# Value of Information Fairness

Frederik Hytting Jørgensen

October 13, 2022

# Background

## Definition: Causal influence diagram [Everitt et al. 2021]

A *causal influence diagram* (CID) is a DAG $\mathcal{G}$ where the nodes $\boldsymbol{V}$ are partitioned into *structure nodes* $\boldsymbol{X}$, *decision nodes* $\boldsymbol{D}$, and *utility nodes* $\boldsymbol{U}$. Utility nodes have no children.

# Background

## Definition: Causal influence diagram [Everitt et al. 2021]

A *causal influence diagram* (CID) is a DAG $\mathcal{G}$ where the nodes $\boldsymbol{V}$ are partitioned into *structure nodes* $\boldsymbol{X}$, *decision nodes* $\boldsymbol{D}$, and *utility nodes* $\boldsymbol{U}$. Utility nodes have no children.

## Definition: Structural causal influence model [Everitt et al. 2021]

A *structural causal influence model* (SCIM) is a tuple $\mathcal{M} = \langle \mathcal{G}, \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{F}}, P \rangle$.

- $\mathcal{G}$ is a CID.
- $\boldsymbol{\mathcal{E}} = \{\mathcal{E}_V\}_{V \in \boldsymbol{V} \setminus \boldsymbol{D}}$ is a set of *noise variables*.
- $\boldsymbol{\mathcal{F}} = \{f^V\}_{V \in \boldsymbol{V} \setminus \boldsymbol{D}}$ is a set of *structural functions*, $V := f^V(\mathbf{PA}^V, \mathcal{E}_V)$ for $V \in \boldsymbol{V} \setminus \boldsymbol{D}$.
- $P_{\boldsymbol{\mathcal{E}}}$ is a probability distribution for $\boldsymbol{\mathcal{E}}$ that makes the noise variables jointly independent.

# Background

We consider the setting with only one utility and decision node, $\boldsymbol{U} = \{U\}$ and $\boldsymbol{D} = \{D\}$, respectively. Once we specify a policy $D := \pi(\mathbf{PA}^D, \mathcal{E}_D)$, we get an SCM $\mathcal{M}_\pi$.

# Background

We consider the setting with only one utility and decision node, $\boldsymbol{U} = \{U\}$ and $\boldsymbol{D} = \{D\}$, respectively. Once we specify a policy $D := \pi(\mathbf{PA}^D, \mathcal{E}_D)$, we get an SCM $\mathcal{M}_\pi$.

**Counterfactual fairness [Kusner et al. 2017].**

Let $S$ be a sensitive feature. A non-random policy $\pi : \text{dom}(\mathbf{PA}^D) \to \text{dom}(D)$ is counterfactually fair if

$$P_\pi(D = D_{S:=s'}) = 1$$

for any $s' \in \text{dom}(S)$.

# Background

We consider the setting with only one utility and decision node, $\boldsymbol{U} = \{U\}$ and $\boldsymbol{D} = \{D\}$, respectively. Once we specify a policy $D := \pi(\mathbf{PA}^D, \mathcal{E}_D)$, we get an SCM $\mathcal{M}_\pi$.

## Counterfactual fairness [Kusner et al. 2017].

Let $S$ be a sensitive feature. A non-random policy $\pi : \mathrm{dom}(\mathbf{PA}^D) \to \mathrm{dom}(D)$ is counterfactually fair if

$$P_\pi(D = D_{S:=s'}) = 1$$

for any $s' \in \mathrm{dom}(S)$.

Often, we consider path-specific counterfactual fairness instead.

# Background

We consider the setting with only one utility and decision node, $\boldsymbol{U} = \{U\}$ and $\boldsymbol{D} = \{D\}$, respectively. Once we specify a policy $D := \pi(\mathbf{PA}^D, \mathcal{E}_D)$, we get an SCM $\mathcal{M}_\pi$.

