatment (population level)

5) Consider existence of **Disparate Impact:**

5a) Indirect effect:

if($W \notin$ BN-set) $H_0^{(x\text{-IE})} : x\text{-IE}_{x_0,x_1}(y \mid x_0) = 0.$

rejected → evidence of disparate impact

not rejected → go to next step

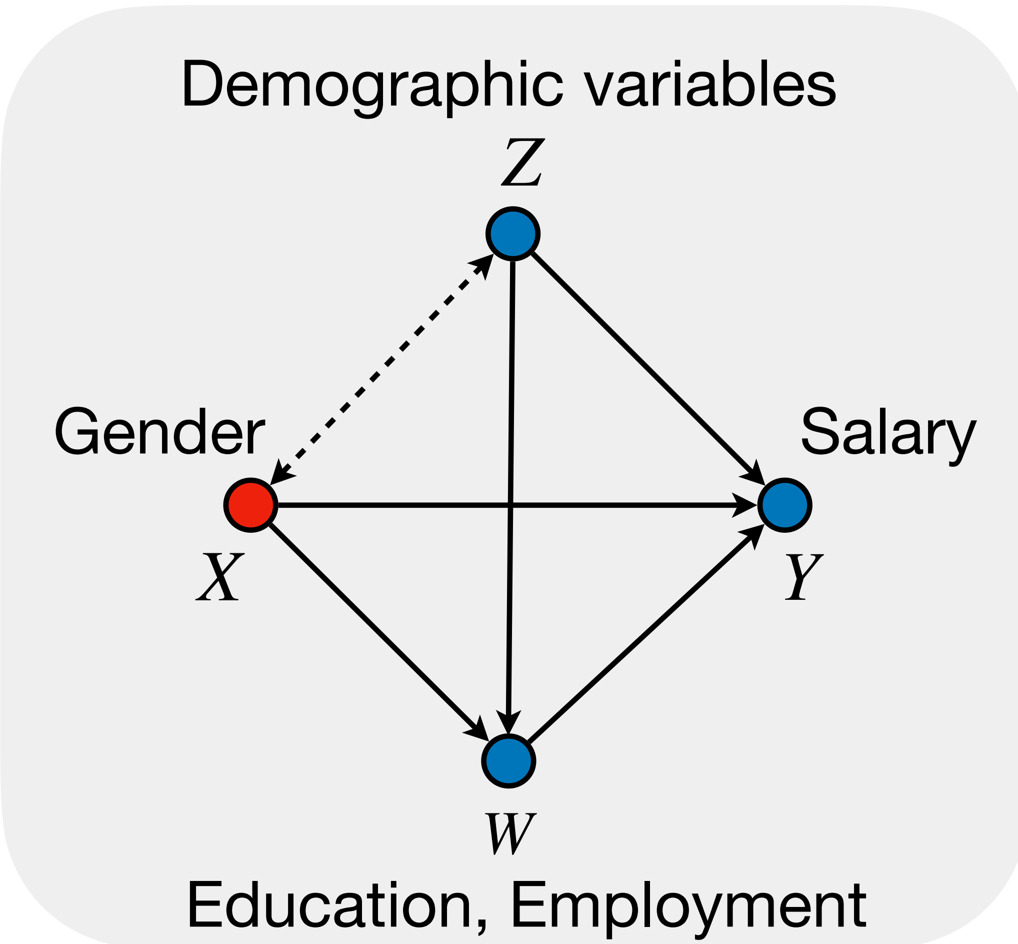
5b) Spurious effect:

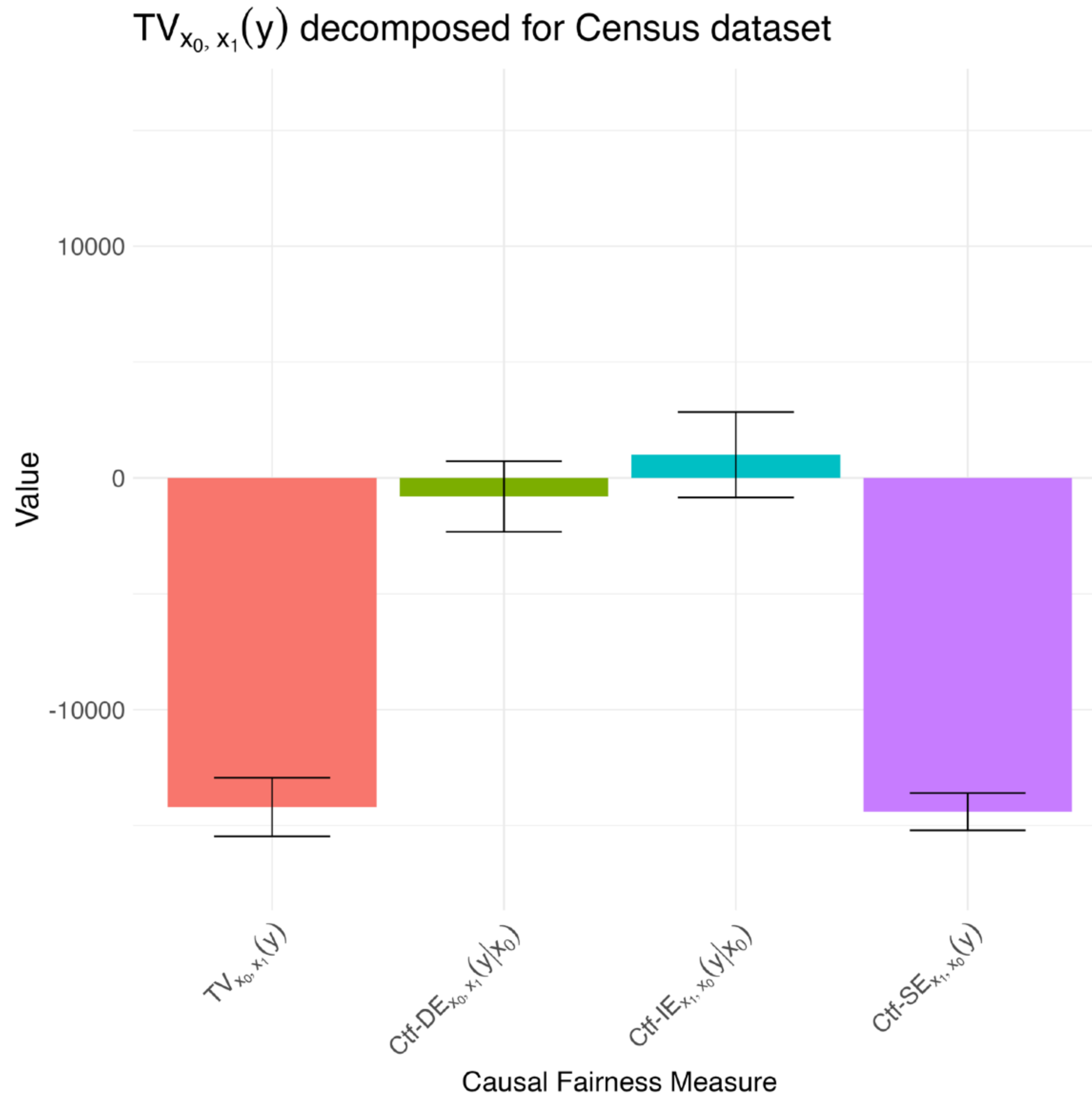
if($Z \notin$ BN-set) $H_0^{(x\text{-SE})} : x\text{-SE}_{x_0,x_1}(y) = 0.$

