Supplementary Materials: Adjusting Machine Learning Decisions for Equal Opportunity and Counterfactual Fairness

A Proof of Theorem 1

Before proving Thm. 1 about the eco-fairness and the optimality of the eco decision-maker, we first prove a lemma about the eco-fairness, which will be useful for proving Thm. 1. It says that a decision-maker $\hat{Y}$ is eco-fair if and only if there is no causal arrow from A to $\hat{Y}$.

Lemma 3. *(eco $\Leftrightarrow$ No $A \to \hat{Y}$) Assume the causal graph in Fig. 2. A decision-maker $\hat{Y}$ satisfies equal opportunities over A if and only if there is no causal arrow between A and $\hat{Y}$.*

The intuition behind Lemma 3 is that all decisions $(\hat{Y}^{\text{ml}}, \hat{Y}^{\text{eco}}, \hat{Y}^{\text{cf}})$ in the causal model Fig. 2 satisfy $\hat{Y}(a, s) \perp A, S$ (Pearl, 2009). Therefore, the eco criterion (Def. 1) reduces to $\hat{Y}(a, s)$ being constant in a. In other words, $\hat{Y}$ does not use the protected attribute to make a decision. Below we prove Lemma 3.

Proof. The goal is to show that eco (Def. 1) is equivalent to

$$\hat{Y}(a, s) \overset{d}{=} \hat{Y}(a', s) \overset{d}{=} \hat{Y}(s) \tag{3}$$

for any $s \in \mathcal{S}$ and $a, a' \in \mathcal{A}$. This equation is equivalent to no causal arrow between A and $\hat{Y}$.

Begin with the definition of eco-fairness,

$$\hat{Y}(a', s) \,|\, A = a, S = s \overset{d}{=} \hat{Y}(a, s) \,|\, A = a, S = s \tag{4}$$

$$\Leftrightarrow \hat{Y}(a', s) \overset{d}{=} \hat{Y}(a, s) \qquad \forall s, a, a' \tag{5}$$

$$\Leftrightarrow \hat{Y}(a', s) \overset{d}{=} \hat{Y}(s). \tag{6}$$

Eq. 5 is due to the observation that $\hat{Y}(a, s) \perp S, A$ in Fig. 2 (Pearl, 2009). Eq. 6 is true because

$$P(\hat{Y}(s)) = \int P(\hat{Y}(A, s)) P(A) \, \mathrm{d}A = P(\hat{Y}(a', s)) \tag{7}$$

for any a', where the last equality of Eq. 7 relies on Eq. 5. Eq. 6 is equivalent to no causal arrow from A to $\hat{Y}$.

□

Given Lemma 3, we next prove Thm. 1.

Proof. The first part of Thm. 1 is that $\hat{Y}^{\text{eco}}$ is eco-fair. It is an immediate consequence of Lemma 3 because $\hat{Y}^{\text{eco}}(a_{\text{new}}, s_{\text{new}}) = \hat{Y}^{\text{eco}}(s_{\text{new}})$ as is defined in Eq. 1.

The second part of Thm. 1 establishes the optimality of $\hat{Y}^{\text{eco}}$:

$$\hat{Y}^{\text{ECO}} = \underset{Y^{\text{ECO}} \in \mathcal{Y}^{ECO}}{\arg\min} \, \mathbb{E}_{P(A)P(S)} \left[\mathrm{KL}(P(\hat{Y}^{\text{ml}}(A, S)) || P(Y^{\text{eco}}(A, S))) \right].$$

Begin by rewriting the goal of the proof. We will show this goal is equivalent to the definition of the eco decision-maker,

$$\underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \, \mathbb{E}_{P(A)P(S)} \left[\mathrm{KL}(P(\hat{Y}^{\text{ml}}(A, S)) || P(Y^{\text{eco}}(A, S)) \right]$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \, \mathbb{E}_{P(S)} \left[\int P(A) \int P(\hat{Y}^{\text{ml}}(A, S)) \left[\log P(\hat{Y}^{\text{ml}}(A, S)) - \log P(Y^{\text{eco}}(A, S)) \right] \mathrm{d}Y \, \mathrm{d}A \right] \tag{8}$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \, -\mathbb{E}_{P(S)} \left[\int P(A) \int P(\hat{Y}^{\text{ml}}(A, S)) \log P(Y^{\text{eco}}(A, S)) \, \mathrm{d}Y \, \mathrm{d}A \right] \tag{9}$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\int P(A) \int P(\hat{Y}^{\text{ml}}(A,S)) \log P(Y^{\text{eco}}(S)) \, \mathrm{d}Y \, \mathrm{d}A \right] \tag{10}$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\int \left[\int P(A) P(\hat{Y}^{\text{ml}}(A,S)) \, \mathrm{d}A \right] \log P(Y^{\text{eco}}(S)) \, \mathrm{d}Y \right] \tag{11}$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\mathbb{E} \left[\int \left[\int P(A) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A \right] \log P(Y^{\text{eco}}(s)) \, \mathrm{d}Y \mid S = s \right] \right] \tag{12}$$

