Adjusting Machine Learning Decisions for Equal Opportunity and Counterfactual Fairness

Anonymous authors
Paper under double-blind review

Abstract

Machine learning (ML) methods have the potential to automate high-stakes decisions, such as bail admissions or credit lending, by analyzing and learning from historical data. But these algorithmic decisions may be unfair: in learning from historical data, they may replicate discriminatory practices from the past. In this paper, we propose two algorithms that adjust fitted ML predictors to produce decisions that are fair. Our methods provide post-hoc adjustments to the predictors, without requiring that they be retrained. We consider a causal model of the ML decisions, define fairness through counterfactual decisions within the model, and then form algorithmic decisions that capture the historical data as well as possible, but are provably fair. In particular, we consider two definitions of fairness. The first is "equal counterfactual opportunity," where the counterfactual distribution of the decision is the same regardless of the protected attribute; the second is counterfactual fairness. We evaluate the algorithms, and the trade-off between accuracy and fairness, on datasets about admissions, income, credit, and recidivism.

1 Introduction

There is growing interest in using machine learning (ML) methods to automate important decisions about people by analyzing and learning from historical data. For example, in the criminal justice system, ML algorithms are routinely used to assess a defendant's risk of recidivism to inform pretrial release and parole decisions (Brennan et al., 2009; Larson et al., 2016; Lin et al., 2020). ML algorithms were also used by a university to help determine which college applicants should be reviewed (Waters & Miikkulainen, 2014).

As a running example, consider an admissions committee that decides which applicants to accept to a university. Given historical data about applicants' traits and the admission decisions, an ML algorithm can learn to predict who will be admitted and who will not, and the resulting ML decision-maker will accurately simulate the committee's decisions. But while these algorithmic decisions would save time and effort, they will also inherit some of the undesirable properties of the admissions committee (Corbett-Davies et al., 2017; Mitchell et al., 2021). If the committee was unfairly biased then its ML replacement will be as well.

In this paper, we develop two methods that adjust existing ML decision-makers to produce algorithmic decisions that are both accurate and fair. We prove that these algorithmic decisions maintain as much fidelity as possible to the decisions of the historical committee, but are adjusted to correct for the biases inherent in past decisions. We demonstrate these algorithmic decisions on datasets about income, credit, and recidivism.

A key contribution of the proposed methods is that they do not require refitting the ML method to data. One can train an ML decision-maker as usual, but then use the proposed methods to adjust the resulting algorithmic decisions to be fair. Consequently, these methods can be employed by practitioners who have a pre-trained ML model but without the original data, or when the ML model is too expensive to refit.

Main idea. To quantify the fairness of a decision maker, we use two causal criteria. Causal criteria for fairness were first introduced in Kusner et al. (2017), and have been operationalized in many algorithms (Zhang & Bareinboim, 2018a;b; Zhang et al., 2017; Wu et al., 2019a;b; Coston et al., 2020; Kilbertus et al., 2017; 2019; Chiappa, 2019). Causal fairness criteria require encoding our assumptions about bias using causal

models. While these assumptions require careful scrutiny (Kohler-Hausmann, 2018), causal models enable us to derive fairness measures to capture different mechanisms of bias. In contrast to causal criteria, statistical measures of fairness cannot distinguish between different mechanisms of unfairness (Hardt et al., 2016).

First, we introduce the equal counterfactual opportunity (ECO) criterion. With an ECO decision-maker, individuals that have the same attributes will receive the same probability of admission, regardless of their demographic group. In particular, for an algorithmic decision that satisfies ECO, there is no causal relationship between the protected attributes (such as race or gender) and the decision. The ECO criterion is individual-level; it requires that each decision is fair.

Second, we consider the counterfactual fairness (CF) criterion proposed in Kusner et al. (2019). Under CF, the probability of admission is not affected by (hypothetical) interventions on protected attributes, including when we account for how those interventions might subsequently change the other attributes. For example, minority applicants may have been historically denied university admission due to lower test scores, but these low scores might be consequences of systemic disadvantages that the minority applicants face. A CF decision-maker adjusts an individuals' attributes (such as test score) to account for such disadvantages. Similar to the ECO criterion, the CF is also individual-level.

