Causal Fairness Analysis

(Causal Inference II - Lecture 2)

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

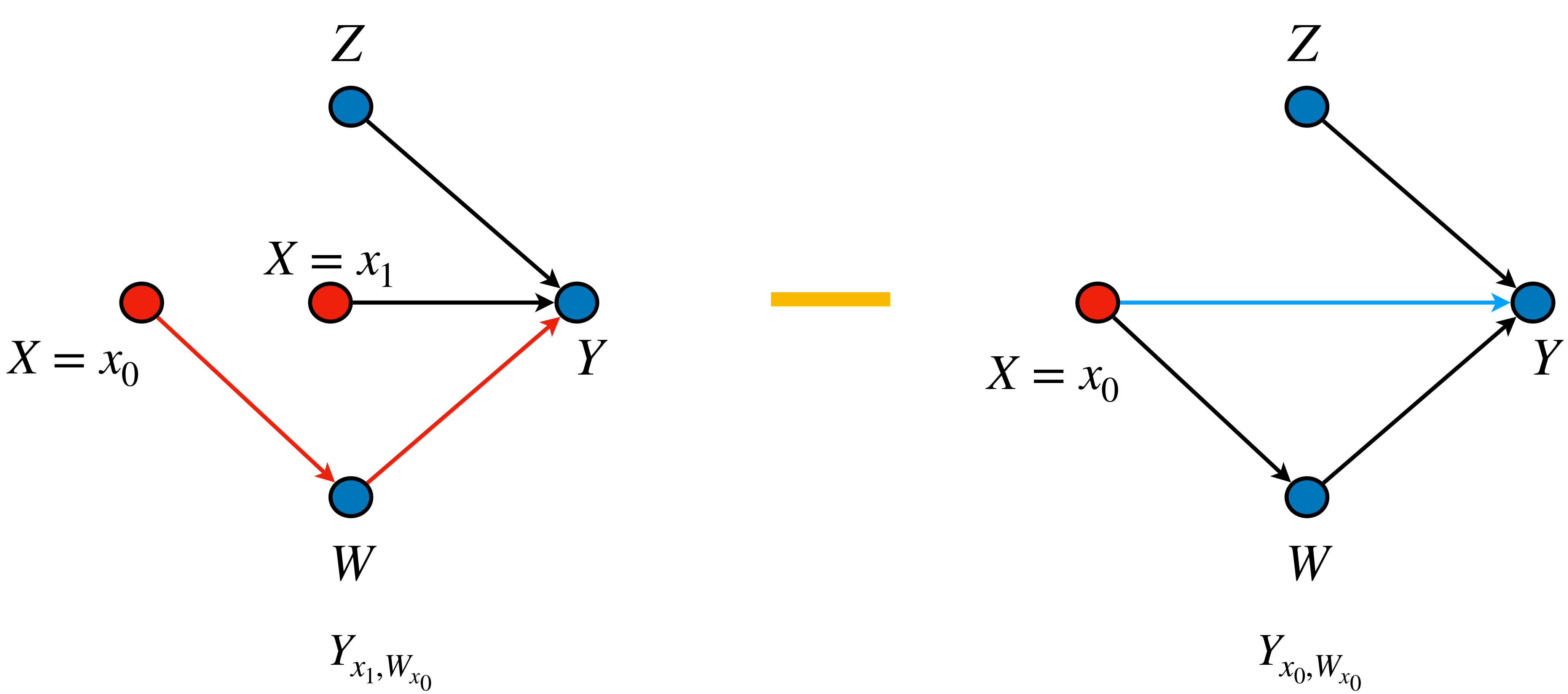
TV family of causal fairness measures

Section 4

Gedankenexperiment (NDE)

- For an individual assigned to male ($X = x_0$) by intervention, how would his salary (Y) change had he been assigned female ($X = x_1$), while keeping the age, nationality, education and employment status unchanged (at the natural level $X = x_0$)?

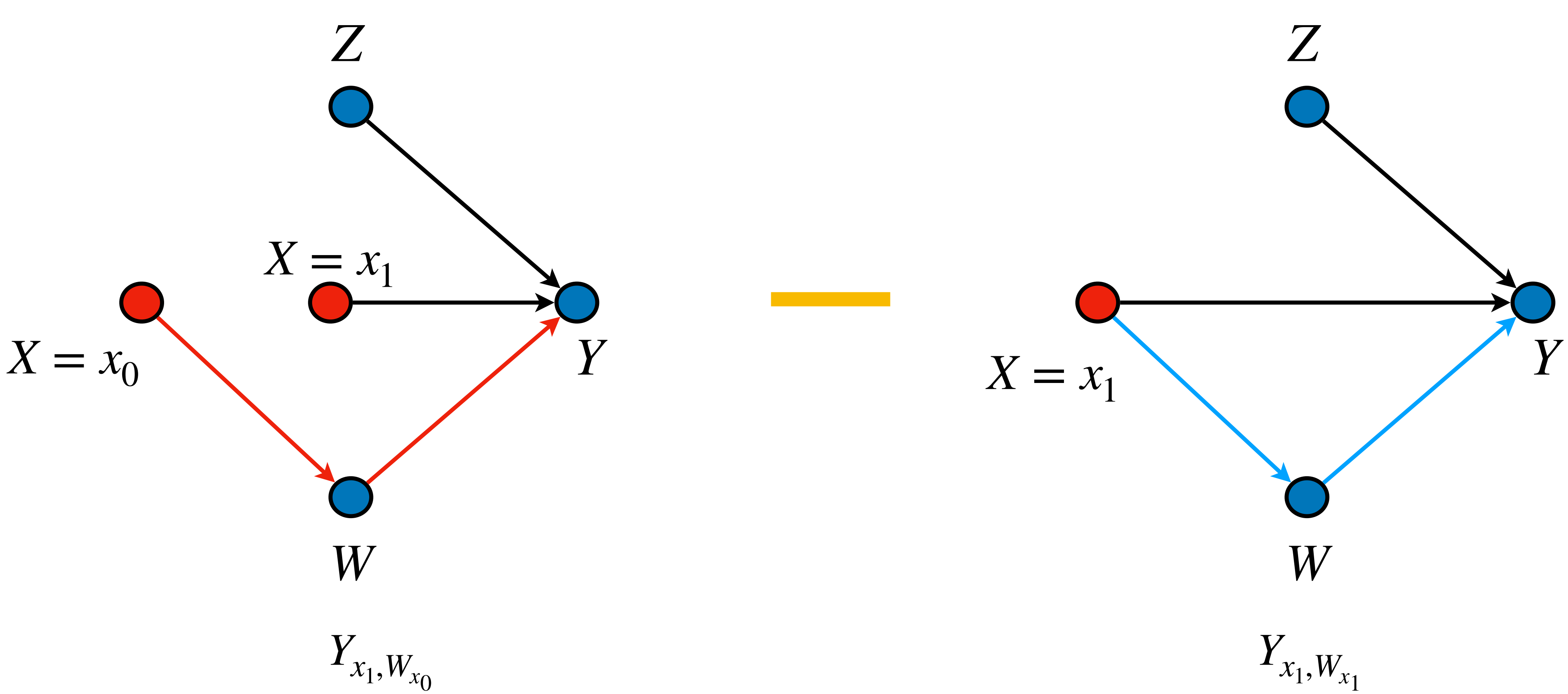
$$\textbf{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})$$



Gedankenexperiment (NIE)

- For an individual assigned to be female ($X = x_1$) by intervention, how would her salary (Y) change had she been assigned to be male ($X = x_0$), while keeping gender unchanged along the direct causal pathway (at the natural level $X = x_1$)?

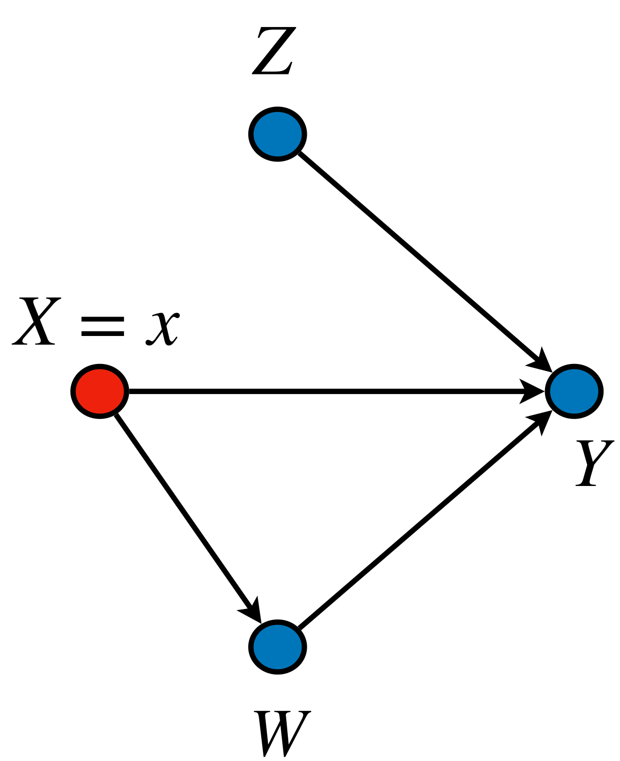
$$\textbf{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1,W_{x_1}})$$



Gedankenexperiment (Exp-SE)

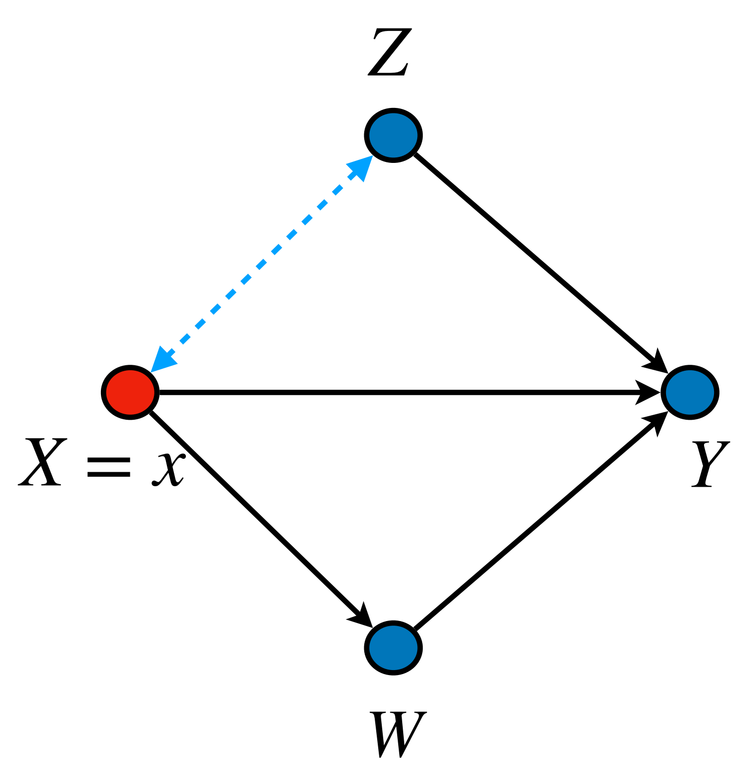
- How would an individuals salary (Y) change if their gender is set to male (or female) by intervention, compared to observing their salary as male (female)?