$$= \underset{Y^{\text{eco}} \in \mathcal{Y}^{\text{eco}}}{\arg\min} \; \mathbb{E}_{P(S)} \left[\mathbb{E} \left[\mathrm{KL}(\int P(A) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A || P(Y^{\text{eco}}(s)) \mid S = s \right] \right] \tag{13}$$

$$= \int P(\hat{Y}^{\text{ml}}(A,s)) P(A) \, \mathrm{d}A \tag{14}$$

$$= \hat{Y}^{\text{eco}}(a,s) \tag{15}$$

Eq. 8 is due to the definition of the kl divergence. Eq. 9 is due to $\hat{Y}^{\text{ml}}$ being a given random variable. Eq. 10 is due to Lemma 3. Eq. 11 switches the integral subject to conditions of the dominated convergence theorem. Eq. 12 is due to Fig. 2 and the tower property. Eq. 13 is due to the definition of the kl divergence and $\hat{Y}^{\text{ml}}$ being given. Eq. 14 is because setting $P(Y^{\text{eco}}(s)) = \int P(A) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A$ has $\mathrm{KL}(\int P(A) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A || P(Y^{\text{eco}}(s)) = 0$ for all s. The expectation is also zero: $\mathbb{E}_{P(S)} \left[\mathrm{KL}(P(\int P(A) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A || P(Y^{\text{eco}}(S)) \right] = 0$.

This calculation implies $\hat{Y}^{\text{eco}} = \int P(\hat{Y}^{\text{ml}}(A,s)) P(A) \, \mathrm{d}A$ minimizes the average kl distance between the ML decision and the eco decision. In other words, the eco decision-maker maximally recovers the ML decision. Put differently, the eco decision-maker minimally modifies the ML decision to achieve eco-fairness.

□

The optimality of ftu under the population $P(A,S)$. We next present a supplementary result of Thm. 1 that establishes the optimality of $\hat{Y}^{\text{FTU}}$ under the same KL divergence objective but the population $P(A,S)$:

$$\hat{Y}^{\text{FTU}} = \underset{Y^{\text{FTU}} \in \mathcal{Y}^{FTU}}{\arg\min} \; \mathbb{E}_{P(A,S)} \left[\mathrm{KL}(P(\hat{Y}^{\text{ml}}(A,S)) || P(Y^{\text{FTU}}(A,S))) \right].$$

Begin by rewriting the goal of the proof. We will show this goal is equivalent to the definition of the ftu decision-maker,

$$\underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; \mathbb{E}_{P(A,S)} \left[\mathrm{KL}(P(\hat{Y}^{\text{ml}}(A,S)) || P(Y^{\text{FTU}}(A,S)) \right]$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; \mathbb{E}_{P(S)} \left[\int P(A \mid S) \int P(\hat{Y}^{\text{ml}}(A,S)) \left[\log P(\hat{Y}^{\text{ml}}(A,S)) - \log P(Y^{\text{FTU}}(A,S)) \right] \mathrm{d}Y \, \mathrm{d}A \right] \tag{16}$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\int P(A \mid S) \int P(\hat{Y}^{\text{ml}}(A,S)) \log P(Y^{\text{FTU}}(A,S)) \, \mathrm{d}Y \, \mathrm{d}A \right] \tag{17}$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\int P(A \mid S) \int P(\hat{Y}^{\text{ml}}(A,S)) \log P(Y^{\text{FTU}}(S)) \, \mathrm{d}Y \, \mathrm{d}A \right] \tag{18}$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\int \left[\int P(A \mid S) P(\hat{Y}^{\text{ml}}(A,S)) \, \mathrm{d}A \right] \log P(Y^{\text{FTU}}(S)) \, \mathrm{d}Y \right] \tag{19}$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; -\mathbb{E}_{P(S)} \left[\mathbb{E} \left[\int \left[\int P(A \mid S) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A \right] \log P(Y^{\text{FTU}}(s)) \, \mathrm{d}Y \mid S = s \right] \right] \tag{20}$$