Let $\mathcal{V}(\mathcal{M}) = \max_\pi \mathbb{E}_\pi(U)$ be the maximum attainable expected utility in $\mathcal{M}$.

# Background

We consider the setting with only one utility and decision node, $\boldsymbol{U} = \{U\}$ and $\boldsymbol{D} = \{D\}$, respectively. Once we specify a policy $D := \pi(\mathbf{PA}^D, \mathcal{E}_D)$, we get an SCM $\mathcal{M}_\pi$.

Let $\mathcal{V}(\mathcal{M}) = \max_\pi \mathbb{E}_\pi(U)$ be the maximum attainable expected utility in $\mathcal{M}$.

**Definition: Value of information (VoI) [Howard 1966, Everitt et al. 2021]**

A node $X \in \boldsymbol{X} \setminus \mathbf{DE}^D$ in an SCIM $\mathcal{M}$ has VoI if $\mathcal{V}(\mathcal{M}_{X \to D}) > \mathcal{V}(\mathcal{M}_{X \not\to D})$.

# Vol-fairness

Let $S \in \boldsymbol{X} \backslash \mathbf{DE}^D$ be a sensitive attribute. Let $\boldsymbol{O} = \mathbf{PA}^D \cup \mathbf{PA}^U \cup \{S, D\}$ denote observed variables. Out of the observed variables $\boldsymbol{O}$, we choose a subset $\boldsymbol{M} \subseteq \boldsymbol{O} \backslash (\mathbf{DE}^D \cup \{S, D\})$ that we call essential features.

# VoI-fairness

Let $S \in \boldsymbol{X} \backslash \mathbf{DE}^{D}$ be a sensitive attribute. Let $\boldsymbol{O} = \mathbf{PA}^{D} \cup \mathbf{PA}^{U} \cup \{S, D\}$ denote observed variables. Out of the observed variables $\boldsymbol{O}$, we choose a subset $\boldsymbol{M} \subseteq \boldsymbol{O} \backslash (\mathbf{DE}^{D} \cup \{S, D\})$ that we call essential features.

## Definition: VoI-fairness

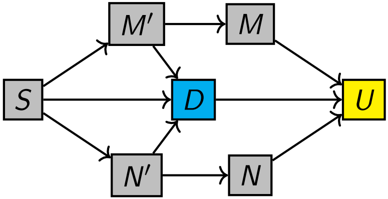
Let an SCIM $\mathcal{M}$ be given. We say that a utility

$$\widetilde{U} := g(\mathbf{PA}^{\widetilde{U}}),\ \mathbf{PA}^{\widetilde{U}} \subseteq \boldsymbol{O},$$

satisfies $\boldsymbol{M}$-VoI-fairness if $S$ does not have VoI in $\mathcal{M}^{do(U:=\widetilde{U})}_{\mathbf{PA}^{D}:=\boldsymbol{M}}$.

**Intuition**: Once the algorithm knows the essential features, it should not have an incentive to know $S$.

# Example




$$S :\sim \text{Unif}\{-1, 1\}$$

$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$

$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$

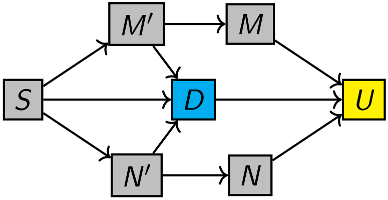
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$

$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$

$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$

# Example




$$S :\sim \text{Unif}\{-1, 1\}$$
$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$
$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$
$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$
$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$
$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$

$S$ : Gender.

$M'$ : Objective measure of medical qualifications.

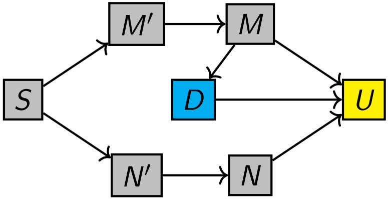
$N'$ : How much the interviewers like the applicant.