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

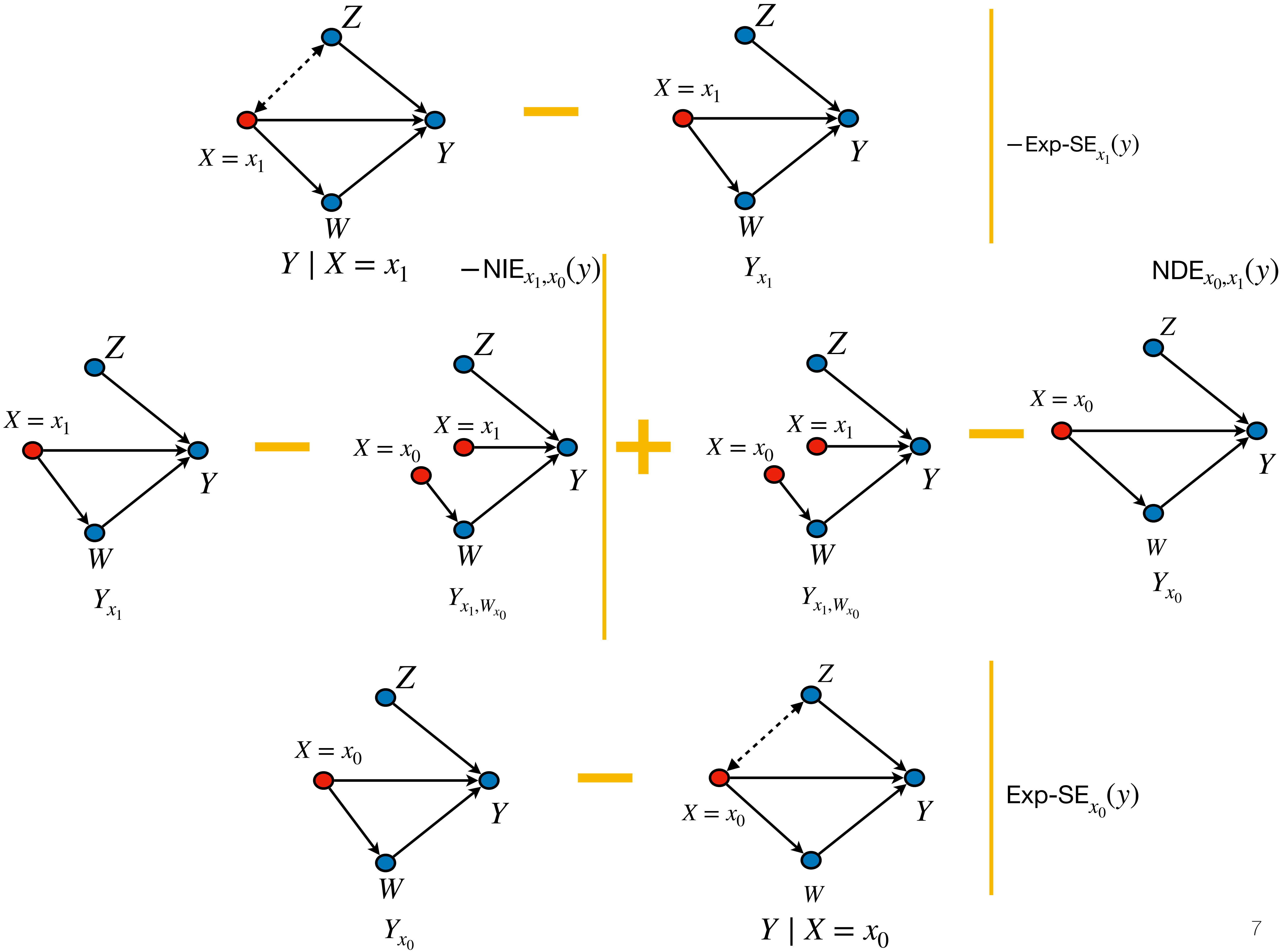


Y_x

—

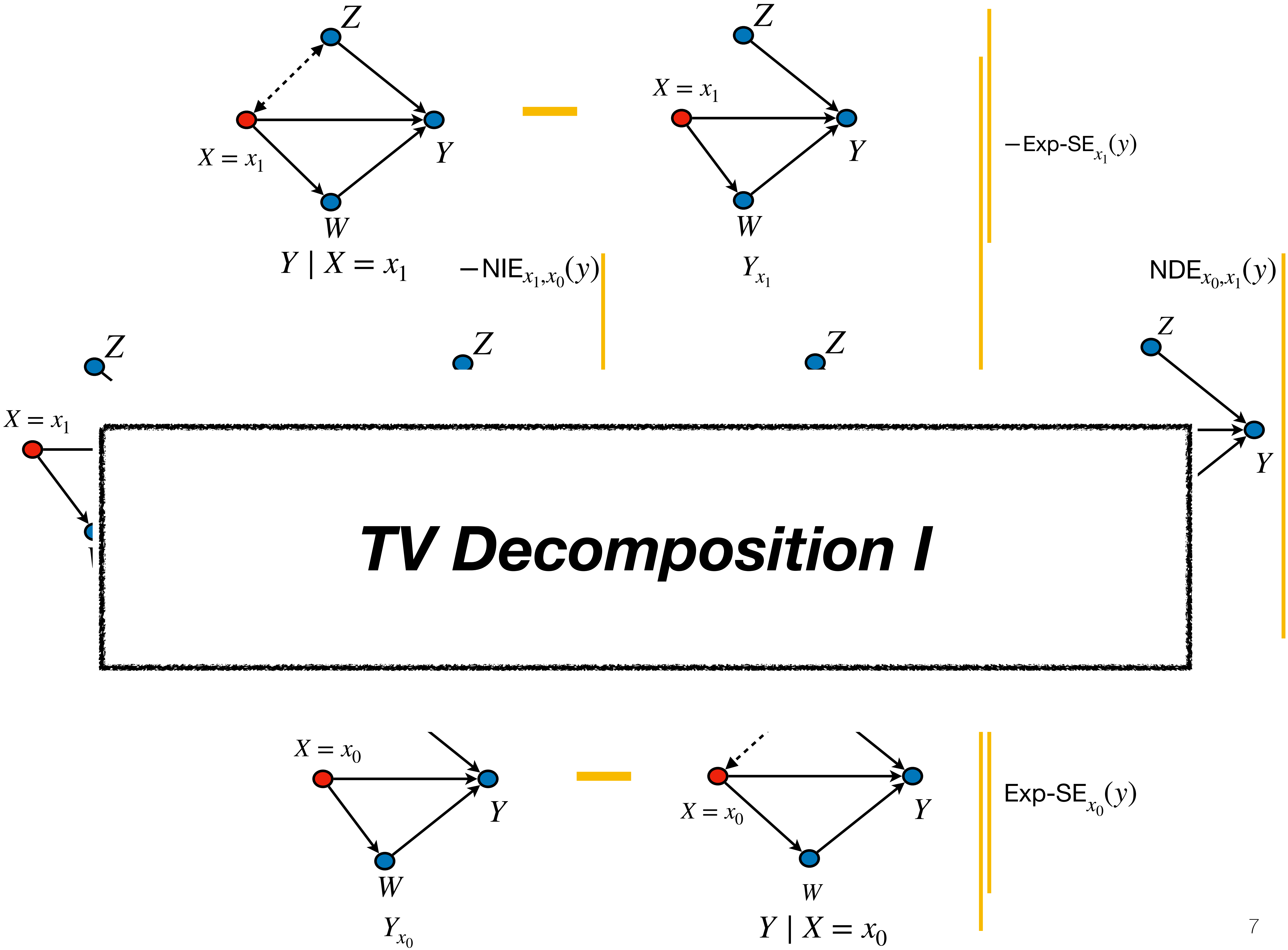


$Y \mid X = x$



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

TV Decomposition I



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:*

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

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,

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

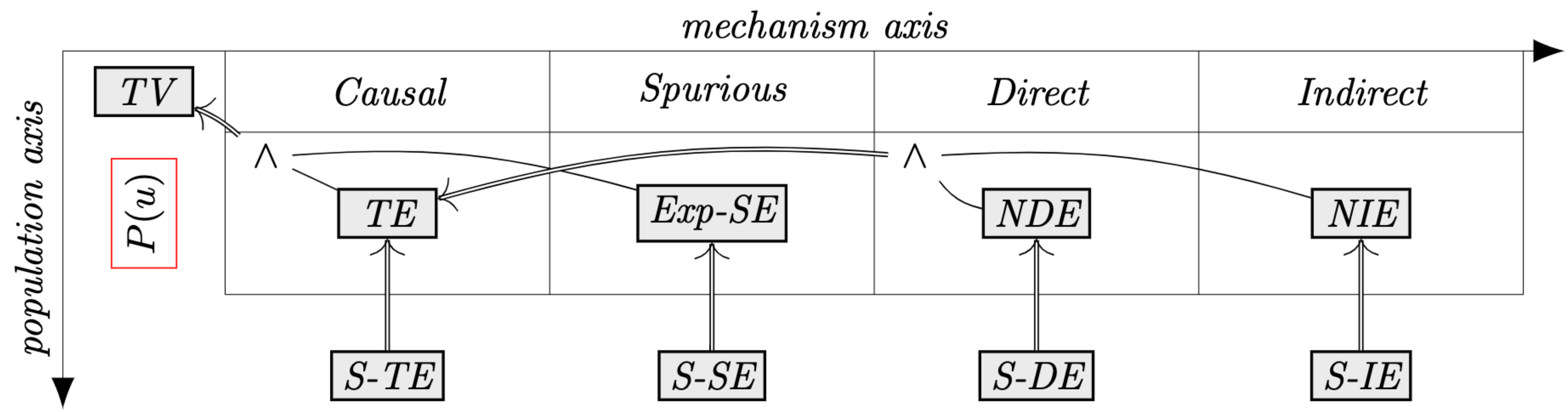
A similar approach can be used for the NIE and Exp-SE since

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

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

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

Fairness Map (prelim version)



- The map is constructed based on the Corollary in the previous page
- We have found fairness measures that are (i) computable from the data; (ii) admissible with respect to structural fairness; (iii) satisfy decomposability with respect to TV; ✓

Section 4.2
Figure 4.2

Does that mean we are done with Causal Fairness Analysis?

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 4.2
Example 4.1

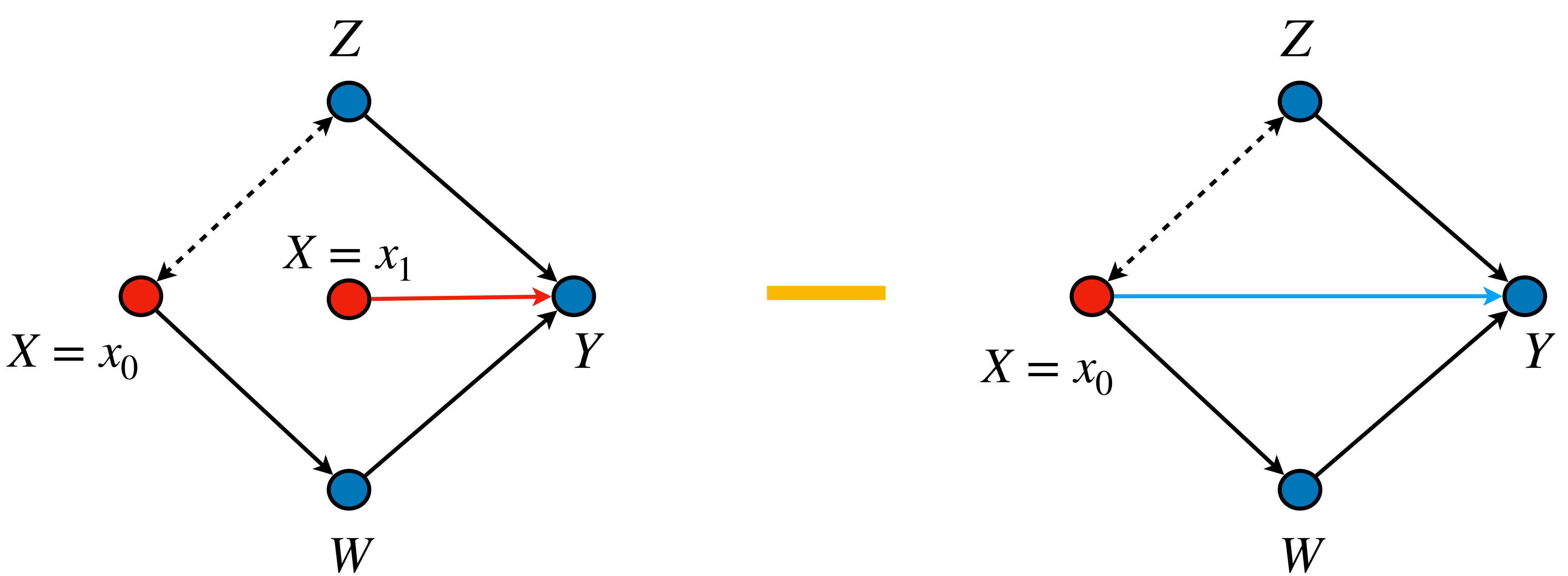
Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and 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, and structural direct effect exists.
Q: Is NDE powerful enough for detecting discrimination?

Gedankenexperiment (Ctf-DE)

- For a male person $X = x_0$, 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 $X = x_0$)?

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



$Y_{x_1,W_{x_0}} \mid X = x_0$ $\qquad$ $Y_{x_0,W_{x_0}} \mid X = x_0$

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

$U \leftarrow N(0,1)$

$X \leftarrow$ Bernoulli$(expit(U))$

$Z \leftarrow$ Bernoulli$(expit(U))$

$W \leftarrow$ Bernoulli(0.3)

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

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

Section 4.2
Example 4.2

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.

Key properties of Ctf-DE:

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

$U \leftarrow N$

$X \leftarrow$ B

$Z \leftarrow$ Bernoulli$(expit(U))$

$W \leftarrow$ Bernoulli(0.3)

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

$\mid x_0)$

$$= \sum_{z\in\{0,1\}} \sum_{w\in\{0,1\}} P(w)P(z \mid x_0)[\frac{1}{5}(1-2z) + \frac{1}{6}w - \frac{1}{6}w]$$

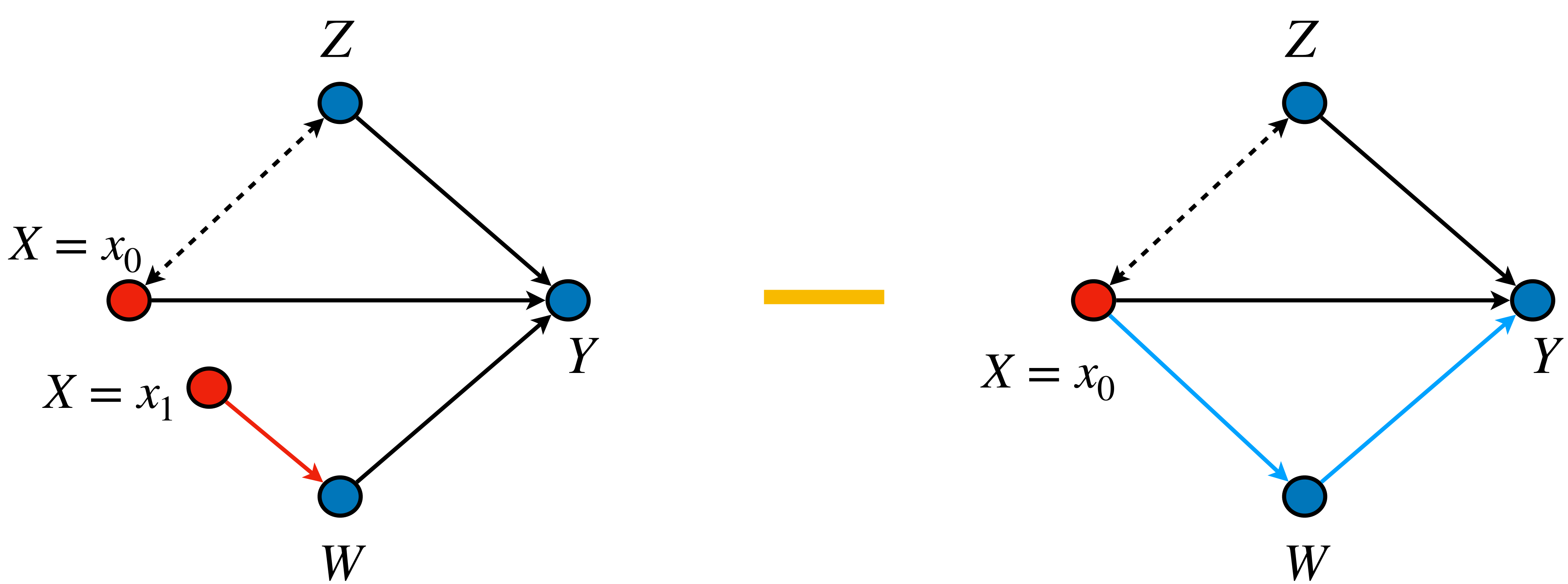
$$= \sum_{z\in\{0,1\}} \frac{1}{5}(1-2z)P(z \mid x_0) = 0.036.$$

Section 4.2
Example 4.2

Gedankenexperiment (Ctf-IE)