With these two criteria in place, the methods we develop draw on the ideas and methods of causal inference (Pearl, 2009; Peters et al., 2017). First we require that the historical decision-making process is captured by a causal model. This model reflects domain knowledge and assumptions about the process, including the probabilistic relationship between an individual's attributes and the decision, and relationships between the attributes themselves. The causal model might be elicited from experts, informed from data, or a combination.

Continuing the example, a causal model of admissions represents the applicants' traits and the committee's decisions. In this paper we use the model illustrated in Fig. 2. There are protected attributes, such as gender, and non-protected attributes, such as test score. The model posits that the (historical) decision was affected by both, e.g., the committee considered the test score but also unfairly considered gender. The model also posits that the non-protected attributes may be affected by the protected ones, e.g., gender may affect the test score through systematic gender bias in schooling. The methods we develop do not require this particular causal model, but it will be a running example.

Given the causal model of historical decisions, we then frame *algorithmic* decisions as new causal variables in the same model, ones whose values are functions of the individuals' attributes. This framing is important because it allows us to consider counterfactuals about the algorithms, e.g., "what would the algorithm decide for this applicant if she was male and had achieved the same test score?" Probabilistic properties of such counterfactuals help formally define the mathematical criteria for ECO and CF.

Finally, we derive methods to produce algorithmic decisions that meet these criteria. The first step is to use the historical data to fit an ML method to predict the decision, such as with a simple logistic regression or a neural network. (Note we can then use the causal model and causal criteria to assess the fairness of the ML decision-maker.) We then show how to adjust the trained ML decision-maker to satisfy ECO, and we show how to adjust the ECO decision-maker to exercise CF. We will see that both adjustment procedures involve estimating causal quantities based on both the assumed causal model and the original ML predictions.

We emphasize again that this method uses a fitted ML decision-maker but does not require retraining it or re-analyzing the historical data, which could require significant computational cost. We prove that, under the assumed causal model, the fair decision-makers we derive are theoretically optimal—they maintain as much fidelity to the historical data as possible, while still being ECO-fair or CF-fair.

We study these approaches on simulated admissions data and on three public datasets, about income, credit, and recidivism. We find that classical ML decisions are the most accurate, but they may be unfair. The fair decision-makers deviate from the ML decisions, but provide decisions that meet the fairness criteria. Compared to other approaches that satisfy ECO and CF (Chen et al., 2019; Kusner et al., 2017), the method developed here provides decisions with higher accuracy while remaining fair.

Contributions. 1) We develop algorithms that modify existing ML predictors to be fair, without requiring retraining. 2) We prove that these decisions optimally recover ML predictors while satisfying the criteria for

ECO and CF. 3) We conduct empirical studies on real and simulated settings to show that our algorithms produce better decisions than existing approaches that satisfy the same fairness criteria.

Related work. This paper contributes to research about causal models for defining and implementing fair ML. Kusner et al. (2017) introduce counterfactual fairness (the CF criterion above) and the FairLearning algorithm to satisfy it. This paper provides ECO-fair decisions and CF-fair decisions that minimally adjust the fitted ML decisions. Like FairLearning Kusner et al. (2017), the CF-fair decisions also satisfy counterfactual fairness. But in contrast to FairLearning, which omits descendants of the protected attributes, the CF decision-maker in this paper uses all available attributes and achieves higher fidelity to historical decisions.

Kilbertus et al. (2017) considers more fine-grained causal notions of fairness and relates paths and variables in causal models to violations of ECO and CF. Other work studies path-specific fairness, where some paths in the causal model are deemed unfair. These works typically propose new training objectives that satisfy fairness (Chiappa, 2019; Zhang et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018b; Wu et al., 2019b; Mhasawade & Chunara, 2021; Dutta et al., 2021; Madras et al., 2019). In contrast to the methods developed for path-specific fairness, this paper targets individual-level fairness, where each decision is guaranteed to be fair. Moreover, the algorithmic decisions we propose in this paper do not require re-training the ML decision-makers.