$$= \underset{Y^{\text{FTU}} \in \mathcal{Y}^{\text{FTU}}}{\arg\min} \; \mathbb{E}_{P(S)} \left[\mathbb{E} \left[\mathrm{KL}(\int P(A \mid S) P(\hat{Y}^{\text{ml}}(A,s)) \, \mathrm{d}A || P(Y^{\text{FTU}}(s)) \mid S = s \right] \right] \tag{21}$$

$$= \int P(\hat{Y}^{\text{ml}}(A,s)) P(A \mid S) \, \mathrm{d}A \tag{22}$$

$$= \hat{Y}^{\mathrm{FTU}}(a, s) \tag{23}$$

Eq. 16 is due to the definition of the kl divergence. Eq. 17 is due to $\hat{Y}^{\mathrm{ml}}$ being a given random variable. Eq. 18 is due to Lemma 3. Eq. 19 switches the integral subject to conditions of the dominated convergence theorem. Eq. 20 is due to Fig. 2 and the tower property. Eq. 21 is due to the definition of the kl divergence and $\hat{Y}^{\mathrm{ml}}$ being given. Eq. 22 is because setting $P(Y^{\mathrm{eco}}(s)) = \int P(A)P(\hat{Y}^{\mathrm{ml}}(A, s))\,\mathrm{d}A$ has $\mathrm{KL}(\int P(A)P(\hat{Y}^{\mathrm{ml}}(A, s))\,\mathrm{d}A || P(Y^{\mathrm{FTU}}(s)) = 0$ for all s. The expectation is also zero: $\mathbb{E}_{P(S)}\left[\mathrm{KL}(P(\int P(A)P(\hat{Y}^{\mathrm{ml}}(A, s))\,\mathrm{d}A || P(Y^{\mathrm{FTU}}(S))\right] = 0$.

This calculation implies $\hat{Y}^{\mathrm{FTU}} = \int P(\hat{Y}^{\mathrm{ml}}(A, s))P(A \mid S = s)\,\mathrm{d}A = P(Y \mid S = s)$ minimizes the average kl distance between the ML decision and the eco decision. In other words, the eco decision-maker maximally recovers the ML decision. Put differently, the eco decision-maker minimally modifies the ML decision to achieve eco-fairness.

B Proof of Theorem 2

Proof. We first prove that the cf decision-maker is cf-fair.

Recall the definition of the cf decision-maker,

$$P(\hat{Y}^{\mathrm{cf}}(a_{\mathrm{new}}, s_{\mathrm{new}})) = \int\int P(\hat{Y}^{\mathrm{eco}}(S(\bar{A})))\ P(S(\bar{A}) \mid A = a_{\mathrm{new}}, S = s_{\mathrm{new}})\ P(\bar{A})\,\mathrm{d}S(\bar{A})\,\mathrm{d}\bar{S}. \tag{24}$$

Note the above equation is the same definition as in Eq. 2 except that the notation a is changed to $\bar{A}$ to distinguish against the notation A used later in the proof.

Following this definition, we have

$$\begin{aligned}
&P(\hat{Y}^{\mathrm{cf}}(a', S(a')) \mid A = a, S = s) && (25)\\
&= \int\int P(\hat{Y}^{\mathrm{eco}}(S(\bar{A})))\ P(S(\bar{A}) \mid A = a', S = S(a')) && (26)\\
&\quad \times P(\bar{A})P(S(a') \mid A = a, S = s)\,\mathrm{d}S(a')\,\mathrm{d}S(\bar{A})\,\mathrm{d}\bar{A} && (27)\\
&= \int\int P(\hat{Y}^{\mathrm{eco}}(S(\bar{A})))\ P(S(\bar{A}) \mid A = a, S = s)\ P(\bar{A})\,\mathrm{d}S(\bar{A})\,\mathrm{d}\bar{A}. && (28)
\end{aligned}$$