rejected → evidence of disparate impact

not rejected → no evidence of disparate impact

Task 1: Census 2018 dataset





- Observed disparity: $\text{TV}_{x_0,x_1}(y) = \$14{,}000/\text{year}$

Task 2. Fair Predictions

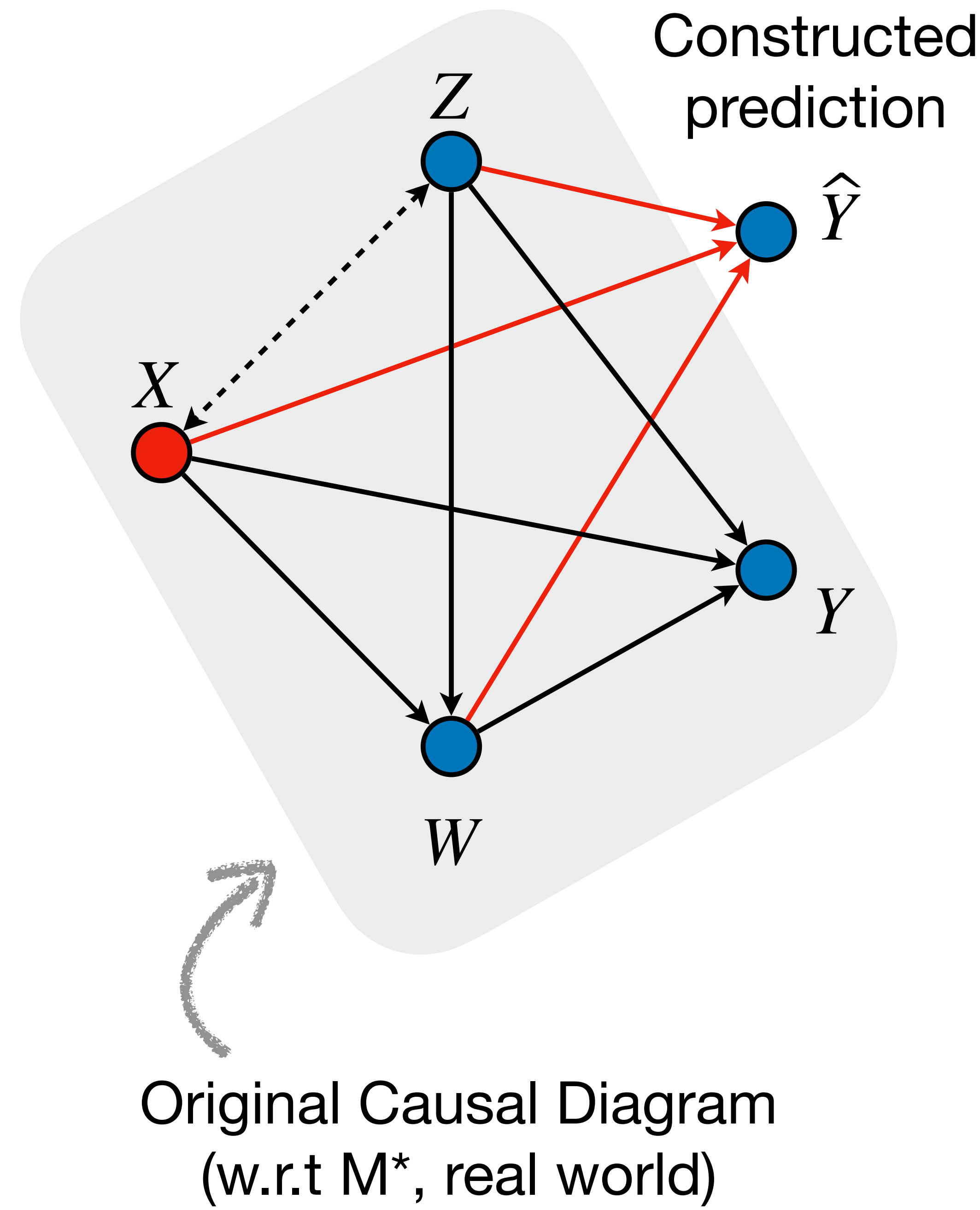
Prediction Task

- The first talk focused on bias detection, where we just analyze the "observed reality", i.e., nature defines f_Y

- When doing prediction, causally speaking, we are constructing a new mechanism $\widehat{Y} \leftarrow f_{\widehat{Y}}(x, z, w)$ that is under our control (i.e., we are selecting it)

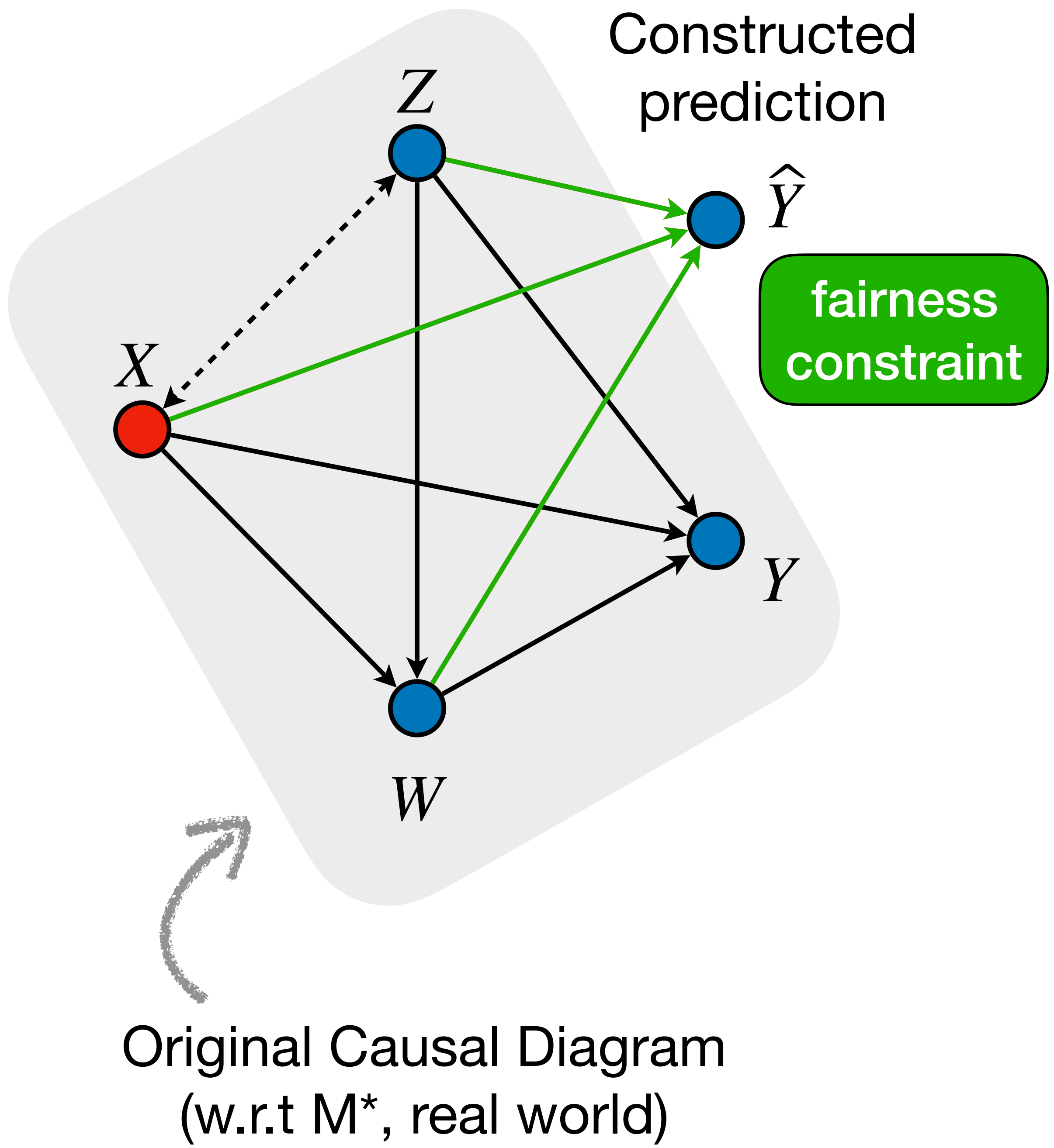
- Typically, in ML, we are simply interested in learning $P(y \mid x, z, w)$