A number of recent papers further extend the idea of counterfactual fairness. Kusner et al. (2019); Mishler & Kennedy (2021) develop decision-makers that target multiple decisions or multiple fairness criteria. Coston et al. (2020) extends the use of counterfactuals for fair risk assessment scores. Galhotra et al. (2021); Parafita & Vitria (2021); Dai et al. (2021); Foster et al. (2021); Artelt et al. (2021); Karimi et al. (2021); Veitch et al. (2021) develop algorithms for counterfactual invariance and counterfactual fairness in model explanation and representation learning. Mishler et al. (2021) extends counterfactual fairness to counterfactual equalized odds and propose post-processing optimization programs for adjustment. Unlike these works, this paper focuses on developing post-hoc fairness adjustments with closed forms, without requiring retraining.

Alongside these causal approaches to algorithmic fairness, there is a vast literature on statistical criteria for fairness. We refer the readers to Mitchell et al. (2021); Corbett-Davies & Goel (2018); Pessach & Shmueli (2022); Berk et al. (2021) for surveys of these related ideas.

The literature on fair ML contrasts fairness criteria that apply to each individual and those that apply to groups. The definitions of ECO and CF in this paper are individual-level criteria. Similarly, Dwork et al. (2012) define a notion of ECO without causality called individual fairness. Given a distance metric between individuals, it requires similar individuals receive similar decisions. The ECO-fair decisions (of Eq. 1) recover this requirement without relying on an explicit distance metric, which can be difficult to construct. Speicher et al. (2018); Chouldechova & Roth (2019) study the tradeoffs between group- and invididual-level criteria.

Finally, the method proposed here crucially relies on the theory and algorithms behind causal inference (Pearl, 2009; Peters et al., 2017). In particular, it requires that the causal model is accurate, that the necessary counterfactuals are identifiable, and that they can be estimated. In practice, these assumptions are strong and have the potential to be misused (Kasirzadeh & Smart, 2021; Kohler-Hausmann, 2018). To alleviate some of this burden, subsequent research can extend this method to employ recent ideas about causal fairness under uncertain causal models (Kilbertus et al., 2018; Russell et al., 2017; Kilbertus et al., 2019; Wu et al., 2019a;b; De Lara et al., 2021).

2 Assessing fairness with counterfactuals

Consider automating the admissions process at a university. By using a dataset of past admissions, the goal is to algorithmically compute the admissions decision for new applicants. The dataset contains n applicants, each with attributes and the committee's decision about whether to admit them. Some attributes are deemed protected, such as gender, religion, or ethnicity; others are not protected, such as a grade point average or a score on a standardized test.

To illustrate the ideas, we will consider a simple setting. Suppose there are only two attributes for each applicant, their gender and their score on a test. (For ease of exposition, we consider two attributes and

a binary gender variable, but the methods in this paper easily handle multiple attributes and non-binary gender.) The gender is a protected attribute; the test score is not. Fig. 1 shows an example of such data. In the figure, the first 5,000 applicants are (simulated) historical data—applicants' attributes and the decisions made by a committee. The three applicants A, B, and C are new applicants for whom we will make decisions.

A causal model of the decisions. The methods we develop require a causal model (Pearl, 2009) of the historical admissions process. This model should capture the decision, the inputs to the decision, and the causal relationships among these variables. Consider the model in Fig. 2. It assumes that an applicant's gender S and test score A can affect the admissions decision Y; the admissions committee might unfairly prefer to admit males. It also assumes the applicant's gender S can affect the test score A; female applicants might receive fewer opportunities for test preparation, a disparity that results in lower scores. (There are also other variables in Fig. 2, which we will discuss below.) Note that positing a causal model is a strong assumption; it is important for the user of the methods to consult domain knowledge and provide a plausible model of the historical decisions. Many causal models are special cases of the causal model in Fig. 2, if we have multiple (or multi-dimensional) protected and unprotected attributes. We discuss these cases in § 3.