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

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

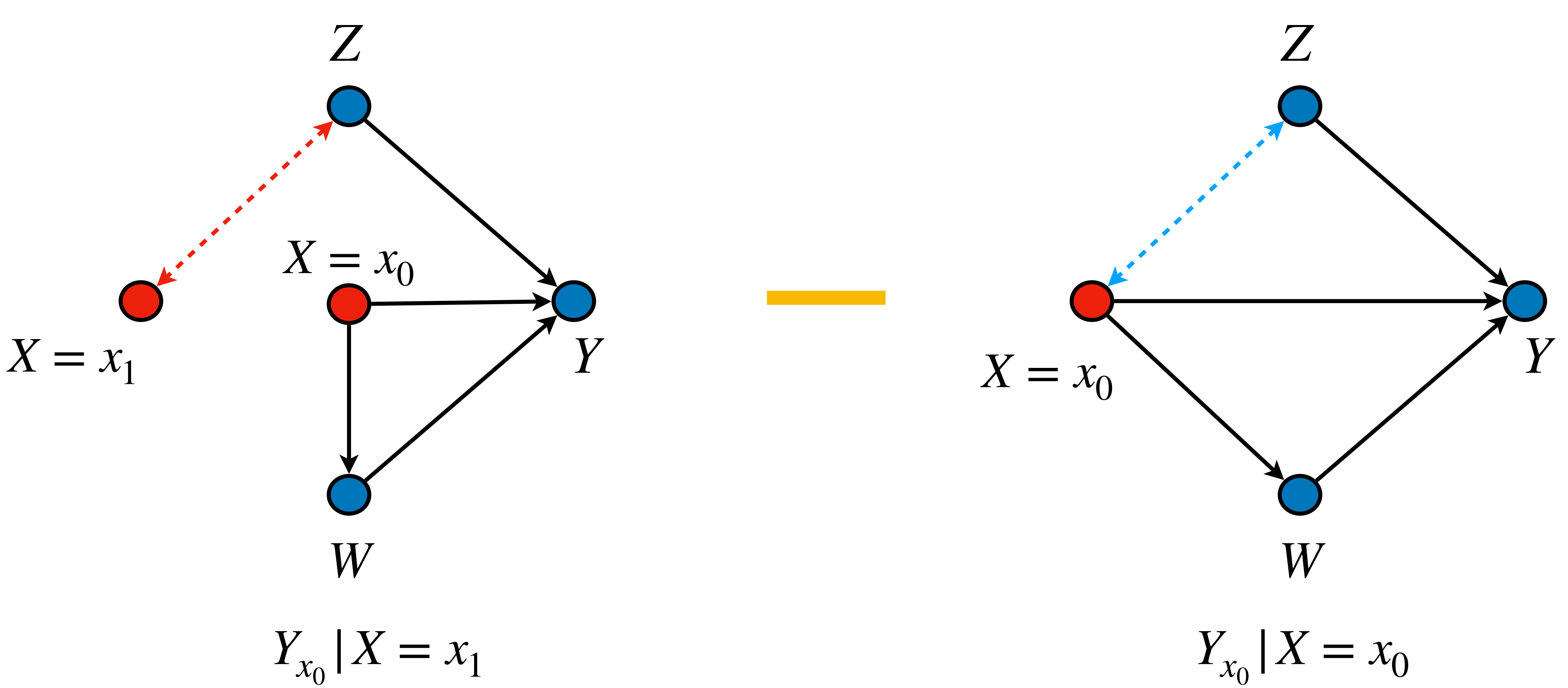


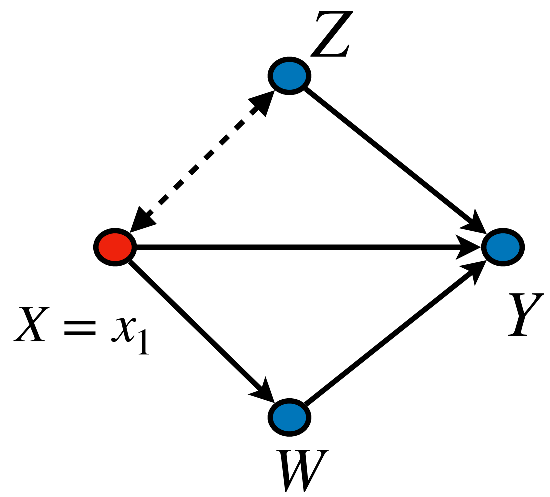
$Y_{x_0,W_{x_1}} | X = x_0$ $\qquad$ $Y_{x_0,W_{x_0}} | X = x_0$

Gedankenexperiment (Ctf-SE)

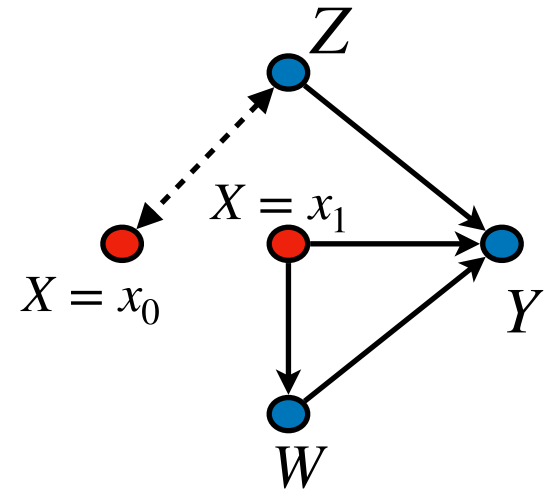
- For a male person $X = x_0$ and a female person ($X = x_1$), how would their salary (Y) differ had they both been male persons $X = x_0$?

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



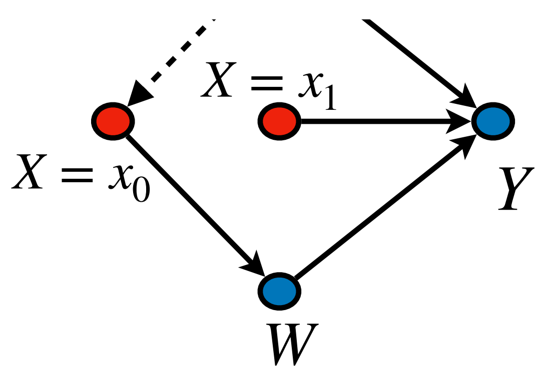


−



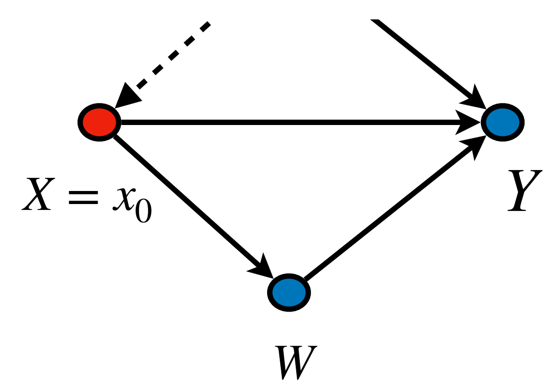
$-$ Ctf-SE$_{x_1,x_0}(y)$

TV Decomposition II (Causal Explanation Formula, ZB18)



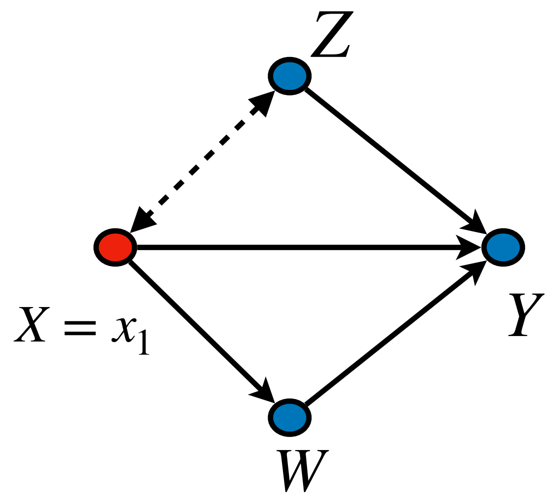
$Y_{x_1} \mid x_0$

−

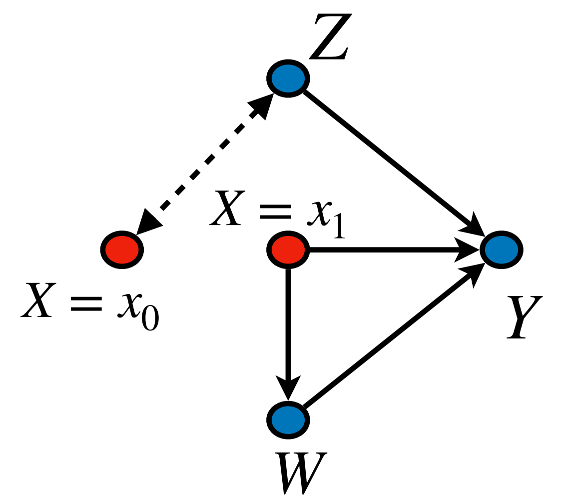


$Y \mid X = x_0$

Ctf-DE$_{x_0,x_1}(y \mid x_0)$

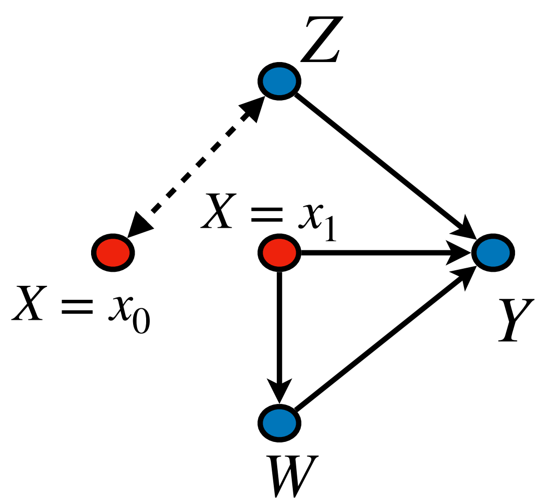


$Y \mid X = x_1$

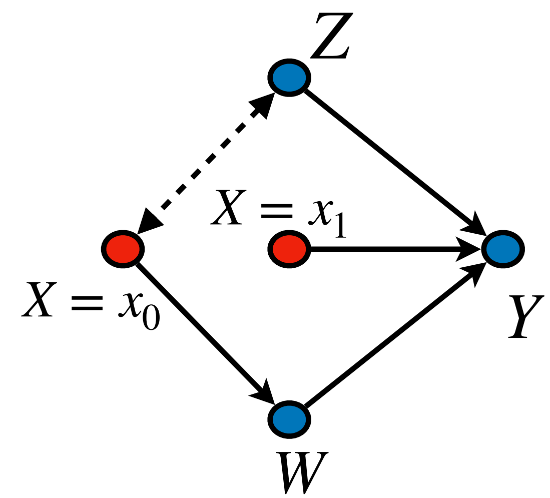


$Y_{x_1} \mid x_0$

$-\text{Ctf-SE}_{x_1,x_0}(y)$

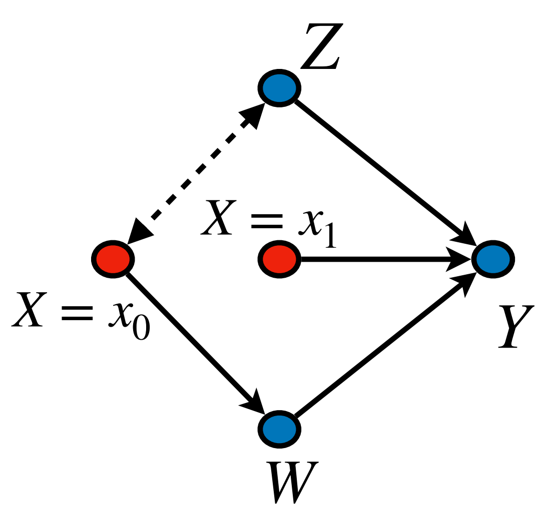


$Y_{x_1} \mid x_0$

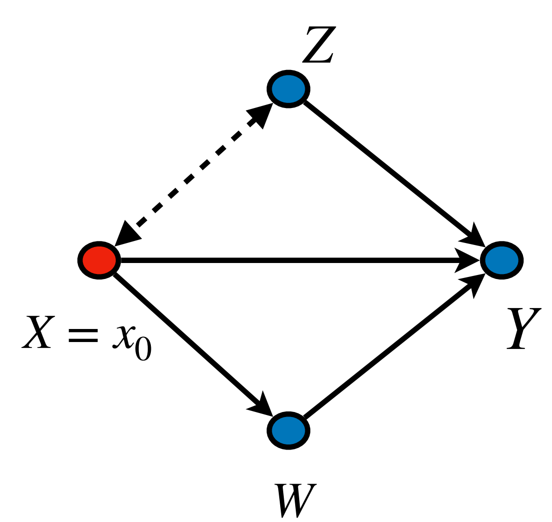


$Y_{x_1} \mid x_0$

$-\text{Ctf-IE}_{x_1,x_0}(y \mid x_0)$

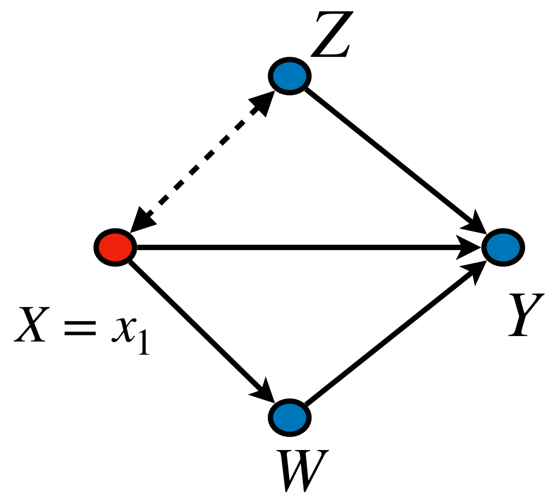


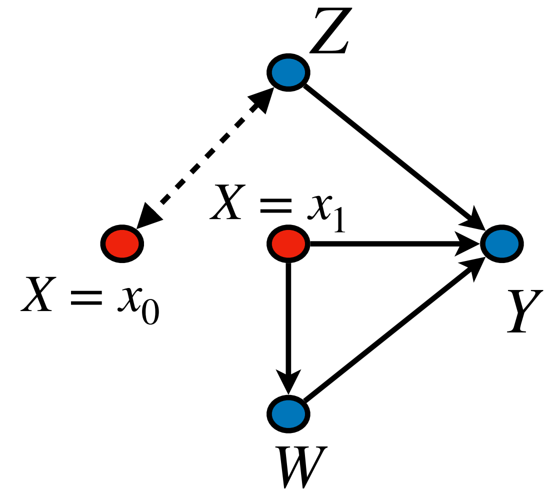
$Y_{x_1} \mid x_0$



$Y \mid X = x_0$

$\text{Ctf-DE}_{x_0,x_1}(y \mid x_0)$

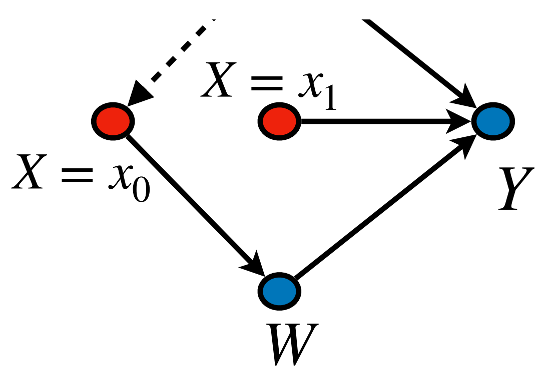




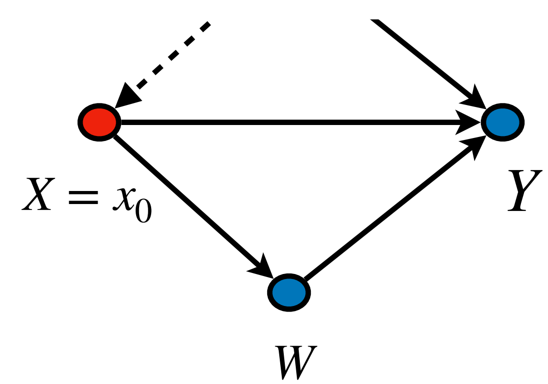
$-$Ctf-SE$_{x_1,x_0}(y)$

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

$$TV_{x_0,x_1}(y) = \underbrace{\text{Ctf-DE}_{x_0,x_1}(y \mid x_0)}_{\textit{direct}} - \underbrace{\text{Ctf-IE}_{x_1,x_0}(y \mid x_0)}_{\textit{indirect}} - \underbrace{\text{Ctf-SE}_{x_1,x_0}(y)}_{\textit{spurious}} .$$