- Does that carry over bias from f_Y?



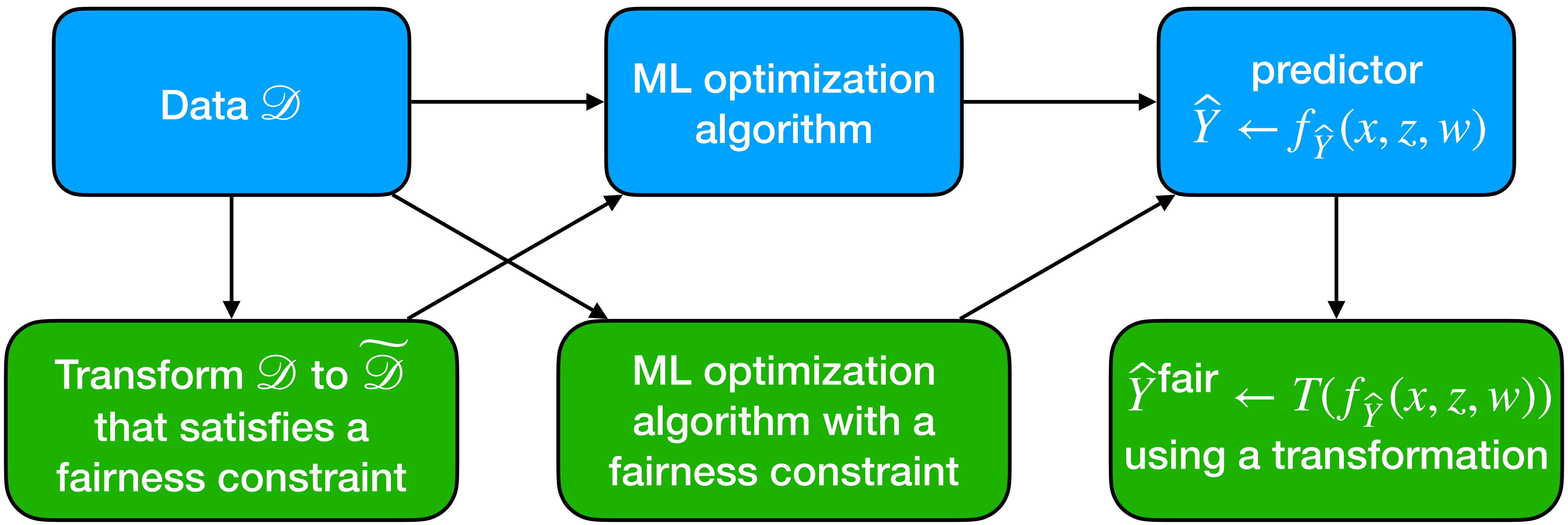
Fair Prediction

- General answer: simply learning $P(y \mid x, z, w)$ will give biased predictions.
- To remove the bias, one might wish for $\widehat{Y}$ to satisfy a pre-specified fairness constraint.
- A commonly considered constraint is to make $\text{TV}_{x_0,x_1}(\widehat{Y}) = 0$.
- In practice, there are different ways to satisfying such a constraint: in particular, we distinguish post-processing, in-processing, and pre-processing methods.



Pre-, In-, Post-Processing

Typical ML framework:



Pre-processing: change the data to satisfy a constraint apriori

In-processing: include a fairness constraint in the learning step

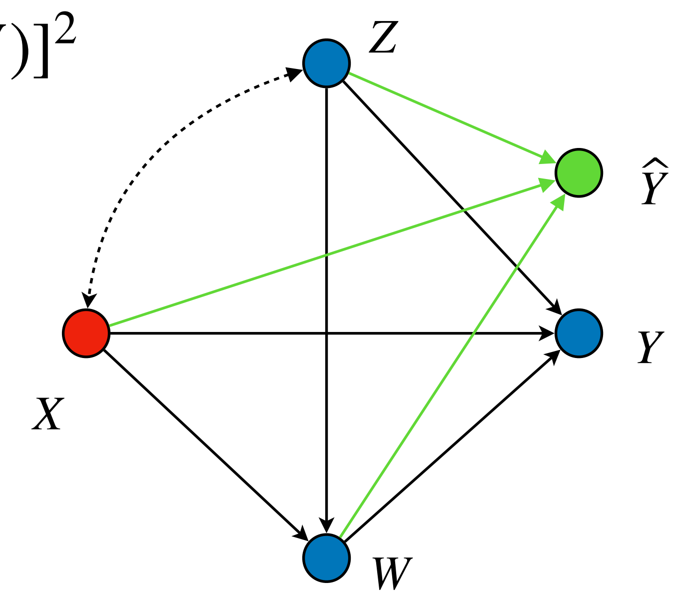
Post-processing: massage the predictions to satisfy a constraint

Fair Prediction Theorem (FPT)

Theorem. Let SFM(n_Z, n_W) be the SFM with $|Z| = n_Z$ and $|W| = n_W$. Let E denote the set of edges of SFM (n_Z, n_W). Further, let $\mathcal{S}^{\text{linear}}_{n_Z,n_W}$ be the space of linear SCMs (but for the variable X, which is a Bernoulli) compatible with the SFM(n_Z, n_W) and whose structural coefficients are drawn uniformly from$[-1,1]^{|E|}$.