Eq. 28 is due to the structural equation representation of the causal graph $S \overset{a.s.}{=} f(A, \epsilon)$, which implies $\int P(S(\bar{A}) \mid A = a', S = S(a'))P(S(a') \mid A = a, S = s)\,\mathrm{d}S(a') = P(S(\bar{A}) \mid A = a, S = s)$. The intuition here is that the observed A, S will provide the same information as $a', S(a')$ for the same person. They all contain the information about the background variable that affects S,

$$P(S(\bar{A}) \mid S(a'), s, a) = P(S(\bar{A}) \mid A = a', S = S(a') = P(S(\bar{A}) \mid S(a'), a', s, a).$$

Notice that the right hand side of Eq. 28 does not depend on a', which implies that

$$P(\hat{Y}^{\mathrm{cf}}(a', S(a')) \mid A = a, S = s) = P(\hat{Y}^{\mathrm{cf}}(a, S(a)) \mid A = a, S = s).$$

This shows that the cf decision-maker is cf-fair, hence establishes the first part of Thm. 2.

We then prove demographic parity for $\hat{Y}^{\mathrm{cf}}(A, S)$. First we notice that

$$\begin{aligned}
&\int P(\hat{Y}^{\mathrm{cf}}(A, S) \mid A)P(S \mid A)\,\mathrm{d}S && (29)\\
&\qquad = \int P(\hat{Y}^{\mathrm{eco}}(S(A'))P(S(A') \mid A, S)P(A')\,\mathrm{d}S(A')\,\mathrm{d}A'P(S \mid A)\,\mathrm{d}S && (30)
\end{aligned}$$

$$= \int P(\hat{Y}^{\text{eco}}(S(A'))P(S(A') \mid A)P(A') \,\mathrm{d}S(A') \,\mathrm{d}A' \tag{31}$$

$$= \int P(\hat{Y}^{\text{eco}}(S(A'))P(S(A'))P(A') \,\mathrm{d}S(A') \,\mathrm{d}A'. \tag{32}$$

The third equality is due to $S(A') \perp A$ by applying the Pearl's twin network strategy (Pearl, 2009). We then compute the marginal distribution of $\hat{Y}^{\text{cf}}(A, S)$.

$$\int P(\hat{Y}^{\text{cf}}(A, S))P(A, S) \,\mathrm{d}A \,\mathrm{d}S \tag{33}$$

$$= \int \int P(\hat{Y}^{\text{eco}}(S(A'))P(S(A') \mid A, S)P(A') \,\mathrm{d}S(A') \,\mathrm{d}A' P(A, S) \,\mathrm{d}S \,\mathrm{d}A \tag{34}$$

$$= \int P(\hat{Y}^{\text{eco}}(S(A'))P(S(A'))P(A') \,\mathrm{d}S(A') \,\mathrm{d}A', \tag{35}$$

which implies $P(\hat{Y}^{\text{cf}}(A, S) \mid A) = P(\hat{Y}^{\text{cf}}(A, S))$, i.e., demographic parity $\hat{Y}^{\text{cf}}(A, S) \perp A$.

We finally prove the second part of Thm. 2. We show that the cf decision-maker $\hat{Y}^{\text{cf}}$ minimizes the average kl distance between the eco decision and the cf decision $\mathbb{E}_{P(S,A)}\left[\mathrm{KL}(P(\hat{Y}^{\text{eco}}(A, S))||P(Y^{\text{cf}}(A, S)))\right]$. In other words, the cf decision-maker maximally recovers the eco decision. Put differently, the cf decision-maker minimally modifies the eco decision to achieve cf-fairness.

To establish this optimality of the cf decision-maker, we first notice that the cf criterion implies that any cf decision-maker must take the form of $Y^{\text{cf}}(\epsilon)$, where $S = f(A, \epsilon)$ is the structural equation of the causal model in Fig. 2. We then compute the minimizer of the kl distance.