$Y_{x_1} \mid x_0$



$Y \mid X = x_0$

Ctf-DE$_{x_0,x_1}(y \mid x_0)$

x-specific measures

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

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

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

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

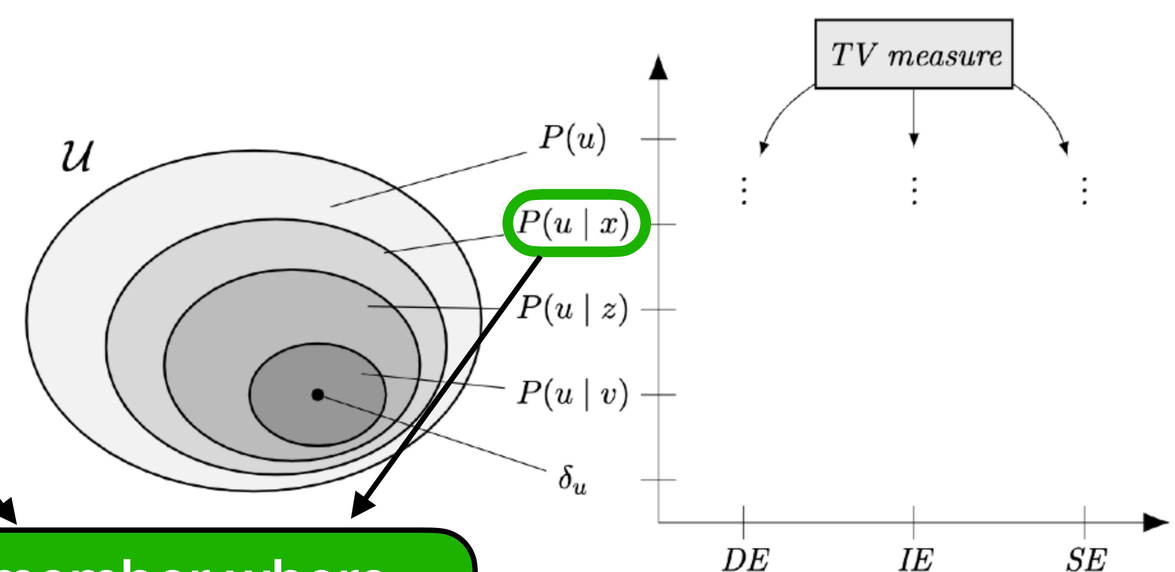
$$Ctf\text{-}SE_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y_{x_0} \mid x_0)\,.$$

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 $\mathcal{U}$

x-specific

$$TE_{x_0,x_1}(y \mid x) = P(y_{x_1}) - P(y_{x_0})$$
$$NDE_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0})$$
$$NIE_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1})$$
$$\text{Exp-SE}_{x_0,x_1}(y) = P(y_x) - P(y_x \mid x)\,.$$

where we came from

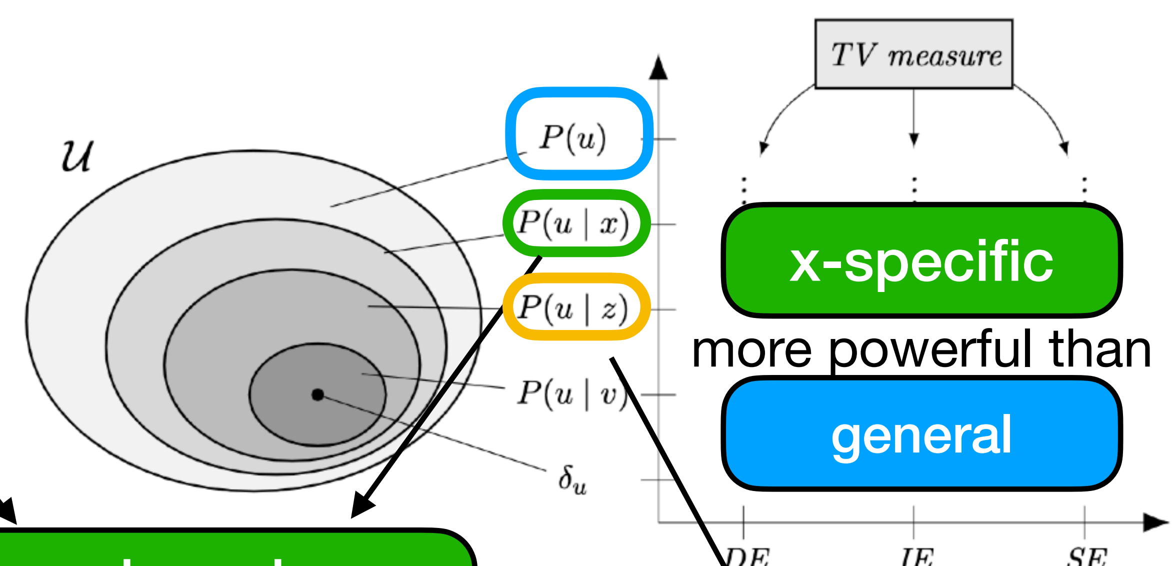
Definition. *The effect of treatment on*
direct, indirect, and spurious effects are

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

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*

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

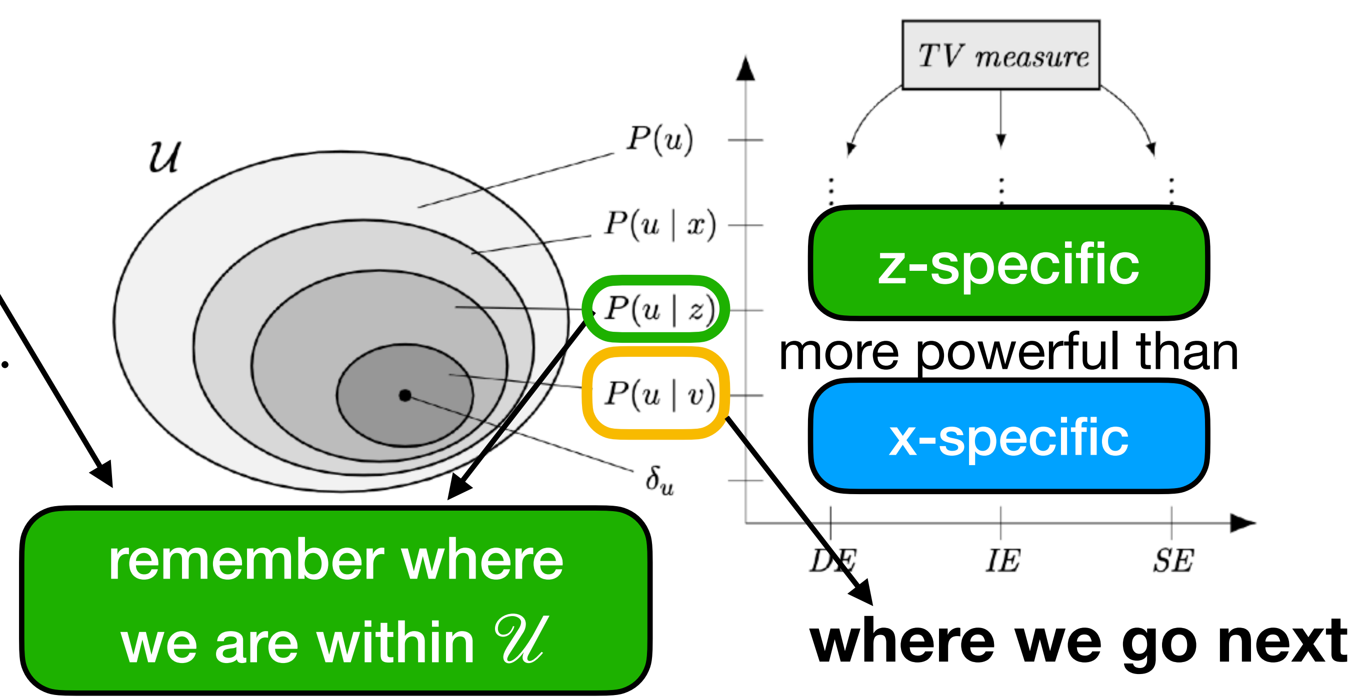
$$z\text{-}DE_{x_0,x_1}(y \mid z) = P(y_{x_1,W_{x_0}} \mid z) - P(y_{x_0} \mid z)$$

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

Structural Basis Expansion:

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

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



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

$U \leftarrow N(0,1)$

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

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

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

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

$$z\text{-}\mathbf{DE}(y \mid Z=0) = P(y_{x_1, W_{x_0}} \mid Z=0) - P(y_{x_0} \mid Z=0)$$

$$= P(\mathbf{Bernoulli}(\frac{1}{5}(1-Z) + \frac{1}{6}W) = 1 \mid Z=0)$$

$$-P(\mathbf{Bernoulli}(\frac{1}{5}(Z) + \frac{1}{6}W) = 1 \mid Z=0)$$

$$= \sum_{w \in \{0,1\}} P(w)[\frac{1}{5} + \frac{1}{6}w - \frac{1}{6}w] = \frac{1}{5}.$$

Section 4.2
Example 4.3

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($\frac{}{5}$ $\frac{}{6}$)

Key properties of z-DE:

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

$= 1 \mid Z = 0)$

$Z = 0)$

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

Section 4.2
Example 4.3

v'-specific measures

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

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

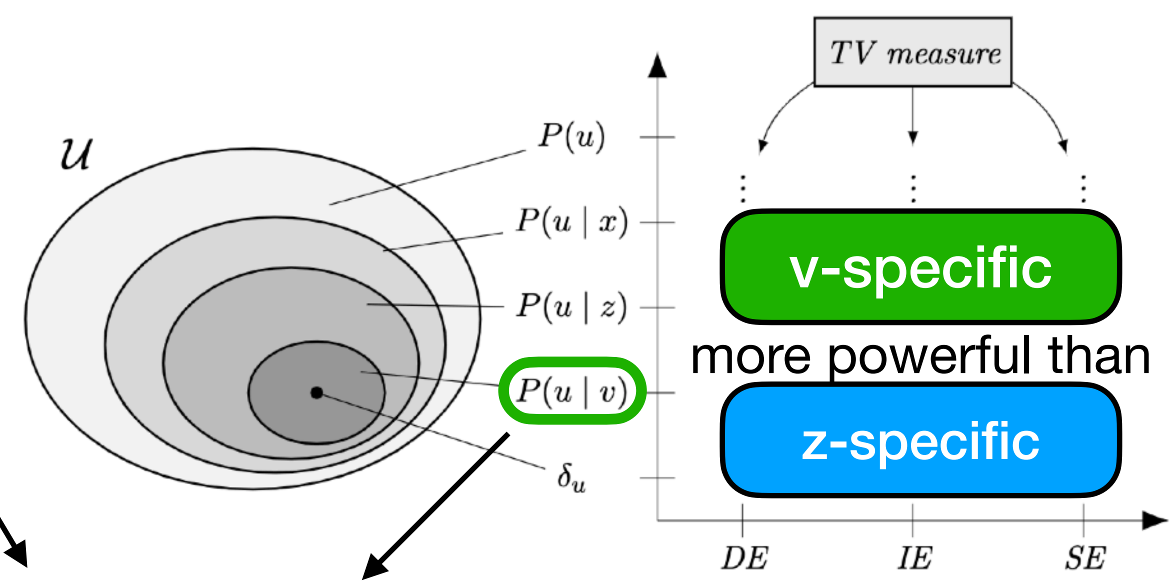
$$v'\text{-}DE_{x_0,x_1}(y \mid v') = P(y_{x_1,W_{x_0}} \mid v') - P(y_{x_0} \mid v')$$

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

Structural Basis Expansion:

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

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



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

Example — Probabilities of Causation (Ch. 9, Pearl, 2000)

By picking $v' = \{x_0, y_0\}$ and the total effect, the measure v'-TE becomes

$$
\begin{aligned}
(x, y)\text{-TE}_{x_0,x_1}(y \mid x_0, y_0) &= P(y_{x_1} \mid x_0, y_0) - P(y_{x_0} \mid x_0, y_0) \\
&= \boxed{P(y_{x_1} \mid x_0, y_0)\,.}
\end{aligned}
$$

Probability of sufficiency!

Similarly, v'-TE for the event $\{x_1, y_1\}$ equals

$$
\begin{aligned}
(x, y)\text{-TE}_{x_0,x_1}(y \mid x_1, y_1) &= P(y_{x_1} \mid x_1, y_1) - P(y_{x_0} \mid x_1, y_1) \\
&= 1 - P(y_{x_0} \mid x_1, y_1) \\
&= \boxed{P(y_{x_0} = 0 \mid x_1, y_1)\,.}
\end{aligned}
$$

Probability of necessity!

Unit-level measures

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

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

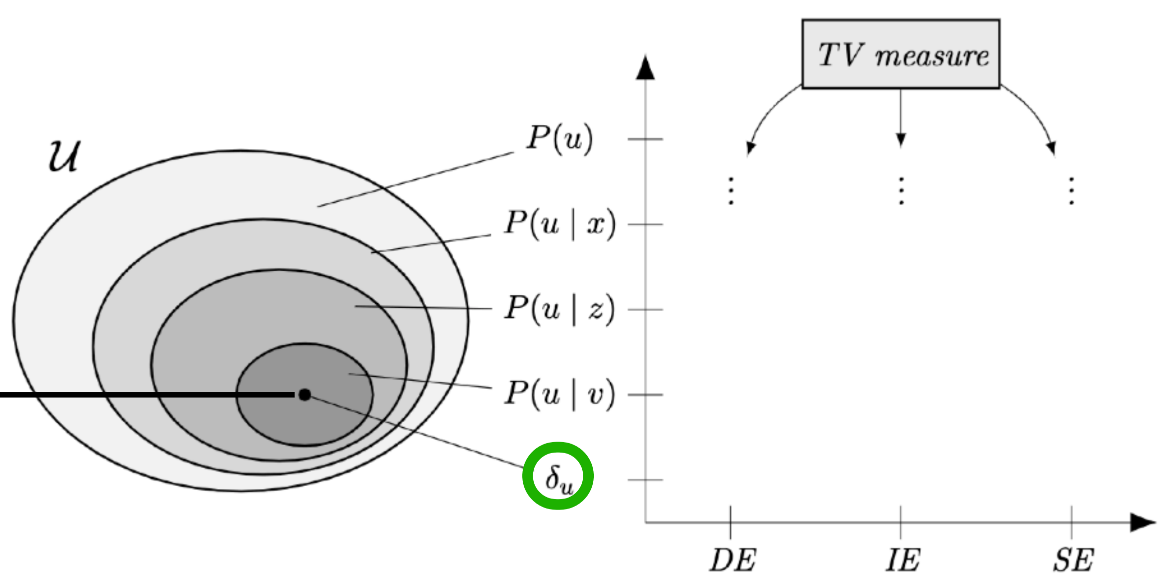
$$\textit{unit-DE}_{x_0,x_1}(y(u)) = y_{x_1,W_{x_0}}(u) - y_{x_0}(u)$$

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

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

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

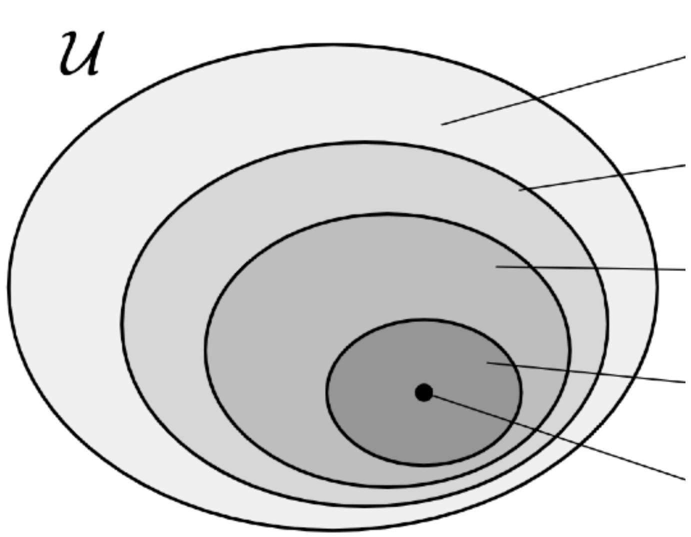
We reached the final, unit-level measures!