$M$ : Recovery rate of patients.

$N$ : An evaluation by colleagues.

$U$ : Job performance measure collected after 1 year.

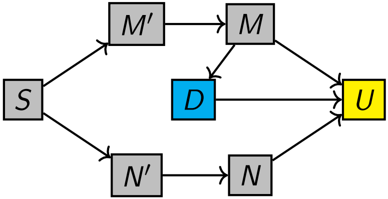
# Example




Assume $\boldsymbol{M} = \{M\}$.

$$S :\sim \text{Unif}\{-1, 1\}$$
$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$
$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$
$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$
$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$
$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$

# Example




$$S :\sim \text{Unif}\{-1, 1\}$$
$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$
$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$
$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$
$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$
$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$

Assume $\boldsymbol{M} = \{M\}$. $S$ has VoI in $\mathcal{M}_{\mathbf{PA}^D := \{M\}}$, so we modify the utility:

$$\begin{aligned}\widetilde{U} &:= \mathbb{1}(D = 1) \cdot (U - \theta_S^{N'} \theta_{N'}^{N} \theta_N^U S) \\ &= \mathbb{1}(D = 1) \cdot (\theta_N^U(\mathcal{E}_N + \theta_{N'}^{N} \mathcal{E}_{N'}) + \theta_M^U M)\end{aligned}$$

$\widetilde{U}$ is $\{M\}$-VoI-fair, and it is easy to show that optimal policies in $\mathcal{M}^{do(U := \widetilde{U})}$ satisfy path-specific counterfactual fairness with unfair paths $\{S \to N' \to D, S \to D\}$.

# Appropriate VoI-fair utility

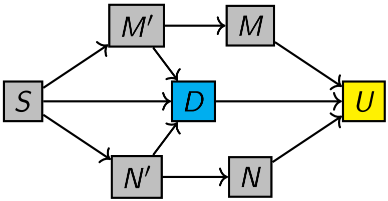
VoI-fair utilities always exist since you can use a constant utility.

## Definition: Appropriate VoI-fair utility

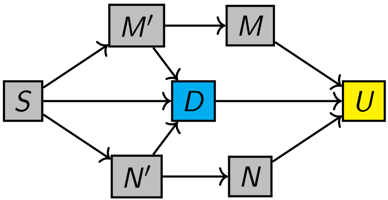
Let a set of essential features $\boldsymbol{M}$ and a set of utilities $\mathcal{U}$ be given. Let $\Pi(\widetilde{U})$ be optimal policies in $\mathcal{M}^{do(U:=\widetilde{U})}$. A utility $\widetilde{U}$ is an appropriate $\boldsymbol{M}$-VoI-fair utility w.r.t. $\mathcal{U}$ if it solves the following optimization problem:

$$\text{Maximize: } \inf_{\pi \in \Pi(\widetilde{U})} \mathbb{E}_{\mathcal{M}_\pi}(U) \quad \text{for } \widetilde{U} \in \mathcal{U}$$

$$\text{Subject to: } \quad \widetilde{U} \text{ satisfies } \boldsymbol{M}\text{-VoI-fairness.}$$




$$S := \mathcal{E}_S$$
$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$
$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$
$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$
$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$
$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$




$$S := \mathcal{E}_S$$
$$M' := \theta_S^{M'} S + \mathcal{E}_{M'}$$
$$N' := \theta_S^{N'} S + \mathcal{E}_{N'}$$
$$M := \theta_{M'}^{M} M' + \mathcal{E}_M$$
$$N := \theta_{N'}^{N} N' + \mathcal{E}_N$$
$$U := \mathbb{1}(D = 1) \cdot (\theta_N^U N + \theta_M^U M)$$
$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I)$$

## Proposition

Let

$$\mathcal{U} = \{(S, N, M, D) \mapsto \mathbb{1}(D = 1)(w_1 S + w_2 N + w_3 M) \mid (w_1, w_2, w_3) \in \mathbb{R}^3\}.$$