An SCM $M \in \mathcal{S}^{\text{linear}}_{n_Z,n_W}$ is said to be ϵ-TV-compliant if

$$\hat{f}_{\text{fair}} = \text{argmin}_{f \text{ linear}} \quad E[Y - f(X, Z, W)]^2$$
$$\text{subject to} \quad TV_{x_0,x_1}(f) = 0$$



also satifies

$$|\text{Ctf-DE}_{x_0,x_1}(\hat{f}_{\text{fair}} \mid x_0)| \leq \epsilon,$$
$$|\text{Ctf-IE}_{x_0,x_1}(\hat{f}_{\text{fair}} \mid x_0)| \leq \epsilon,$$
$$|\text{Ctf-SE}_{x_0,x_1}(\hat{f}_{\text{fair}})| \leq \epsilon\,.$$

Under the Lebesgue measure ... SF ... e 0.

Furthermore, for any n_Z, n_W t ... TV-

non-vanishing probability of things "going wrong"

Section 5.2 Theorem 10

FPT proof sketch

Objective:

$$Y = \sum_{V_i \in X,Z,W} a_{V_iY}V_i + \epsilon_Y, \quad f(X,Z,W) = \sum_{V_i \in X,Z,W} \tilde{a}_{V_iY}V_i.$$

$$E[Y - f(X,Z,W)]^2 = E\Big[\sum_{V_i \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY})V_i + \epsilon_Y\Big]^2$$

$$= E[\epsilon_Y^2] + E\Big[\sum_{V_i,V_j \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY})(a_{V_jY} - \widetilde{a}_{V_jY})V_iV_j\Big]$$

$$= 1 + (a_{VY} - \widetilde{a}_{VY})^T E[VV^T](a_{VY} - \widetilde{a}_{VY}),$$

optimizing over $\widetilde{a}_{VY}$

ellipsoid

$$TV_{x_0,x_1}(f) = (E[V \mid x_1] - E[V \mid x_0])^T \widetilde{a}_{VY} = 0.$$

what the constraint is

$$Ctf\text{-}DE = \widetilde{a}_{XY}(x_1 - x_0) = 0,$$

$$Ctf\text{-}IE = \sum_{W_i} \widetilde{a}_{W_iY}(E[W_i \mid x_1] - E[W_{i_{x_0}} \mid x_1]) = 0,$$

$$Ctf\text{-}SE = \sum_{W_i} \widetilde{a}_{W_iY}(E[W_{i_{x_0}} \mid x_1] - E[W_i \mid x_0]) +$$

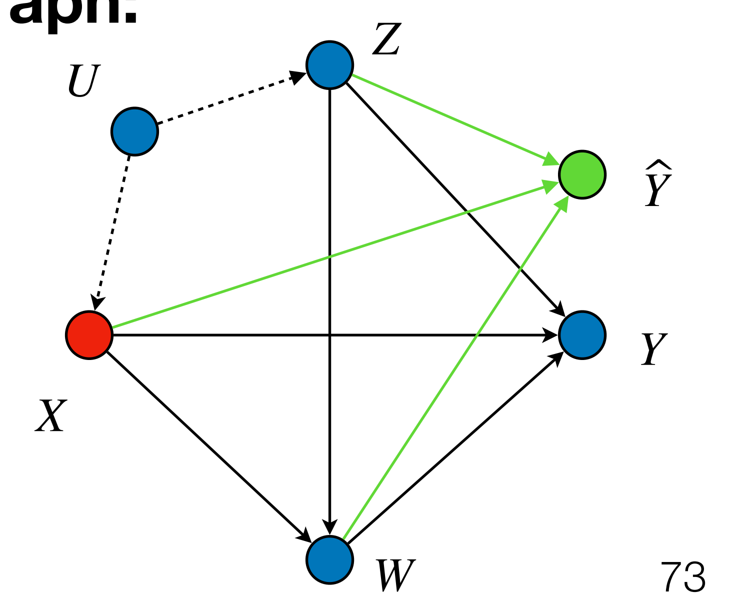
$$\sum_{Z_i} \widetilde{a}_{Z_iY}(E[Z_i \mid x_1] - E[Z_i \mid x_0]) = 0.$$

what we actually want

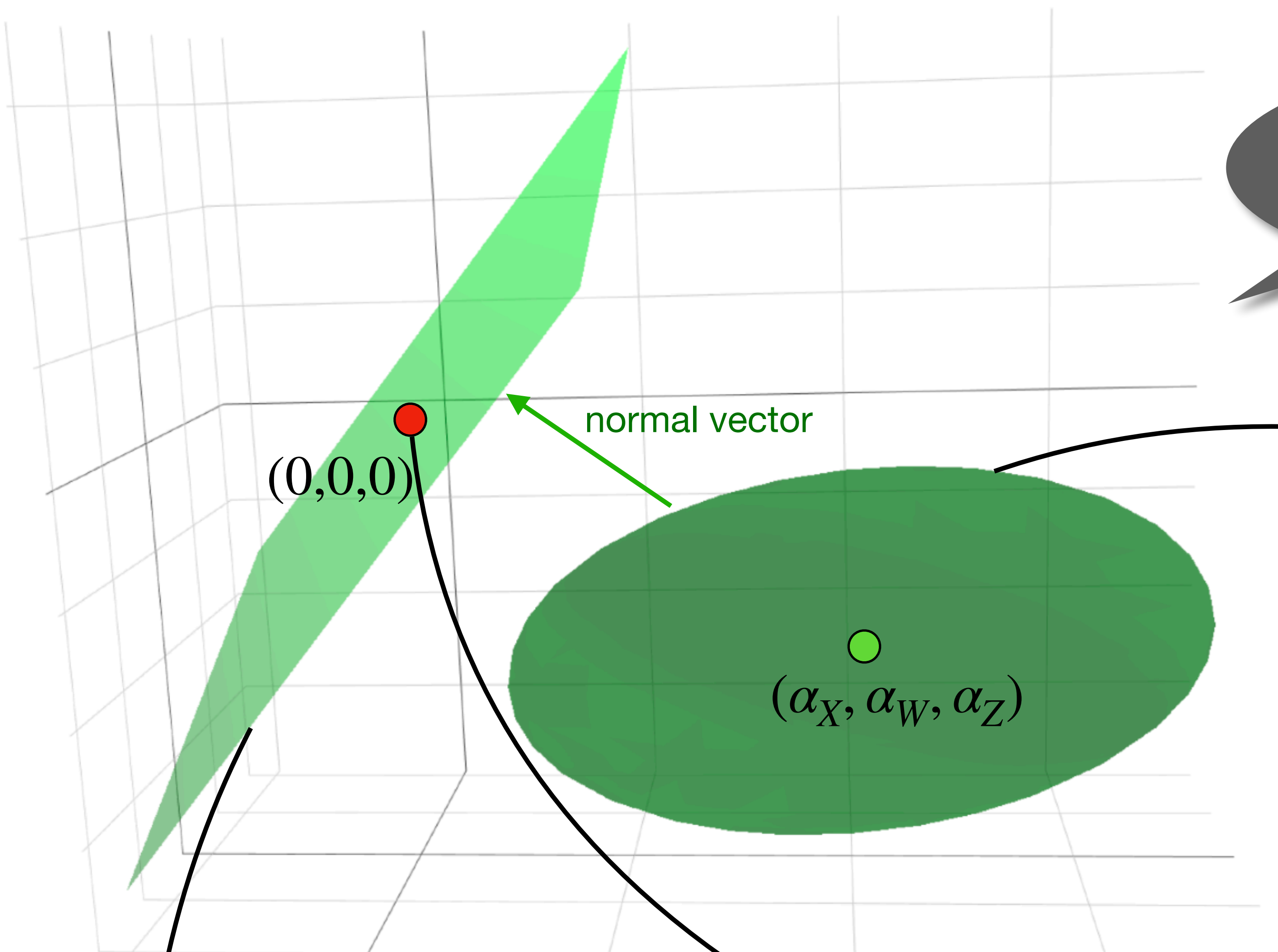
Linear SCM:

$$U \leftarrow N(0,1)$$
$$X \leftarrow \textbf{Bernoulli}(\textbf{expit}(U))$$
$$Z \leftarrow a_{UZ}U + a_{ZZ}Z\epsilon_Z$$
$$W \leftarrow a_{XW}X + a_{ZW}Z + a_{WW}W + \epsilon_W$$
$$Y \leftarrow a_{XY}X + a_{ZY}Z + a_{WY}W + \epsilon_Y$$