$$\begin{aligned}
&\arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} \mathbb{E}_{P(A)P(S)}\left[\mathrm{KL}(P(\hat{Y}^{\text{eco}}(A, S))||P(Y^{\text{cf}}(A, S))\right] \\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} \mathbb{E}_{P(S)}\left[\int P(A) \int P(\hat{Y}^{\text{eco}}(A, S))\left[\log P(\hat{Y}^{\text{eco}}(A, S)) - \log P(Y^{\text{cf}}(A, S))\right] \mathrm{d}Y \,\mathrm{d}A\right] && (36)\\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} -\mathbb{E}_{P(S)}\left[\int P(A) \int P(\hat{Y}^{\text{eco}}(A, S)) \log P(Y^{\text{cf}}(A, S)) \,\mathrm{d}Y \,\mathrm{d}A\right] && (37)\\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} -\mathbb{E}_{P(S)}\left[\int P(A) \int P(\hat{Y}^{\text{eco}}(f(A, \epsilon))) \log P(Y^{\text{cf}}(\epsilon)) \,\mathrm{d}Y \,\mathrm{d}A\right] && (38)\\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} -\mathbb{E}_{P(S)}\left[\int \left[\int P(A')P(\hat{Y}^{\text{eco}}(f(A', \epsilon))) \,\mathrm{d}A'\right] \log P(Y^{\text{cf}}(\epsilon)) \,\mathrm{d}Y\right] && (39)\\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} -\mathbb{E}_{P(S)}\left[\mathbb{E}\left[\int \left[\int P(A')P(\hat{Y}^{\text{eco}}(f(A', \epsilon))) \,\mathrm{d}A'\right] \log P(Y^{\text{cf}}(s)) \,\mathrm{d}Y \mid A = a, S = s\right]\right] && (40)\\
&= \arg\min_{Y^{\text{cf}} \in \mathcal{Y}^{\text{cf}}} \mathbb{E}_{P(S)}\left[\mathbb{E}\left[\mathrm{KL}(\int P(A')P(\hat{Y}^{\text{eco}}(f(A', \epsilon))) \,\mathrm{d}A'||P(Y^{\text{cf}}(s)) \mid A = a, S = s\right]\right] && (41)\\
&= \int \int P(\hat{Y}^{\text{eco}}(f(A', \epsilon)))P(A')P(\epsilon \mid a, s) \,\mathrm{d}A' \,\mathrm{d}\epsilon && (42)\\
&= \int \int P(\hat{Y}^{\text{eco}}(S(A'))\, P(S(A') \mid A = a, S = s)p(A') \,\mathrm{d}S(A') \,\mathrm{d}A' && (43)\\
&= \hat{Y}^{\text{cf}}(a, s) && (44)
\end{aligned}$$

Eq. 36 is due to the definition of the kl divergence. Eq. 37 is due to $\hat{Y}^{\text{eco}}$ being a given random variable. Eq. 38 is due to the observation that eco decision must only depend on $S = f(A, \epsilon)$ and cf decisions must only depend on ϵ. Eq. 39 switches the integral subject to conditions of the dominated convergence theorem. Eq. 40 is due to Fig. 2 and the tower property. Eq. 41 is due to the definition of the kl divergence and $\hat{Y}^{\text{eco}}$ being given. Eq. 42 is because setting $P(Y^{\text{cf}}(\epsilon)) = \int P(A')P(\hat{Y}^{\text{eco}}(f(A', \epsilon))) \,\mathrm{d}A'$ has $\mathrm{KL}(\int P(A')P(\hat{Y}^{\text{eco}}(f(A', \epsilon))) \,\mathrm{d}A'||P(Y^{\text{cf}}(\epsilon)) =$

0 for all ϵ. The expectation is also zero: $\mathbb{E}\left[\mathrm{KL}\left(\int P(A')P(\hat{Y}^{\mathrm{eco}}(f(A',\epsilon)))\,\mathrm{d}A' || P(Y^{\mathrm{cf}}(\epsilon)\right)\right] = 0$. Eq. 43 rewrites the ϵ in terms of A and S. Eq. 44 is due to the definition of the cf decision-maker.

In addition, we show that $\hat{Y}^{\mathrm{cf}}$ recovers the marginal distribution of $\hat{Y}^{\mathrm{eco}}$,

$$\int P(\hat{Y}^{\mathrm{cf}}(A,S))P(A,S)\,\mathrm{d}A\,\mathrm{d}S \tag{45}$$

$$= \int P(\hat{Y}^{\mathrm{eco}}(S(A'))P(S(A') \mid A,S)P(A,S)P(A')\,\mathrm{d}S(A')\,\mathrm{d}A'\,\mathrm{d}A\,\mathrm{d}S \tag{46}$$

$$= \int P(\hat{Y}^{\mathrm{eco}}(S)P(A,S)\,\mathrm{d}S(A)\,\mathrm{d}A\,\mathrm{d}S \tag{47}$$

$$= \int P(\hat{Y}^{\mathrm{eco}}(S))P(S)\,\mathrm{d}S \tag{48}$$

$$= \int P(\hat{Y}^{\mathrm{eco}})P(A,S)\,\mathrm{d}A\,\mathrm{d}S. \tag{49}$$

The second equality is because $P(A) = P(A')$ by the construction of the cf decision-maker.