TV family measures as contrasts

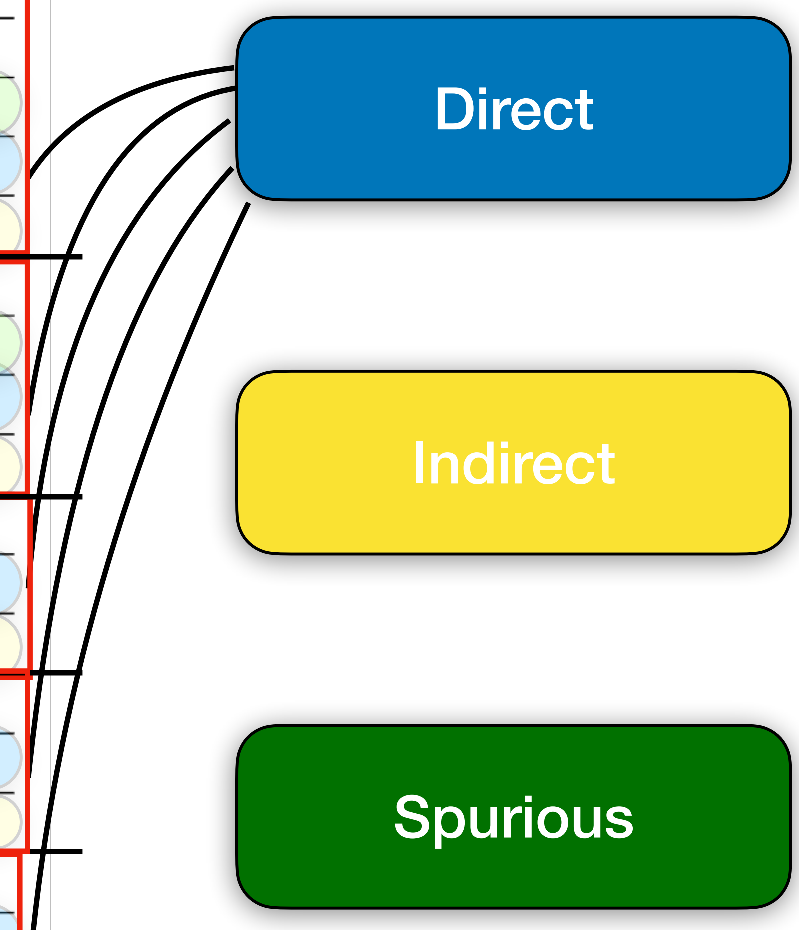
Lemma. *Under the Standard fairness model, all the measures within the TV family can be written as contrasts* $P(y_{C_1} \mid E_1) - P(y_{C_0} \mid E_0)$ *, following he constructions indicated below.*

units



	Measure	mechanism C_0	C_1	unit 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

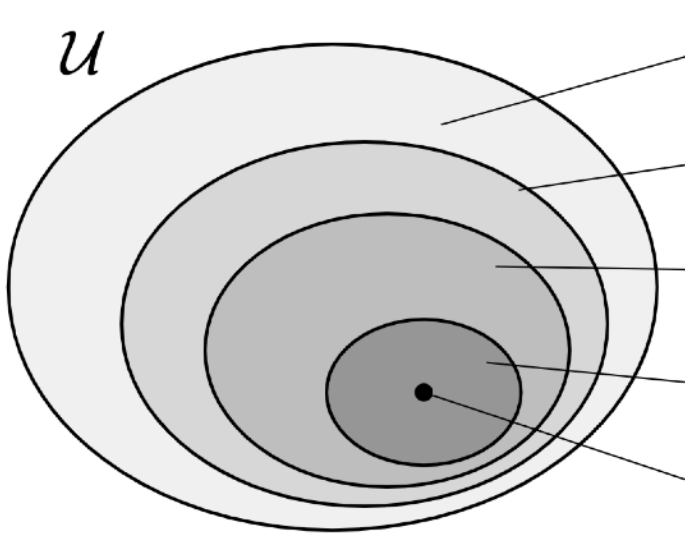


TV family measures as contrasts

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.*

	Measure	C_0	C_1	E_0	E_1
		mechanism		unit	
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

units



mechanisms

Direct

Indirect

Spurious

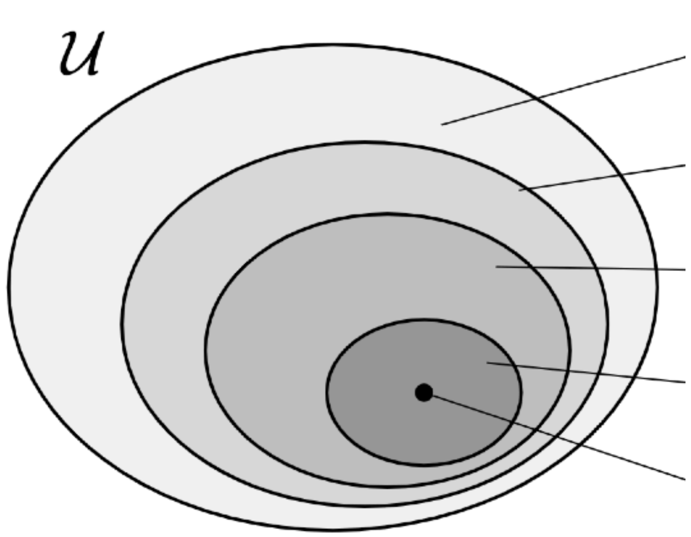
TV family measures as contrasts

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.*

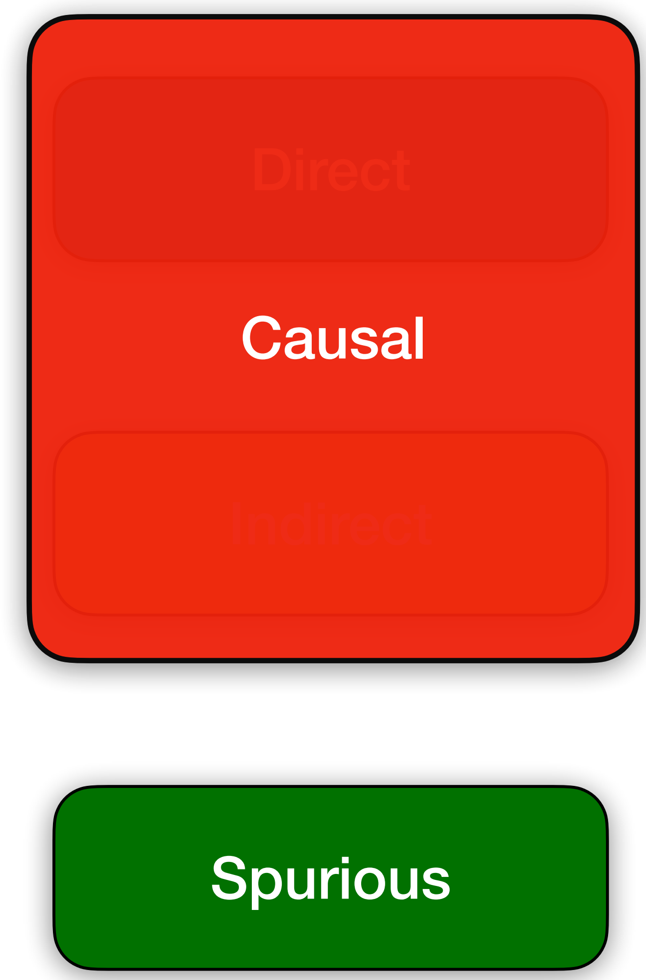
mechanism *unit*

	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	$\emptyset$	$\emptyset$
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

units



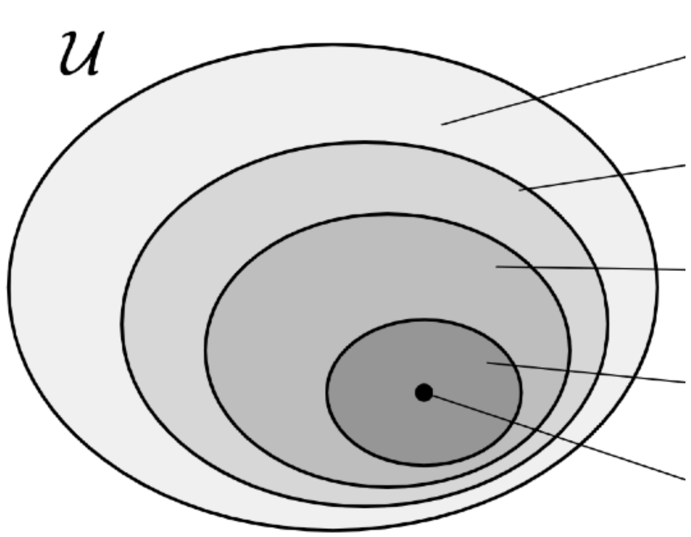
mechanisms



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.*

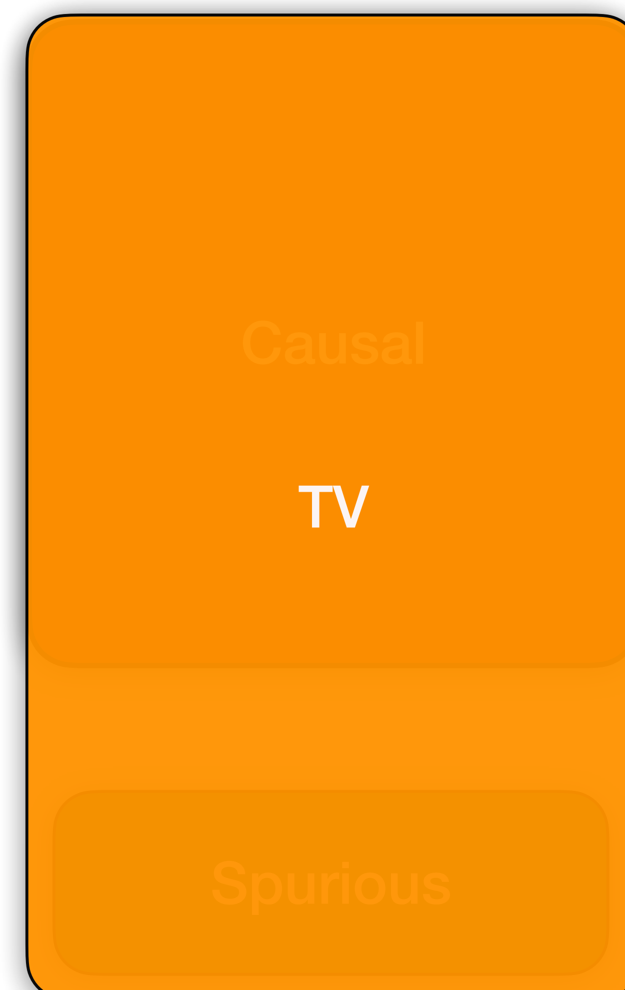
units



mechanism *unit*

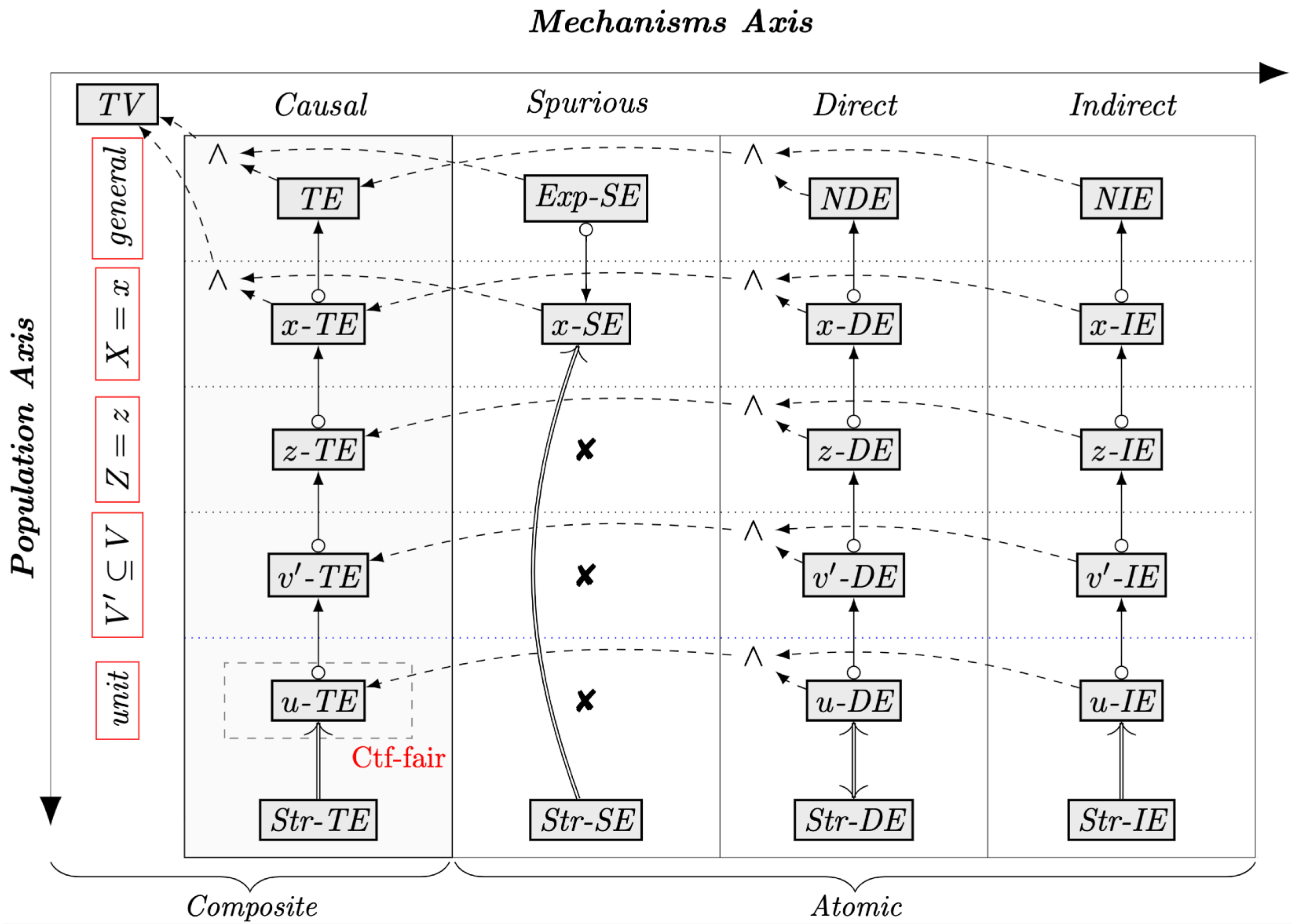
	Measure	C_0	C_1	E_0	E_1
	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
general	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
general	Exp-SE_{x}	x	x	$\emptyset$	x
general	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
general	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	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