Graph:



FPT visualization



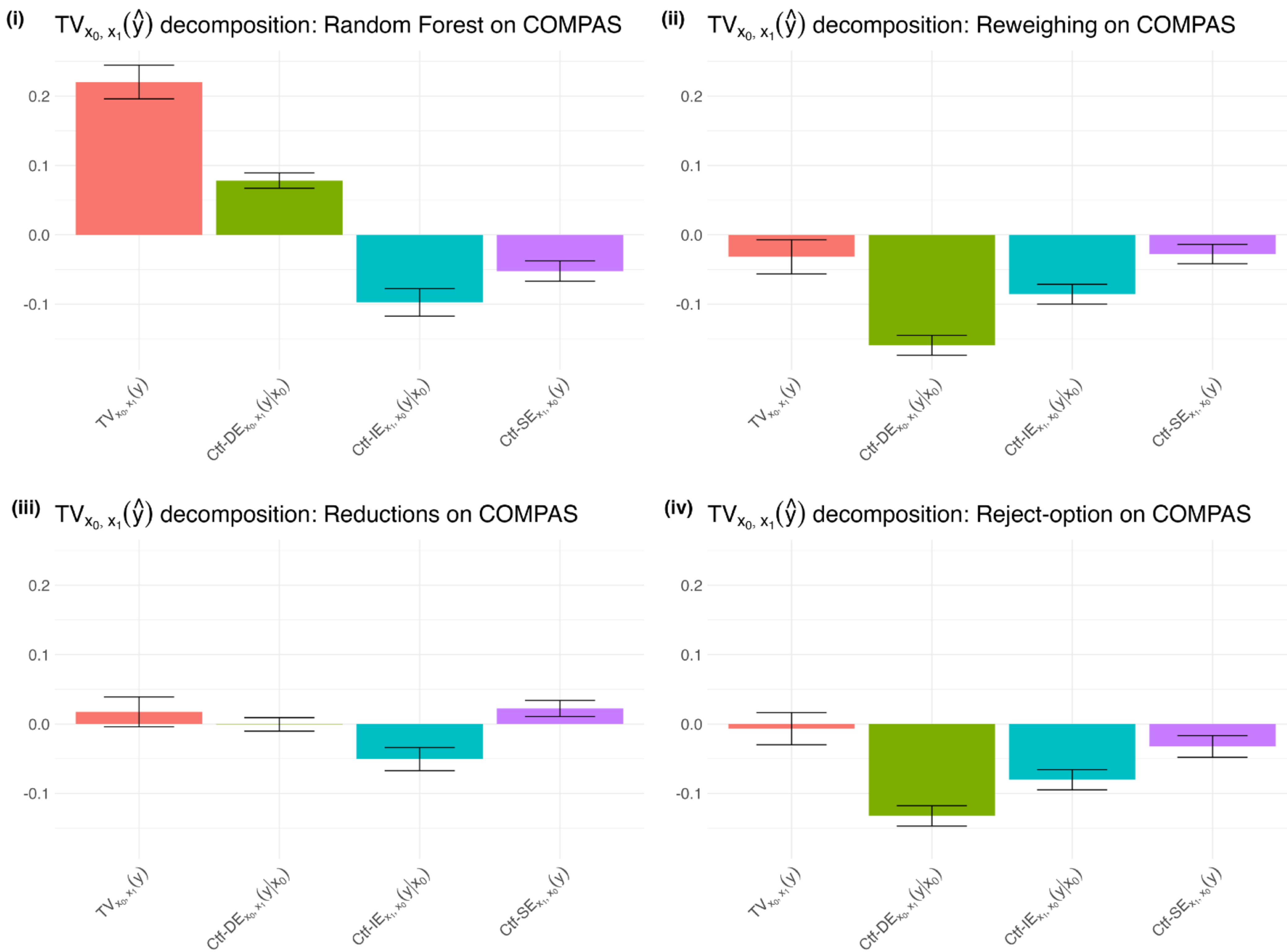
measuring the probability of ellipsoid hitting a point

ellipsoid with characteristic matrix $\Sigma = E[VV^T]$

What the constraint is:
$\text{TV}_{x_0,x_1}(\widehat{y}) = 0$
represents a hyperplane through origin.

What we want:
3 linear constraints
Ctf-DE $= 0$, Ctf-IE $= 0$, Ctf-SE $= 0$.
represents a single point

Fair Prediction Theorem in Practice (COMPAS dataset)



Failure of Optimal Transport (in the Individual Fairness framework)

- A common approach for pre-processing is to use optimal transport
- The distribution $P(V \mid x_1)$ is transported onto $P(V \mid x_0)$

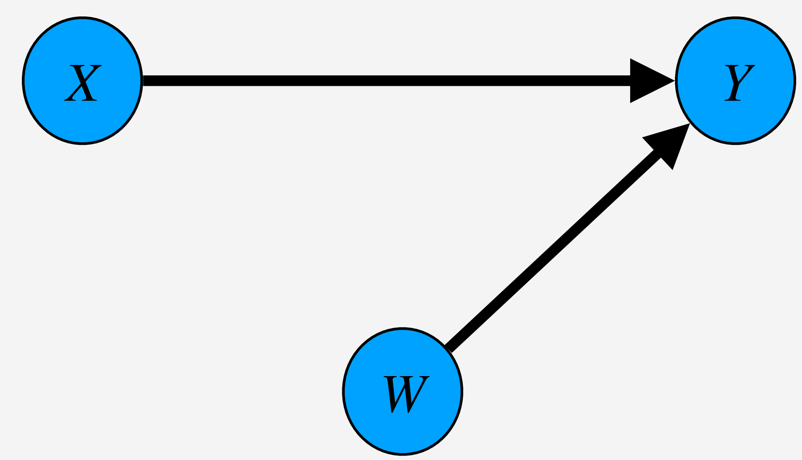
Example.