This calculation shows that the cf decision-maker $\hat{Y}^{\mathrm{cf}}$ preserves the marginal distribution of $\hat{Y}^{\mathrm{eco}}$. □

C The correctness of Algorithm 1

We prove the identifiability of the eco and cf decision given observational data, which establishes the correctness of Alg. 1.

For the eco decision-maker, we utilize the backdoor adjustment formula (Pearl, 2009).

Proposition 4. *Assume the causal graph Fig. 2. The eco decision-maker can be computed using the observed data* (A_i, S_i, Y_i)*,*

$$P(\hat{Y}^{\mathrm{ECO}}(a_{\mathrm{new}}, s_{\mathrm{new}})) = \int P(Y_i \mid A_i, S_i = s_{\mathrm{new}})P(A_i)\,\mathrm{d}A_i, \tag{50}$$

if $P(S_i \in \mathbf{S} \mid A_i = a) > 0$ *for all* $a \in \mathcal{A}$ *and sets* $\mathbf{S}$ *such that* $P(\mathbf{S}) > 0$ *and* $\mathbf{S} \subset \mathcal{S}$*.*

Proof. Proposition 4 is a direct consequence of Theorem 3.3.2 of Pearl (2009). □

Eq. 50 reiterates the fact that $\hat{Y}^{\mathrm{ECO}}$ does not simply ignore the protected attribute A; estimating $P(Y_i \mid A_i, S_i)$ relies on A in the training data. However, $\hat{Y}^{\mathrm{ECO}}(a_{\mathrm{new}}, s_{\mathrm{new}})$ does not rely on the protected attribute a_{new} in predicting on the test data.

For the cf decision-maker, we adopt the abduction-action-prediction approach (Pearl, 2009).

Proposition 5. *Assume the causal graph Fig. 2. Write* $S \overset{a.s}{=} f_S(A, \epsilon_S)$ *for some function* f_S *and zero mean random variable* ϵ_S *satisfying* $S \perp \epsilon_S$*. The cf decision-maker can be computed using the observed data* (A_i, S_i, Y_i)*:*

$$P(\hat{Y}^{\mathrm{cf}}(a,s)) = \int P(\hat{Y}^{\mathrm{eco}}(a,s')) \cdot P(S = s' \mid A = a', \epsilon_S) \tag{51}$$

$$\times\, P(\epsilon_S \mid A = a, S = s) \cdot P(A = a')\,\mathrm{d}s'\,\mathrm{d}a' \tag{52}$$

$$\approx \int P\left(\hat{Y}^{\mathrm{eco}}\left(\mathbb{E}\left[S_i \mid A_i = a'\right] + \hat{\epsilon}_S\right), a\right) \cdot P(A = a')\,\mathrm{d}a', \tag{53}$$

where $\hat{\epsilon}_S = s - \mathbb{E}\left[S_i \mid A_i = a\right]$*.*

Proof. Eq. 52 is a direct consequence of Theorem 7.1.7 of Pearl (2009). Eq. 53 is due to a linear approximation of f_S. □

D Datasets in empirical studies

Each dataset in empirical studies consists of training and test sets, potentially with multiple protected and unprotected attributes.

The simulated college admissions data in Fig. 1 comes from the structural model in § 4. We set $\beta_s = 2.0$ and $\beta_a = 1.0$ so that for the same test score the committee is more likely to admit males than females. We set $\lambda = 2.0$ to model that females perform less well on the test.

The case studies involve three public datasets.

- **Adult income data (Dua and Graff, 2017a)** The task is to decide whether an individual is loan worthy; the decision is binary. The data do not initially contain decisions about loan worthiness; instead, each individual's income is provided. To evaluate the algorithms, we produce a decision about loan worthiness by thresholding income; individuals that make more than \$50K are decided as being loan worthy. The protected attributes are gender and race; other attributes include education and marital status. The dataset has 32,561 training samples and 16,282 test samples.
- **ProPublica's COMPAS recidivism data.** The task is to decide an individual's risk score of recidivating; the decision is real-valued. The protected attributes are gender and race; other attributes include the priors count, juvenile felonies count, and juvenile misdemeanor count. The dataset has 6,907 complete samples. We split them into 75% training and 25% testing.
- **German credit data. (Dua and Graff, 2017b)** The task is again to decide whether an individual is loan worthy. In this setting, each individual is initially labeled as having good or bad credit; we use this covariate to produce a decision. Individuals with good credit are decided as being loan worthy. The protected attributes are gender and marital status; other attributes include credit history, savings, and employment history. The dataset has 1,000 samples. We split them into 75% training and 25% testing.