Assume that all $\theta$s are strictly positive. Then, $(w_1, w_2, w_3) = \left(-\theta_N^U \theta_N^{N'} \theta_S^{N'}, \theta_N^U, \theta_M^U + \frac{\theta_S^{N'} \theta_{N'}^{N} \theta_N^U \theta_S^{M'}}{((\theta_S^{M'})^2 + 1)\theta_{M'}^{M}}\right)$ corresponds to an appropriate $\{M\}$-VoI-fair utility w.r.t. $\mathcal{U}$.

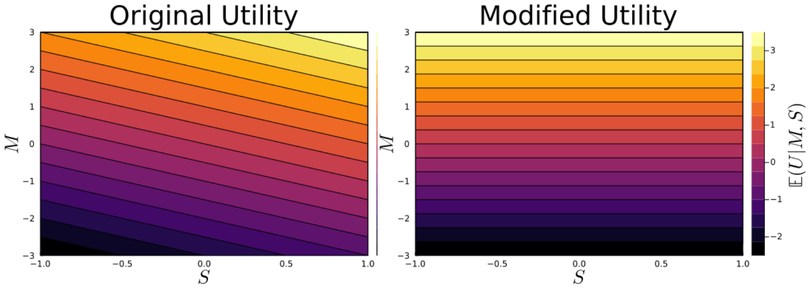
## Proposition

Let

$$\mathcal{U} = \{(S, N, M, D) \mapsto \mathbb{1}(D = 1)(w_1 S + w_2 N + w_3 M) \mid (w_1, w_2, w_3) \in \mathbb{R}^3\}.$$

Assume that all $\theta$s are strictly positive. Then, $(w_1, w_2, w_3) = \left(-\theta_N^U \theta_N^{N'} \theta_S^{N'}, \theta_N^U, \theta_M^U + \frac{\theta_S^{N'} \theta_{N'}^N \theta_N^U \theta_S^{M'}}{((\theta_S^{M'})^2 + 1)\theta_{M'}^M}\right)$ corresponds to an appropriate $\{M\}$-VoI-fair utility w.r.t. $\mathcal{U}$.

*Proof sketch:* Maximize $E(U \mid E\left(w_1 S + w_2 N + w_3 M \mid M', S, N'\right) > 0)$ under the constraint $w_1 = -w_2 \theta_{N'}^N \theta_S^{N'}$.


Original Utility
Modified Utility
$M$
$S$
$\mathbb{E}(U|M,S)$
-1.0
-0.5
0.0
0.5
1.0
3
2
1
0
-1
-2
-3

# Why do I think this definition is interesting?

- The definition is intuitive and gives intuitive results in concrete cases.

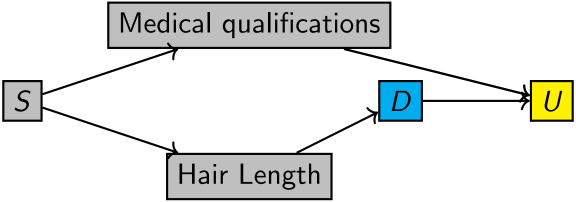
# Why do I think this definition is interesting?

- The definition is intuitive and gives intuitive results in concrete cases.
- The definition does not rely on conceptually problematic interventions.

# Why do I think this definition is interesting?

- The definition is intuitive and gives intuitive results in concrete cases.
- The definition does not rely on conceptually problematic interventions.
- Formalizes the notion of a fair label.

# Limitations (Discussion)




$U$ is {Medical qualifications}-VoI-fair.

# Bibliography

Everitt, Tom, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg (2021). "Agent Incentives: A Causal Perspective". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35.

Howard, Ronald A. (1966). "Information Value Theory". In: *IEEE Transactions on Systems Science and Cybernetics* 2.

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva (2017). "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Thank you!