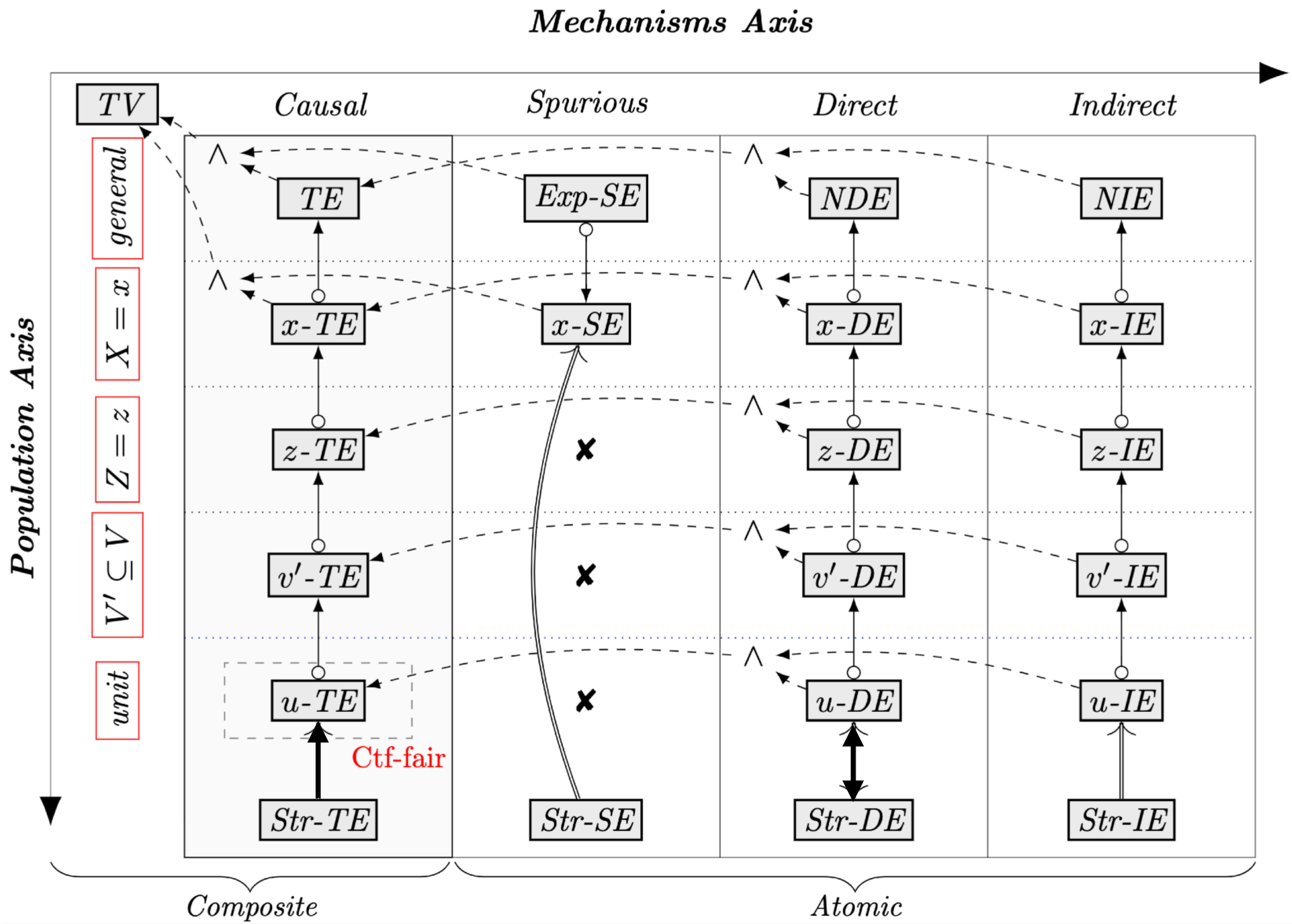


Fairness Map

Fairness Map

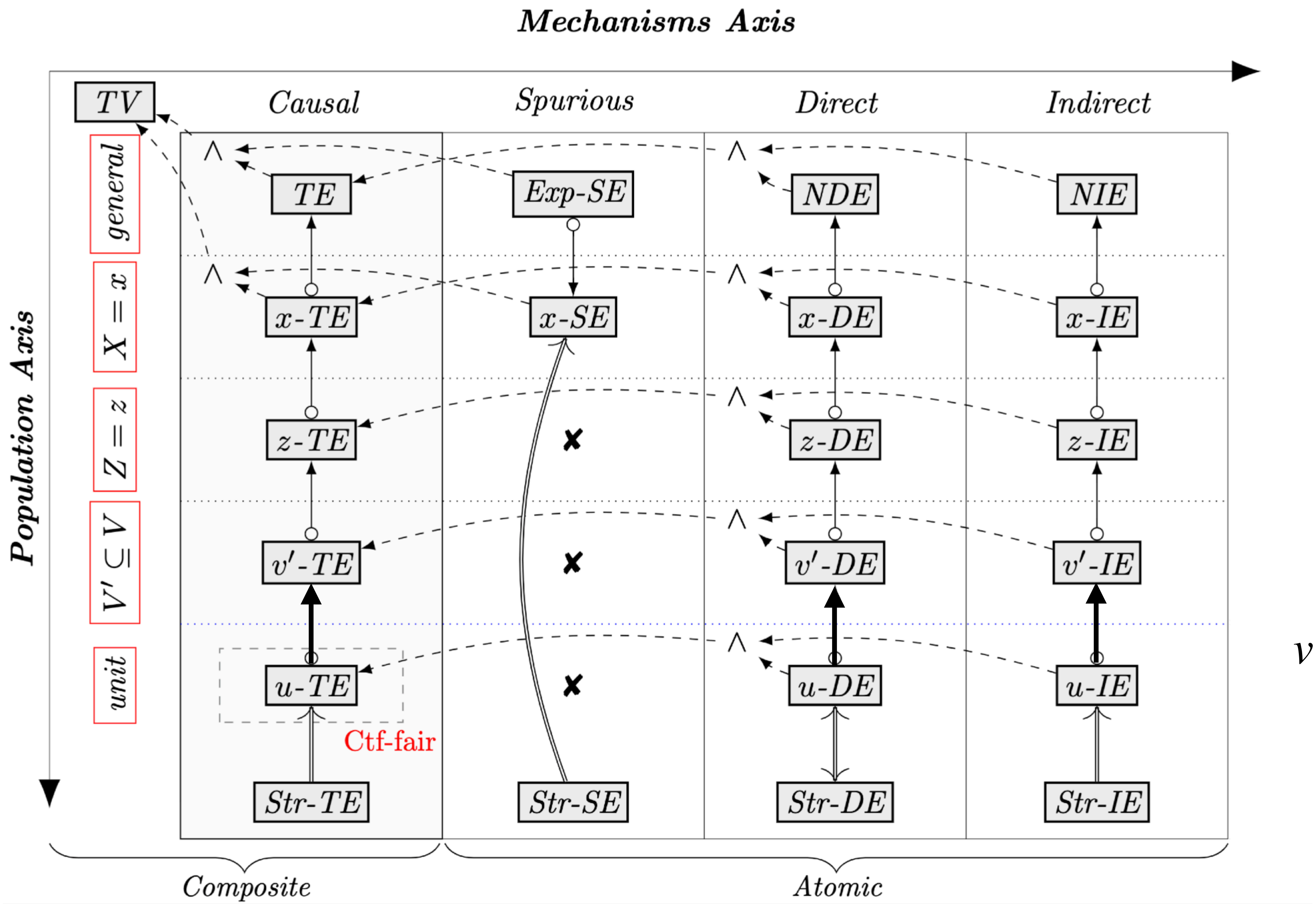


Fairness Map



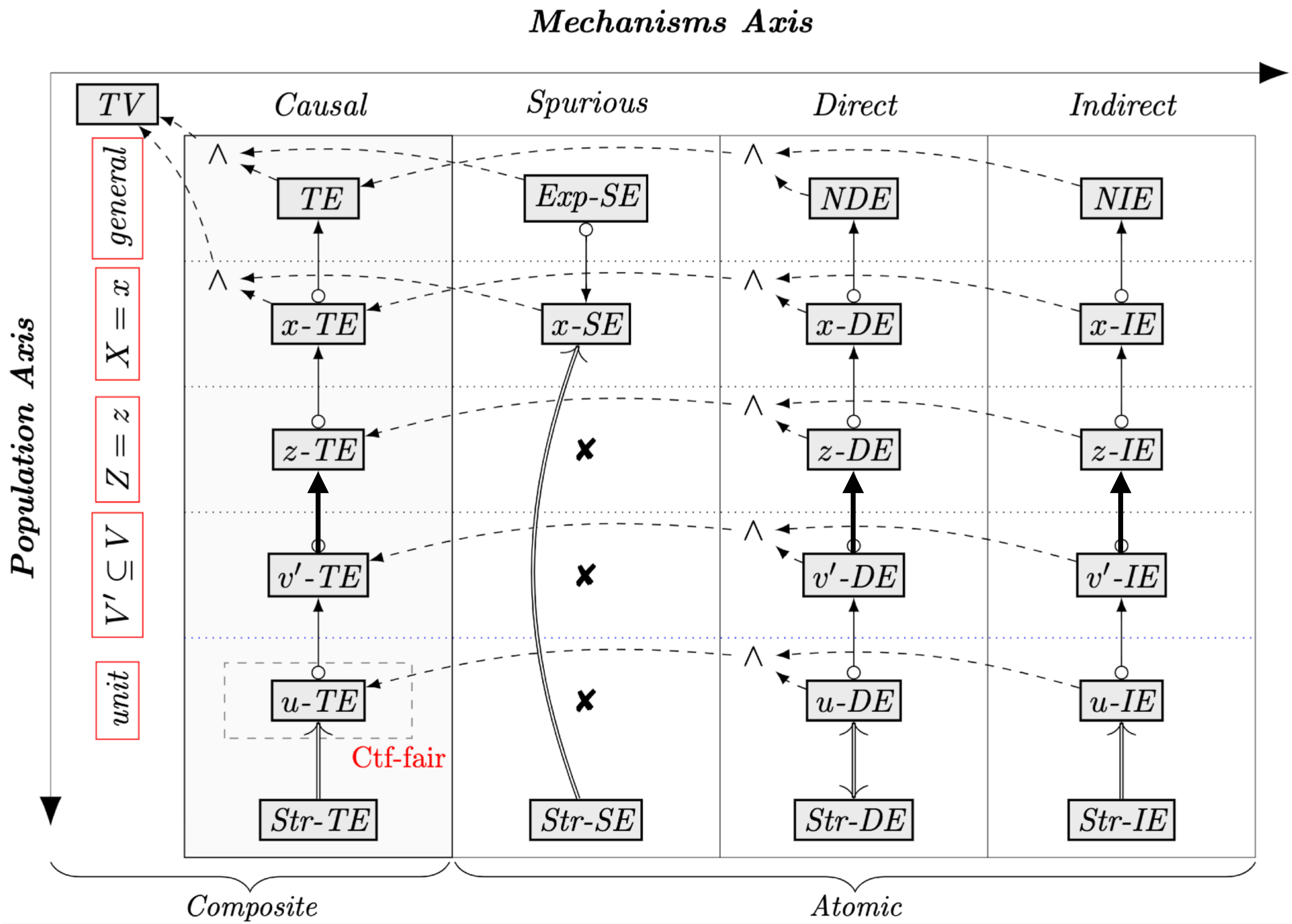
structural to unit

Fairness Map



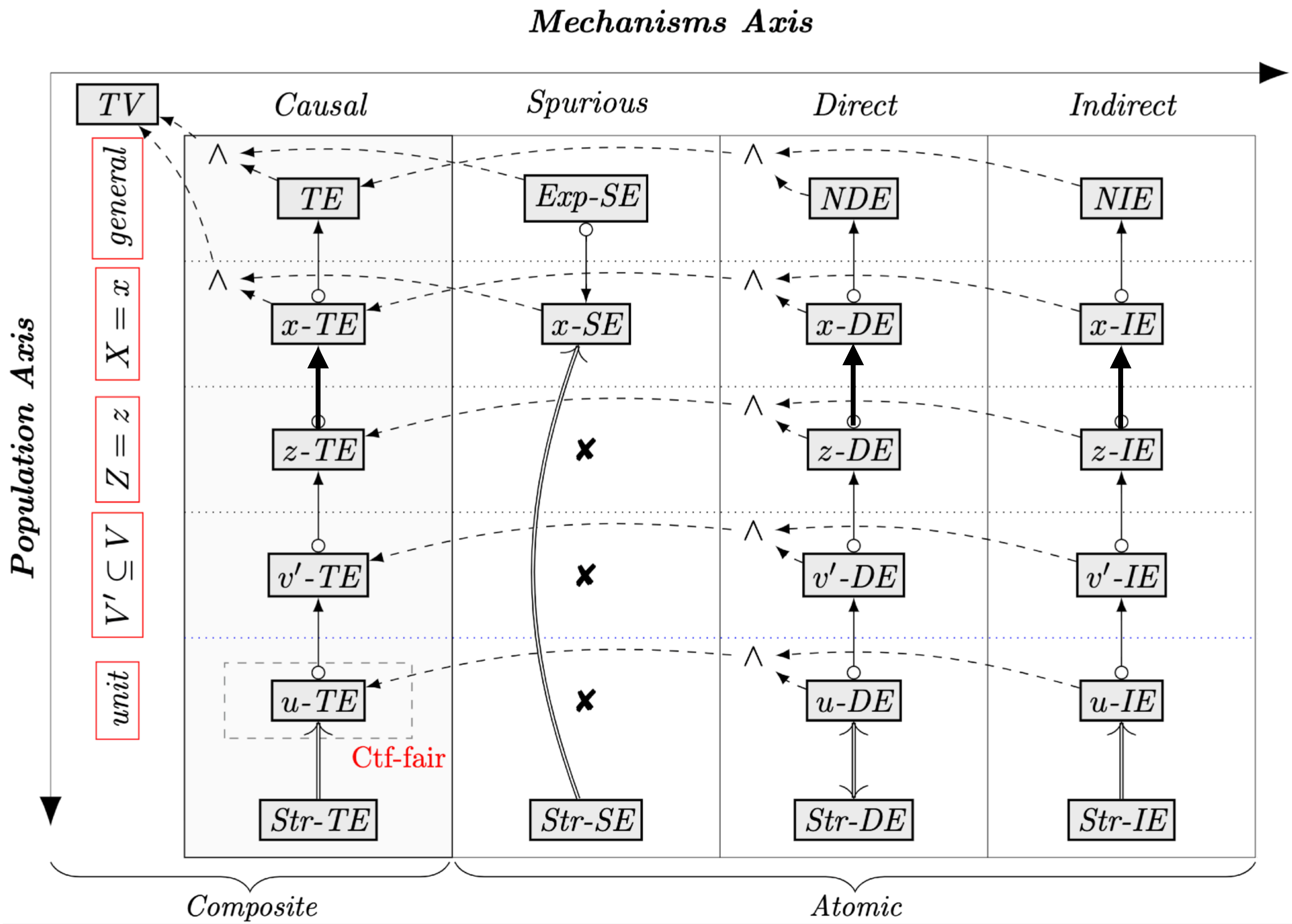
unit to
v'-specific

Fairness Map



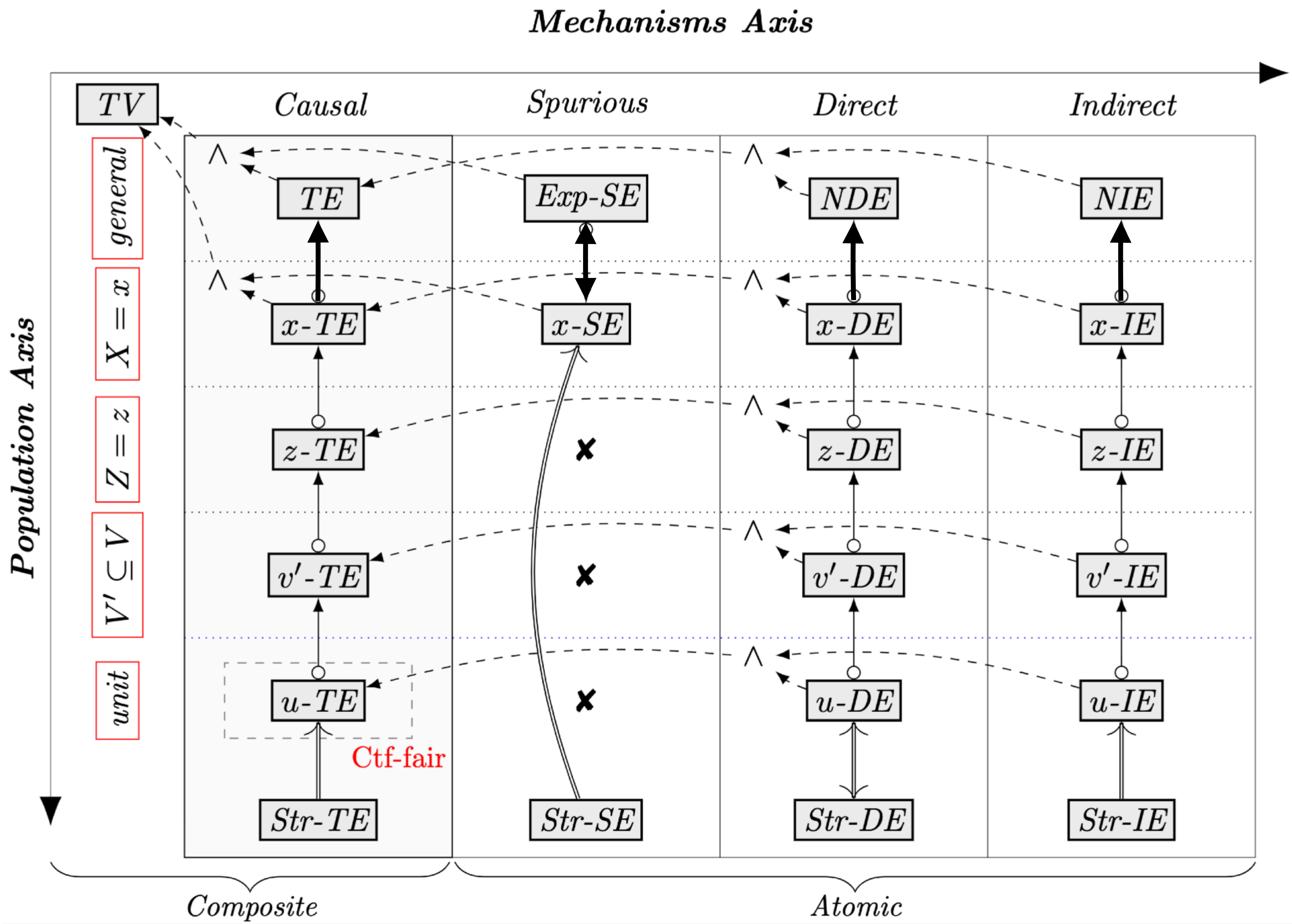