E Implementation of decision-makers in the empirical studies

We implement classical ML decision-makers with linear/logistic regression using all attributes (including the protected ones). The ftu decision-makers perform regression but omit the protected attributes. The eco and cf decision-makers adjust the classical ML decision-makers using Alg. 1. FairLearning infers background causal variables by fitting a linear regression of A against S and compute the residuals $\epsilon_i = S_i - \mathbb{E}[S_i \mid A_i]$; it then performs linear/logistic regression against the background variables. (This is the same procedure used in the abduction step for calculating f_{cf}.) With these decision-makers, we make algorithmic decisions $\hat{y}$ about the test set units.

F Additional empirical studies

In the adult income data and COMPAS data, race and gender are the protected attributes. In the German credit data, gender and marital status are protected attributes. The decisions about loan worthiness are binary (as in the admissions example); the decisions in the COMPAS data are real-valued assessments of the recidivism score.

Tables 1 and 2 present detailed results of the eco and cf decision-makers on the COMPAS and the German credit datasets.

	Metrics ($\times 10^{-2}$) on COMPAS			
	ρ_{eco}	ρ_{cf}	**KL**	**Prediction**
ML predictor f_{ml}	-68.1(42.1)	-104.9(59.7)	17.1	28.0
ftu	**0(0)**	-52.9(30.1)	0.5	**25.6**
eco predictor f_{eco}	**0(0)**	-36.8(21.2)	0.5	**25.6**
FairLearning	-41.2(28.9)	**0(0)**	**0.2**	**22.7**
cf predictor f_{cf}	-41.2(29.4)	**0(0)**	**0.2**	**22.7**

Table 1: Both the eco predictor f_{eco} and ftu are eco-fair; they achieve zero in the eco metric (lower is better). cf predictor f_{cf} and FairLearning (Kusner et al., 2017) are cf-fair; they achieve zero in the cf metric (lower is better). cf predictor f_{cf} achieves demographic parity; it has close-to-zero kl divergence (Lower is better.) ML predictor f_{ml} predicts best; eco predicts best among the fair predictors (higher prediction scores are better.) We report mean values across individuals with the standard deviation in parentheses. eco and cf metric standard deviations are ≤ 0.42 and ≤ 0.6, respectively. We also report the KL divergence and prediction accuracy, which are both distributional metrics.

	Metrics ($\times 10^{-2}$) on German Credit			
	ρ_{eco}	ρ_{cf}	**KL**	**Prediction**
ML predictor f_{ml}	-5.4(2.5)	-3.9(1.8)	18.6	64.7
ftu	**0(0)**	-2.0(1.0)	15.6	63.4
eco predictor f_{eco}	**0(0)**	1.5(0.7)	13.2	**64.5**
FairLearning	-1.4(0.7)	**0(0)**	8.3	63.5
cf predictor f_{cf}	-1.3(0.6)	**0(0)**	**7.8**	**64.3**

Table 2: Both eco predictor f_{eco} and ftu are eco-fair; they achieve zero in the eco metric (lower is better). cf predictor f_{cf} and FairLearning (Kusner et al., 2017) are cf-fair; they achieve zero in the cf metric (lower is better). cf predictor f_{cf} achieves demographic parity; it has close-to-zero kl divergence (Lower is better.) ML predictor f_{ml} predicts best; eco predicts best among the fair predictors (higher prediction scores are better.) We report mean values across individuals with the standard deviation in parentheses. eco and cf metric standard deviations are ≤ 0.02 and ≤ 0.02, respectively. We also report the KL divergence and prediction accuracy, which are both distributional metrics.

References

Dheeru Dua and Casey Graff. 2017a. UCI Machine Learning Repository: Adult Data Set. https://www.kaggle.com/wenruliu/adult-income-dataset

Dheeru Dua and Casey Graff. 2017b. UCI Machine Learning Repository: Statlog (German Credit Data) Data Set. https://www.kaggle.com/uciml/german-credit

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Neural Information Processing Systems* (2017).

Judea Pearl. 2009. *Causality*. Cambridge University Press.