$$X \leftarrow U_X$$

$$W \leftarrow \epsilon(2U_W - 1)$$

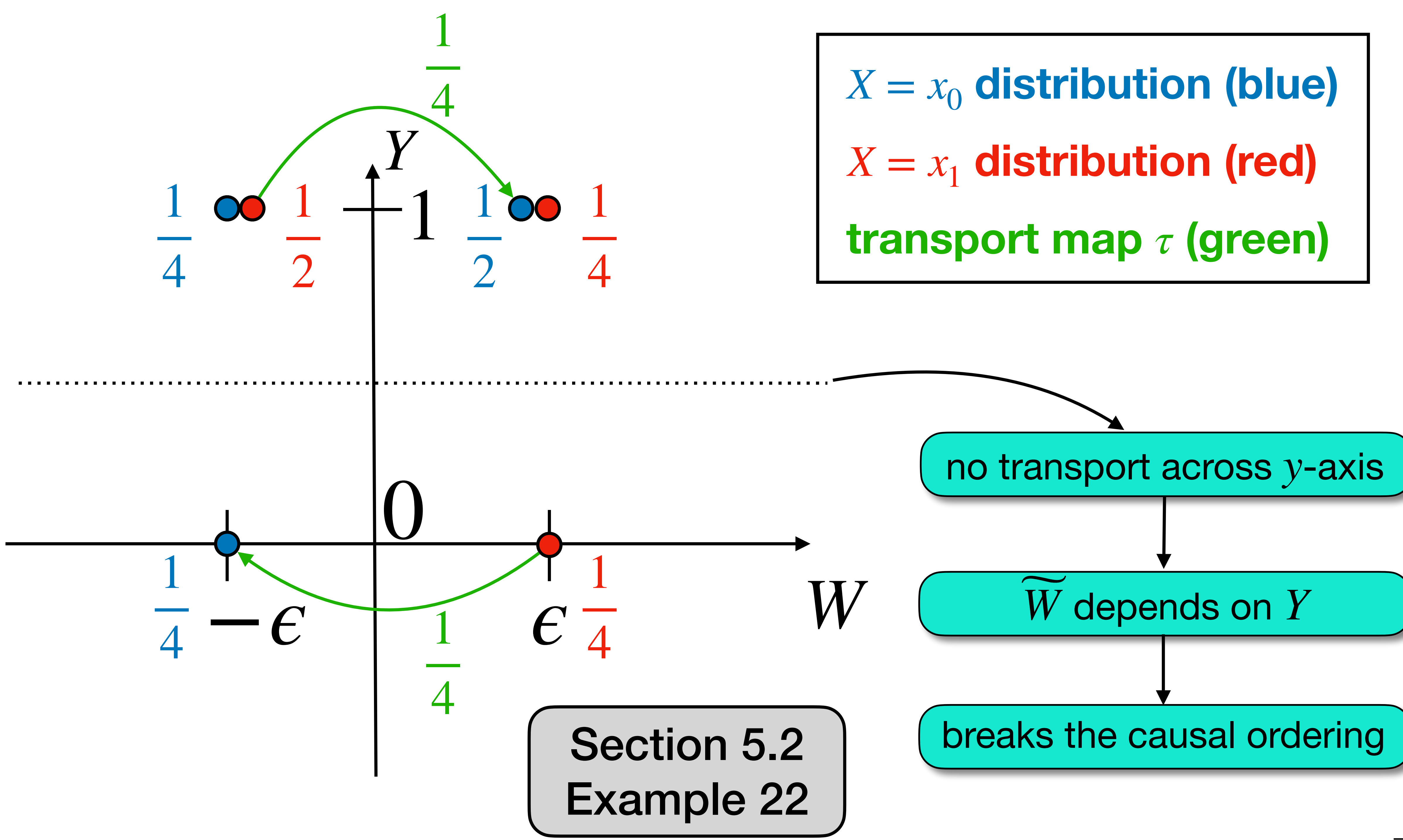
$$Y \leftarrow \begin{cases} U_Y \vee 1(W > 0) \textbf{ if } X = x_0 \\ U_Y \vee 1(W < 0) \textbf{ if } X = x_1 \end{cases}$$

U_X, U_W, U_Y **Bernoulli(0.5)**



- In the example, we wish to compute $\text{NIE}_{x_0,x_1}(\widetilde{y}) = P(\widetilde{y}_{x_0,\widetilde{W}_{x_1}}) - P(\widetilde{y}_{x_0})$

Failure of Optimal Transport (in the Individual Fairness framework)



Failure of Optimal Transport (in the Individual Fairness framework)

$$P(\widetilde{y}_{x_0,\widetilde{W}_{x_1}}) = P(\widetilde{y}_{x_0,\epsilon}, \widetilde{W}_{x_1} = \epsilon) + P(\widetilde{y}_{x_0,-\epsilon}, \widetilde{W}_{x_1} = -\epsilon) \quad - \quad P(\widetilde{y}_{x_0}) = P(y_{x_0})$$

using the SCM

$\widetilde{y}_{x_0,\epsilon} = 1$ for any u

$\widetilde{W}_{x_1} = \epsilon$ for $U_W = 1$ w.p. $\frac{1}{2}$

$U_W = 0$ w.p. $\frac{1}{2}$ (1/4 for each U_Y)

$y_{x_0,-\epsilon} = U_Y$

for $U_Y = 1$, $\widetilde{W}_{x_1} = -\epsilon$

with prob. $\frac{1}{4}$ (0 for $U_W = 1$)

$y_{x_0} = U_Y \vee 1(W > 0)$

for $U_Y = 1$, $y_{x_0} = 1$

for $U_Y = 0$, $y_{x_0} = 1$

with prob. $\frac{1}{2}$

putting together

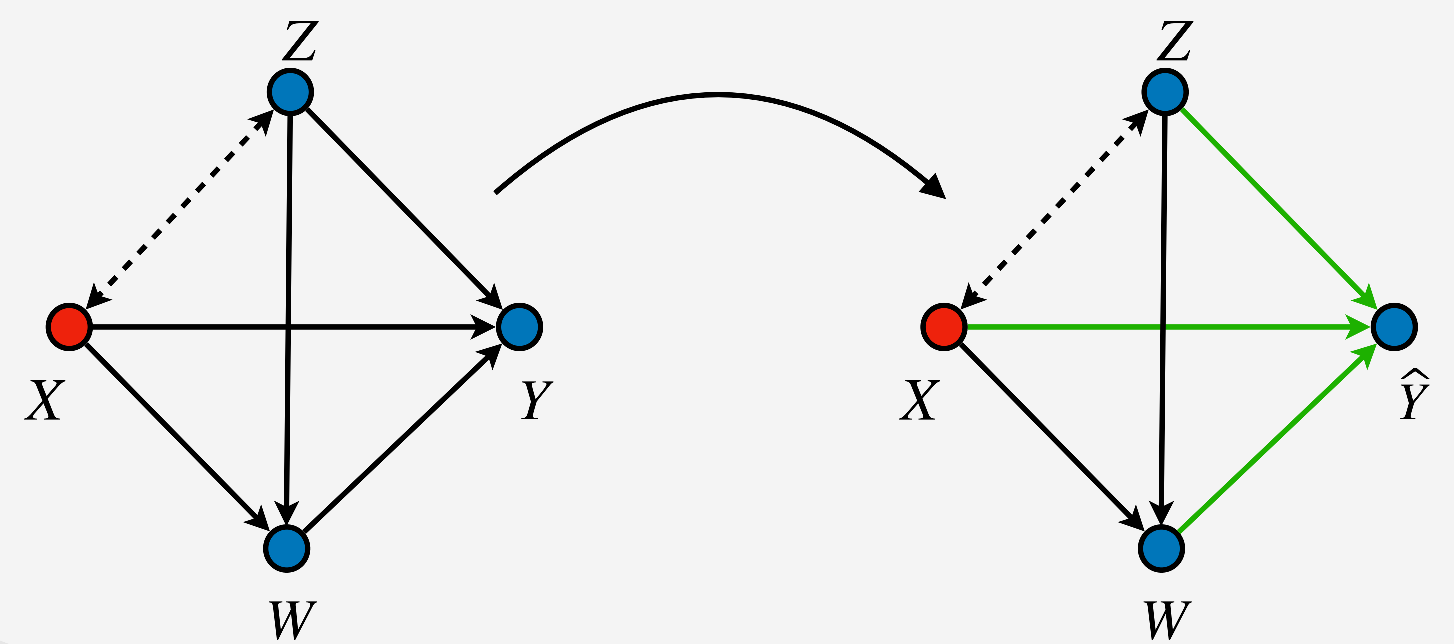
$$\boxed{P(\widetilde{y}_{x_0,\widetilde{W}_{x_1}}) - P(\widetilde{y}_{x_0})} = \frac{1}{2} + \frac{1}{8} - \frac{3}{4} = -\frac{1}{8} \implies$$