v'-specific to z-specific

Fairness Map



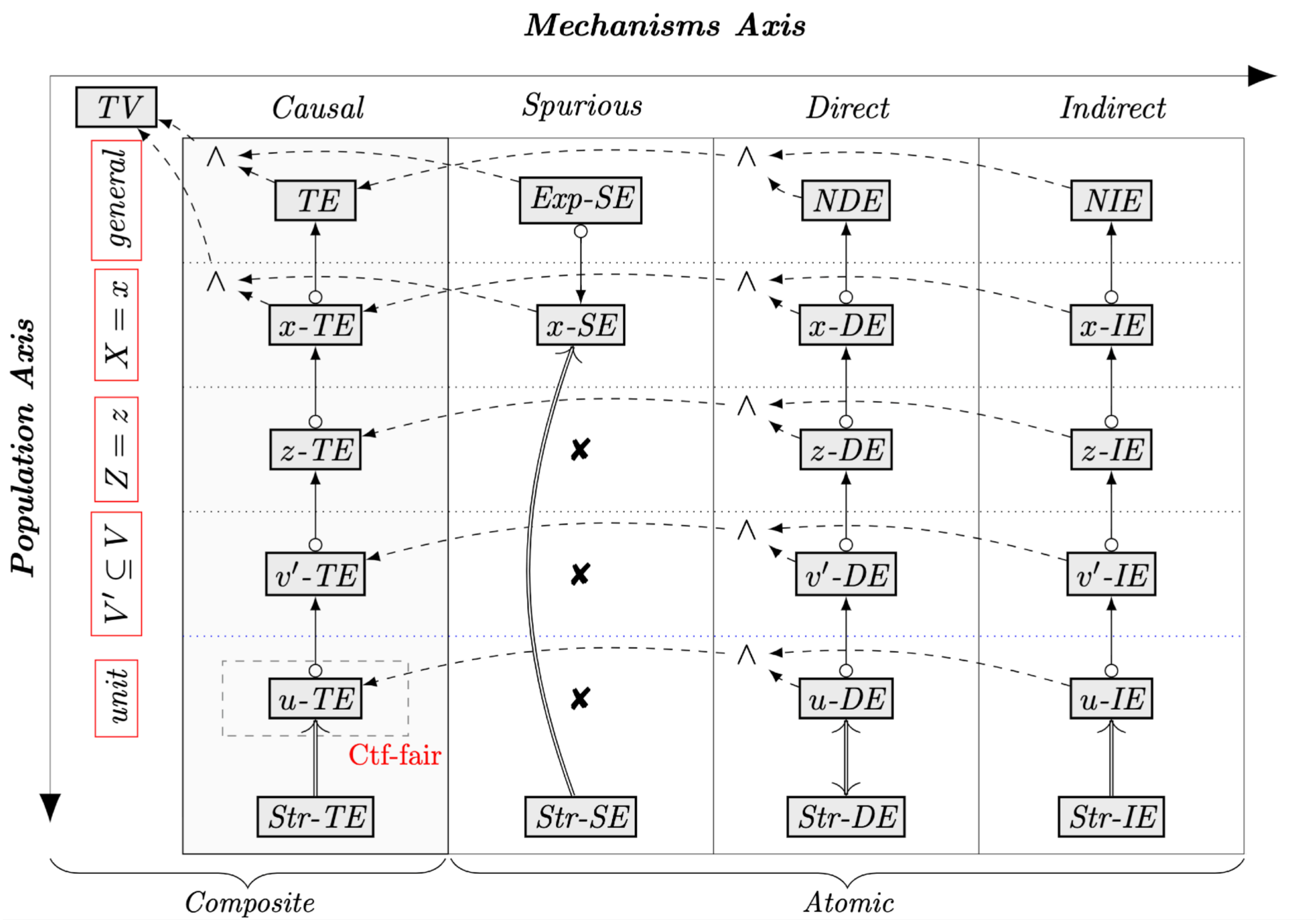
z-specific to x-specific

Fairness Map

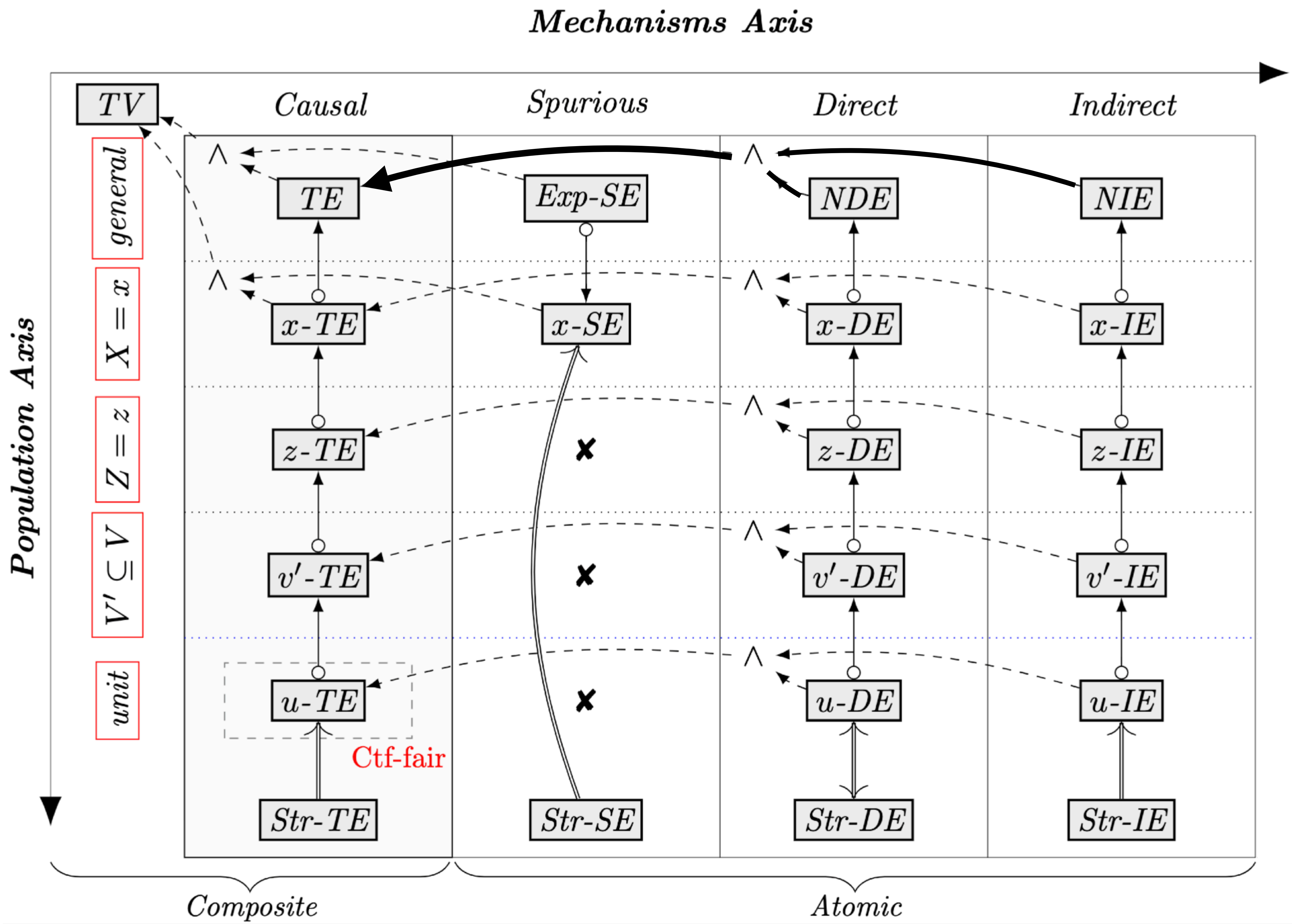


x-specific to general

Fairness Map

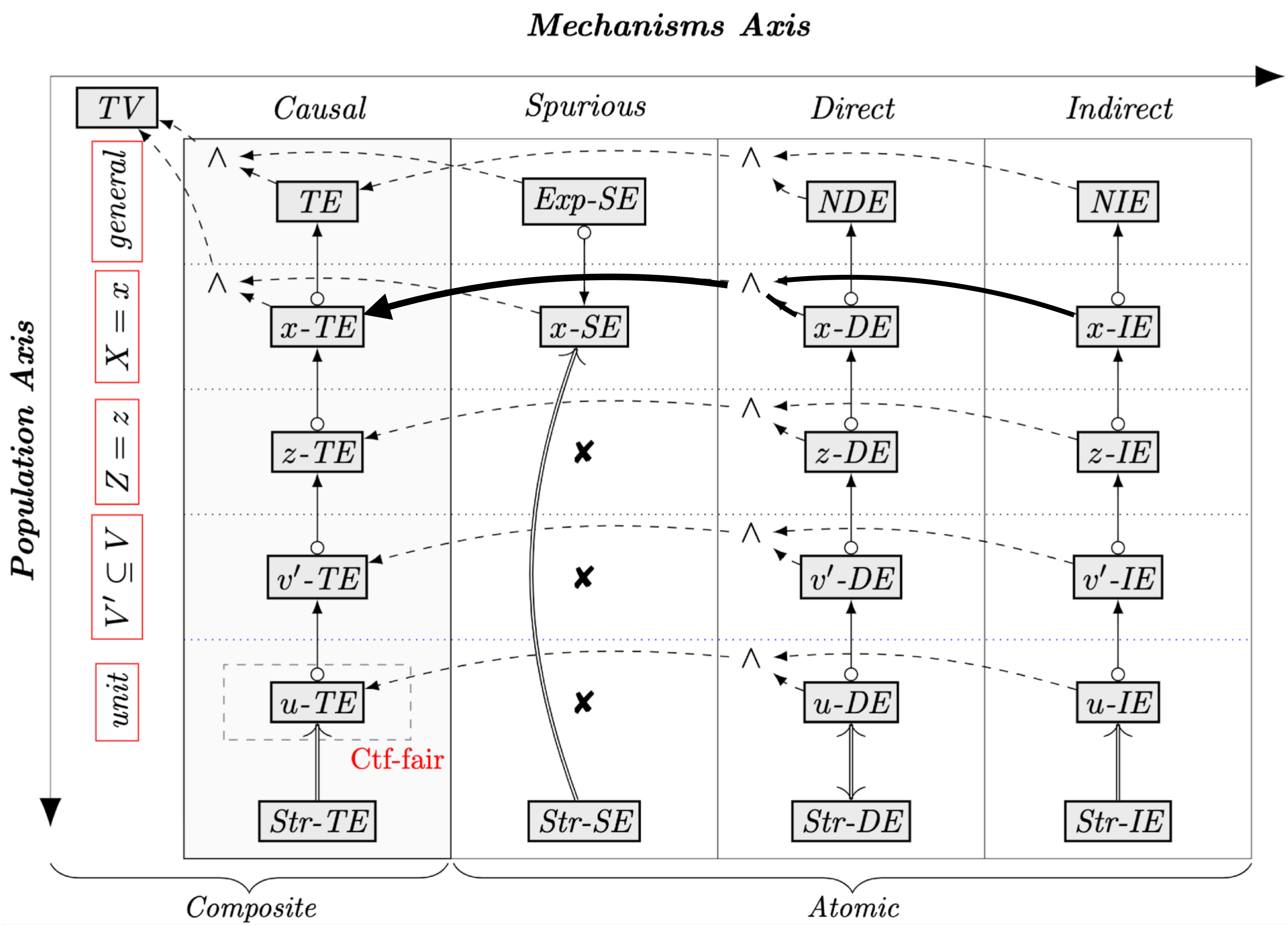


Fairness Map



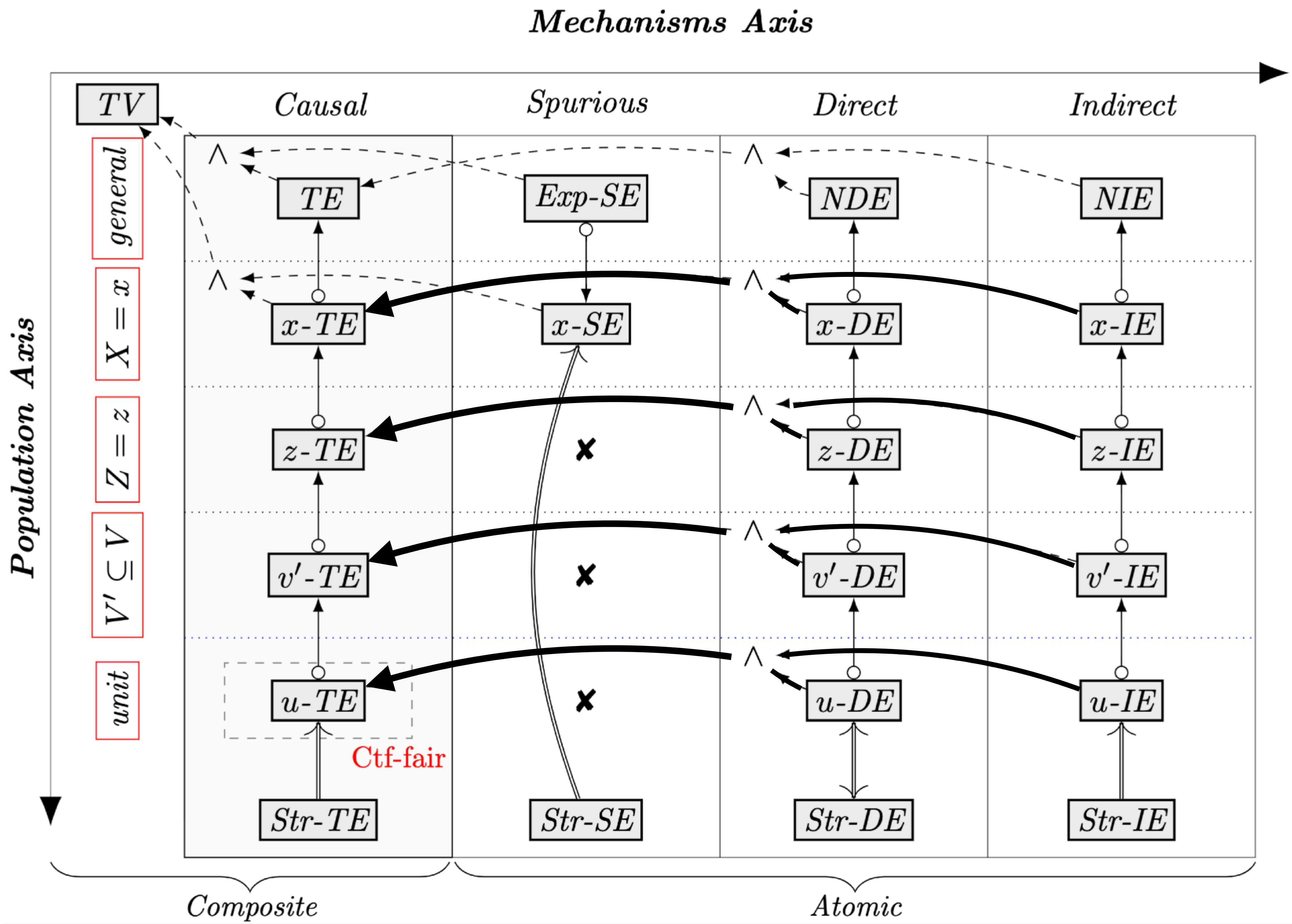
Mediation formula (Pearl, 2012)

Fairness Map



Extended Mediation Formula

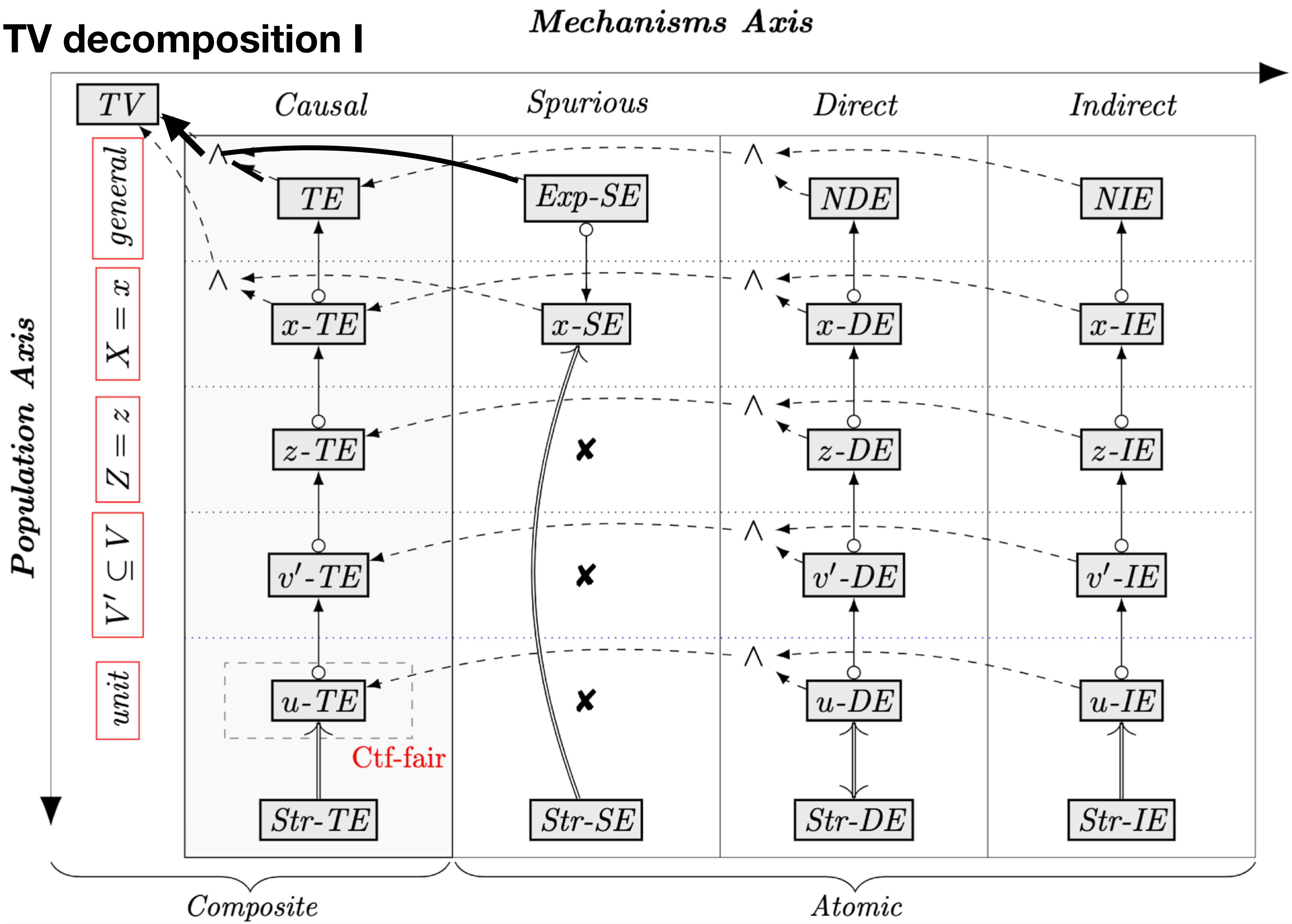
Fairness Map



Extended Mediation Formula

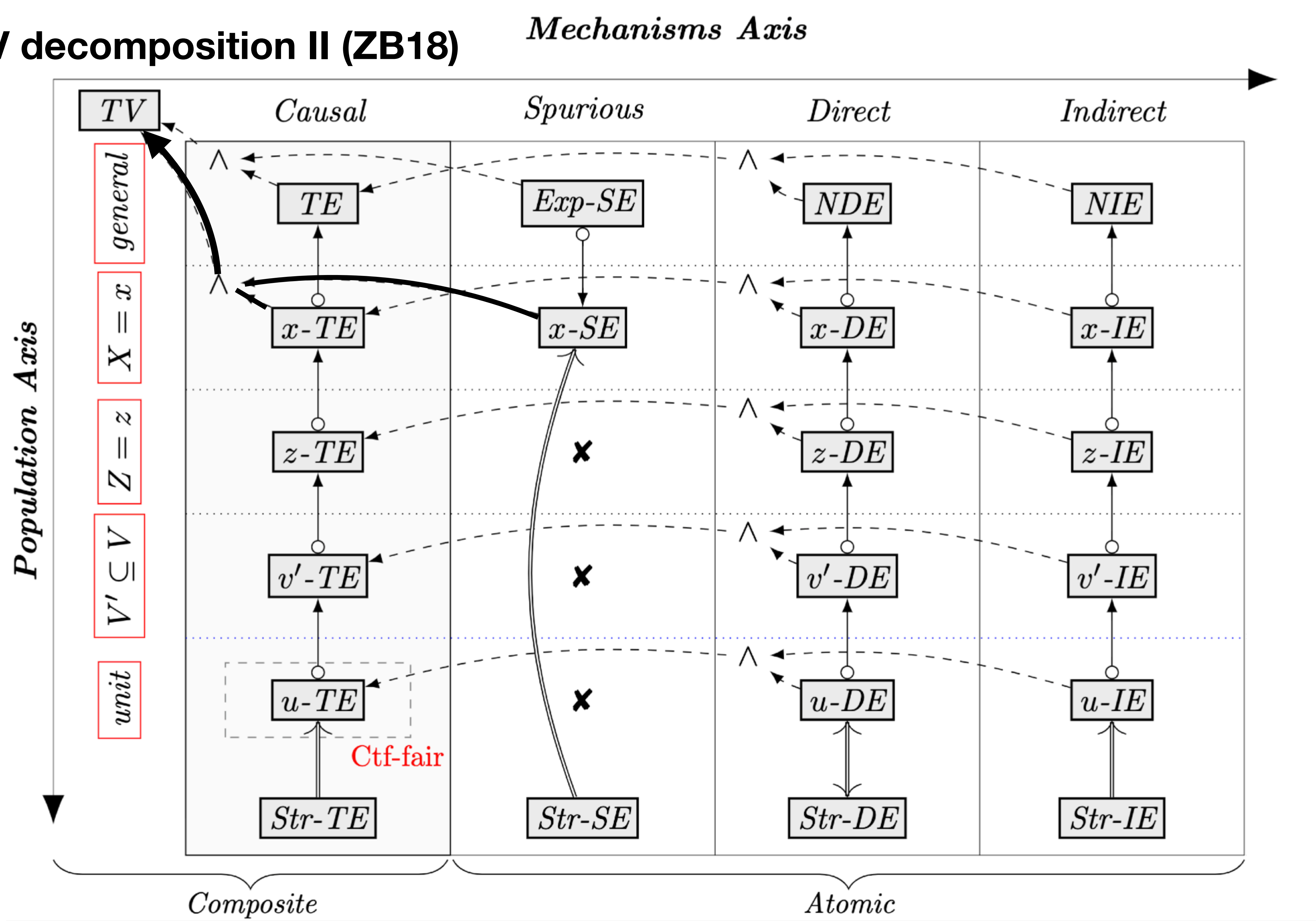
Fairness Map

TV decomposition I

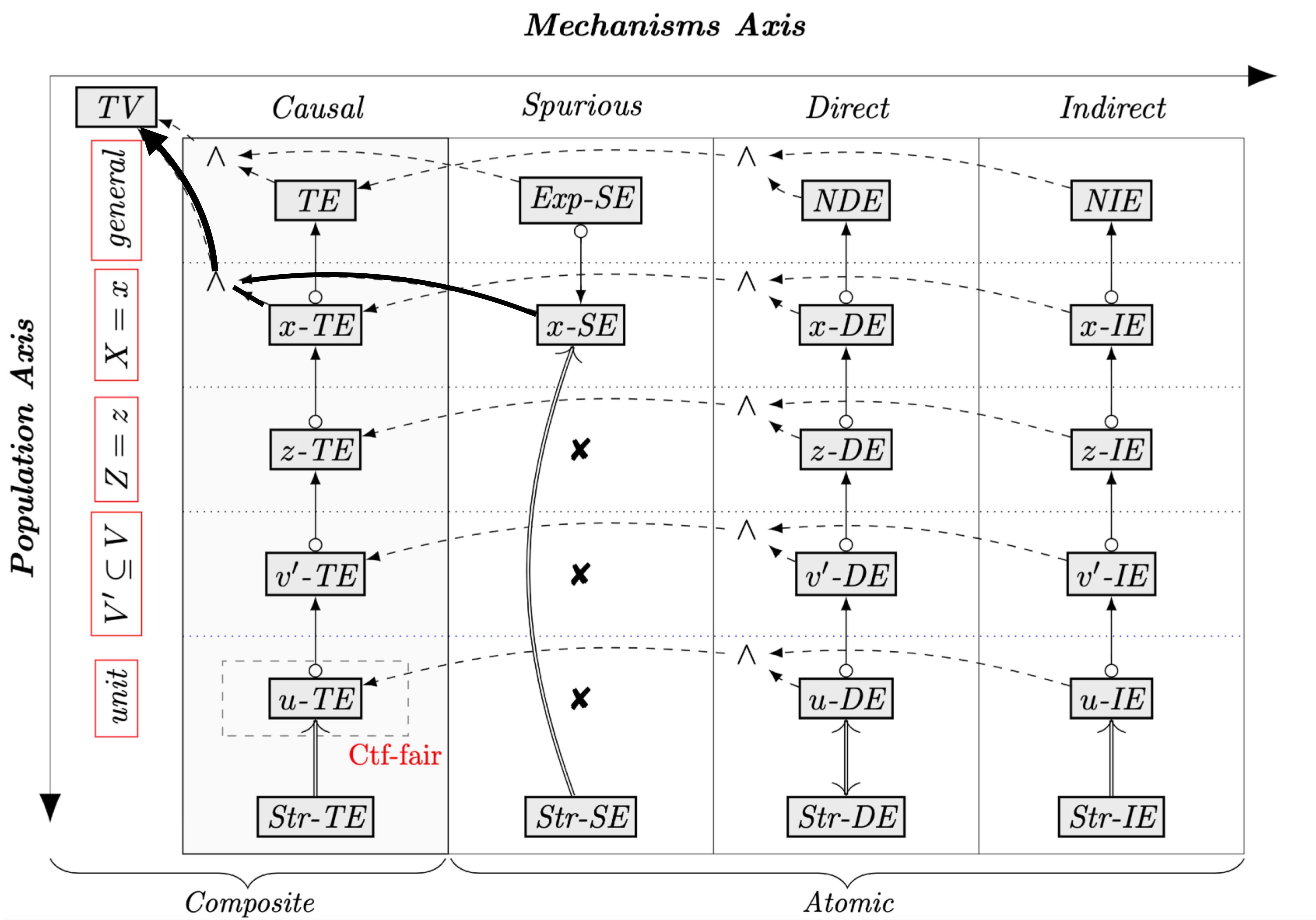


Fairness Map

TV decomposition II (ZB18)



Fairness Map



Fairness Map

Section 4.3
Theorem 4.9