Indirect Effect $\neq 0$!

Towards the solution

- how can we construct "causal" fair predictions?

(i) causal structure of the SFM is preserved for the predictor $\widehat{Y}$



Section 5.2.8

(ii) identification expressions for x-DE, x-SE, and x-IE equal 0 for the predictor $\widehat{Y}$

$$x\text{-}\mathbf{DE}^{ID}_{x_0,x_1}(\hat{y}) = \sum_{z,w} [P(\hat{y} \mid x_1, z, w) - P(\hat{y} \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x_0) = 0$$

$$x\text{-}\mathbf{IE}^{ID}_{x_0,x_1}(\hat{y}) = \sum_{z,w} P(\hat{y} \mid x_1, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x) = 0$$

$$x\text{-}\mathbf{SE}^{ID}_{x_1,x_0}(\hat{y}) = \sum_{z} P(\hat{y} \mid x_1, z)[P(z \mid x_1) - P(z \mid x_0)] = 0.$$

In-processing solution

Theorem. Let $\widehat{Y}$ be the solution to the following optimization problem:

$$\widehat{Y} = \textbf{argmin}_f \quad E[Y - f(X, Z, W)]^2$$

$$\textbf{subject to} \quad x\text{-}\textbf{DE}^{\textbf{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0$$

$$x\text{-}\textbf{DE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0$$

$$x\text{-}\textbf{IE}^{\textbf{ID}}_{x_0,x_1}(\widehat{y} \mid x_0) = 0$$

$$x\text{-}\textbf{IE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y} \mid x_0) = 0$$

$$x\text{-}\textbf{SE}^{\textbf{ID}}_{x_1,x_0}(\widehat{y}) = 0$$

Section 5.2.9

Then $\widehat{Y}$ satisfies

$$x\text{-DE}_{x_0,x_1}(\widehat{y} \mid x_0) = x\text{-IE}_{x_1,x_0}(\widehat{y} \mid x_0) = x\text{-SE}_{x_1,x_0}(\widehat{y}) = 0.$$

Pre-processing solution (Causal IF)

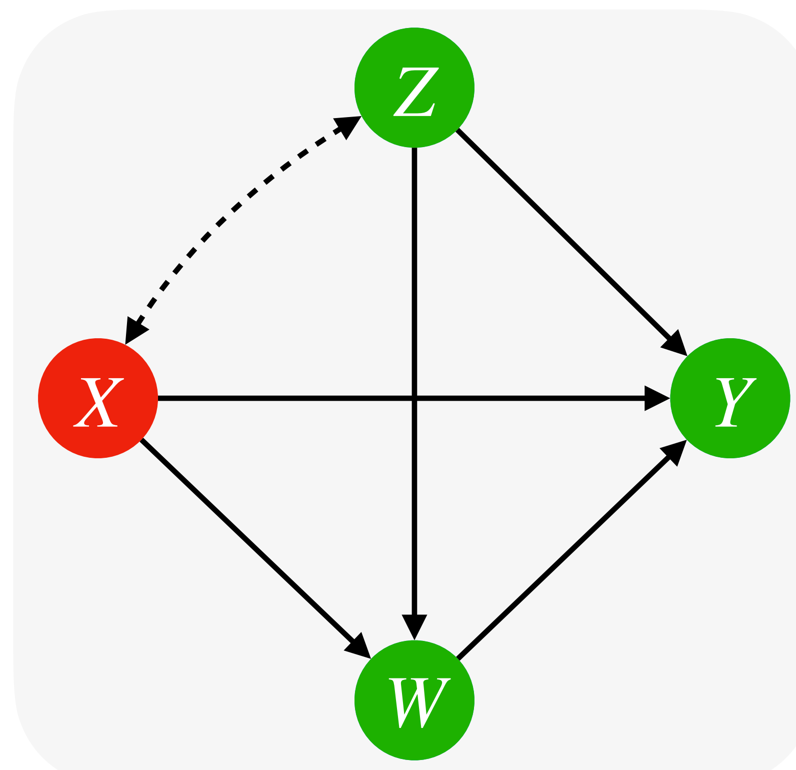
Definition. The Causal Individual Fairness (Causal IF, for short) algorithm is performed on a data coming from an SCM $\mathscr{M}$ compatible with the standard fairness model (SFM), in the following way:

1) if $Z \notin$ BN-set, transport
$Z \mid x_1 \mapsto Z \mid x_0$

2) if $W \notin$ BN-set, transport
$W \mid x_1, Z = z \mapsto W \mid x_0, Z = z$

3) transport
$Y \mid x_1, Z = z, W = w \mapsto Y \mid x_0, Z = z, W = w$

SFM



Data $\mathscr{D}$

X	Z	W	Y
$x^{(1)}$	$\tilde{z}^{(1)}$	$\tilde{w}^{(1)}$	$\tilde{y}^{(1)}$
$x^{(1)}$	$\tilde{z}^{(2)}$	$\tilde{w}^{(2)}$	$\tilde{y}^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x^{(n)}$	$\tilde{z}^{(n)}$	$\tilde{w}^{(n)}$	$\tilde{y}^{(n)}$

Section 5.2
Theorem 11

Pre-processing solution (Causal IF)

Theorem. Let $\mathscr{M}$ be an SCM compatible with the SFM. Let τ be the optimal transport map obtained when applying Causal IF. Define a new, additional mechanism of the SCM $\mathscr{M}$ such that

$$\widetilde{Y} \leftarrow \tau^Y(Y; X, Z, W)\,.$$

For the transformed outcome $\widetilde{Y}$ we can then claim:

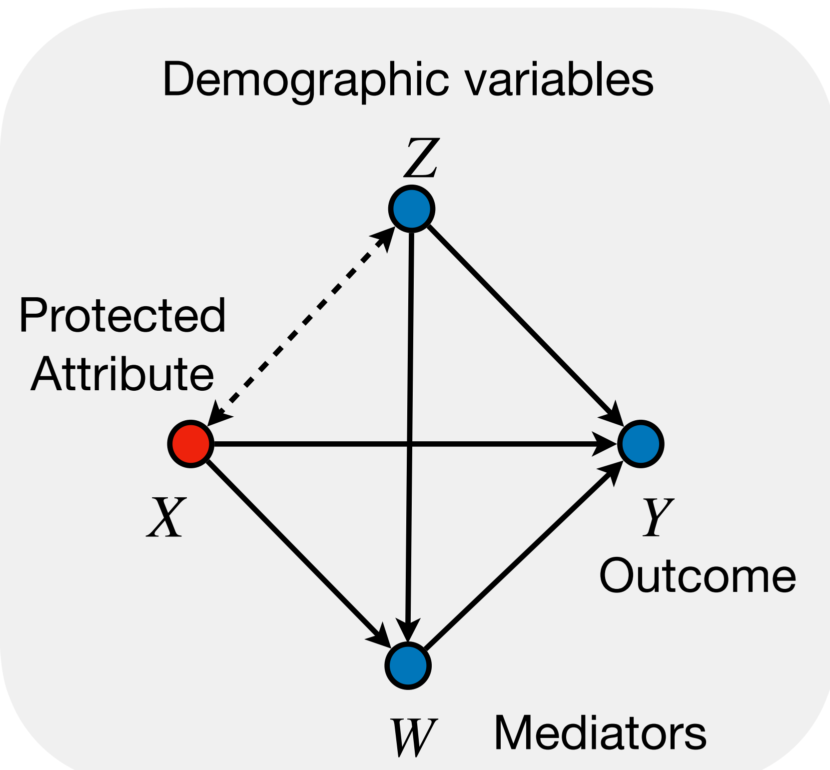
$$\text{if } Z \notin \text{BN-set} \implies x\text{-SE}_{x_1,x_0}(\widetilde{y}) = 0.$$
$$\text{if } W \notin \text{BN-set} \implies x\text{-IE}_{x_1,x_0}(\widetilde{y} \mid x_0) = 0.$$

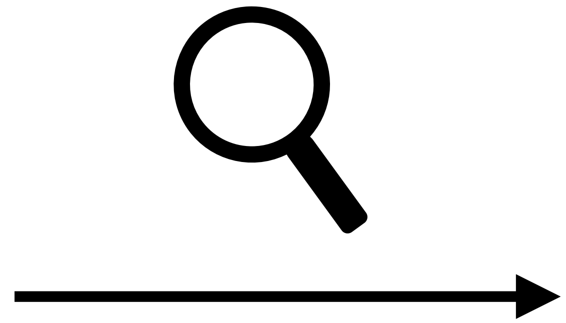
Furthermore, the transformed outcome $\widetilde{Y}$ also satisfies

$$x\text{-DE}_{x_0,x_1}(\widetilde{y} \mid x_0) = 0.$$

Moving beyond SFM

SFM

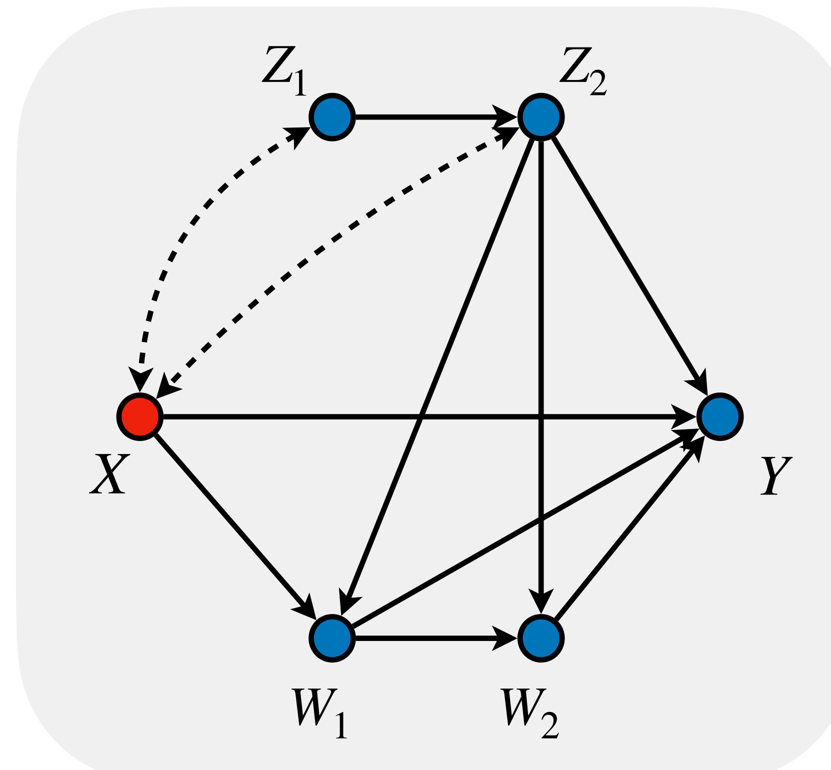




better resolution

Section 6 TBD

Diagram $\mathscr{G}$



Measures	direct, indirect, spurious
Business Necessity	$\{\{\varnothing\}, \{Z\}, \{W\}, \{Z, W\}\}$
Fair Prediction	Causal IF

Measures	variable specific
Business Necessity	any $V' \subseteq V$
Fair Prediction	fairadapt

Conclusions

- Well-founded disparate treatment and impact claims require the plaintiff to establish a causal connection between a defendant's policy and the statistical disparities found in the observed data.

  SCOTUS: No fairness claim can be made without solid causal underpinnings.

- We introduced a framework for fairness analysis based on causal inference to support such claims. In particular, we showed

  A. - how the total variation can be decomposed into variations that can be easily associated with the underlying causal mechanisms, and mapped to disparate impact and disparate treatment doctrines.

  B. - how the developed foundations of Causal Fairness Analysis can be applied in practice, in the context of bias detection and fair prediction.

- We hope these results can help towards the development of the next generation of AI systems to be more fair, accountable, and transparent.

Thank you!

References: Causal Fairness

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. & Wallach, H. (2018), A reductions approach to fair classification, *in* 'International Conference on Machine Learning', PMLR, pp. 60–69.

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 507-556).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, *12*, 317-337.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, *30*.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1), 3-3.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2001), Direct and indirect effects, *In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*.

Pearl, J. & Mackenzie, D. (2018), *The Book of Why: The New Science of Cause and Effect*, 1st edn, Basic Books, Inc., New York, NY, USA.

Plečko, D., & Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *The Journal of Machine Learning Research*, *21*(1), 9776-9819.

Rutherglen, G. (1987). Disparate impact under title VII: an objective theory of discrimination. *Va. L. Rev.*, *73*, 1297.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In *International conference on machine learning* (pp. 325-333). PMLR.

Zhang, J., & Bareinboim, E. (2018, April). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Plečko, D., & Bareinboim, E. (2022). Causal Fairness Analysis. *Columbia University Technical Report R-90, CausalAI Lab*.