A causal model implies a distribution of *counterfactuals*, how each variable would change under hypothetical intervention on other variables (Pearl, 2009). (Note this "intervention" is only a mathematical construction, a hypothetical change that helps articulate counterfactuals. It may be about an immutable characteristic (Greiner & Rubin, 2011).) Denote $Y_i(s, a)$ to be the counterfactual decision of the ith applicant when we set their test score to be $A_i = a$ and gender to be $S_i = s$. For example, consider the counterfactual decision of a female applicant had she been a male and had the same test score. A key idea of this paper is that we can use the assumed model to ask counterfactual questions about any decisions, either those made by a committee or those made by an algorithm. We first show how we can interpret fairness through properties of such counterfactual decisions. We then show how to adjust an algorithmic decision-maker, such as one produced by classical ML, to produce decisions that satisfy fairness properties.

ID	Sex	Test	Admit	$\hat{Y}^{\text{ml}}$	$\hat{Y}^{\text{eco}}$	$\hat{Y}^{\text{cf}}$
1	f	54	yes			
2	m	66	no			
⋮	⋮	⋮	⋮			
5000	f	44	no			
A	f	85	?	0.67	0.77	0.78
B	m	85	?	0.84	0.77	0.76
C	f	65	?	0.57	0.69	0.70

Figure 1: Simulated data of past university applicants (1-5000) and their admissions decisions. The (fictional) admissions committee violates ECO and does not enforce CF. For three new applicants (A, B, C), a machine learning classifier trained on this biased data yields unfair decisions, denoted by $\hat{Y}^{\text{ml}}$. For example, a male and female applicant with the same test score receive different admission probabilities, violating ECO. We algorithmically adjust this decision to produce $\hat{Y}^{\text{eco}}$ decisions that are ECO-fair. This decision provides the equally qualified two candidates A and B an equal probability of being admitted. From this $\hat{Y}^{\text{eco}}$ decision, we derive $\hat{Y}^{\text{cf}}$, an algorithmic decision that enforces CF. It increases the probability of admission for applicant C, by accounting for her systematic disadvantage.

The equal counterfactual opportunity criterion. A decision-maker provides equal counterfactual opportunity (ECO) when equally qualified people receive the same decisions regardless of their protected attributes. In the causal model, we say that a decision-maker satisfies ECO if changing the protected attribute S does not change its distribution of the decision Y. Consider an applicant with gender s and test score a. An ECO decision-maker gives the same probability of admission even after changing the gender to s' (but keeping the test score at a). In Fig. 1, candidates A and B have the same test score but different genders; if given equal opportunities, they should receive the same probability of admission.

Definition 1 (Equal counterfactual opportunity (ECO))**.** *A decision Y satisfies ECO in the protected attribute S if, for all possible values of s, s' and a,*

$$Y(s, a) \,|\, \{S = s, A = a\} \stackrel{d}{=} Y(s', a) \,|\, \{S = s, A = a\}.$$

(The notation $Q \stackrel{d}{=} R$ means Q and R are equal in distribution.)

An algorithmic decision that satisfies ECO asserts that there is no causal relationship between the protected attributes S and the decision Y. If a committee or algorithm produces ECO-fair decisions then the causal

model contains no arrow between S and Y; two applicants from different demographic groups, but who are otherwise similarly qualified, will have the same probability of admission.

Other research has posited mathematical criteria to capture the legal notion of equal opportunity. Hardt et al. (2016) proposed the equality of opportunity criterion to evaluate the fairness of decisions predicted by a classifier. Their criterion requires the misclassification rates for the disadvantaged group to be the same as those for the other group. In contrast, the ECO criterion focuses on the individual; it requires that each individual receive the same decision probability regardless of interventions on their protected attributes. Another individual-level criterion appears in Dwork et al. (2012), which defines a notion of equal opportunity (without causality) called individual fairness. Given a distance metric between individuals, it requires that similar individuals receive similar decisions. Algorithmic decisions that satisfy the ECO criterion also recover this requirement, but with a distance metric implied by the causal model.

The counterfactual fairness criterion. The counterfactual fairness (CF) criterion was introduced by Kusner et al. (2017). A decision-maker that targets CF corrects for historical disadvantages that may stem from protected attributes. Female applicants may have fewer opportunities for test preparation, leading to lower test scores and a lower likelihood of admission. Had they been male, they may have had more opportunities for preparation, achieved a higher score, and been admitted.

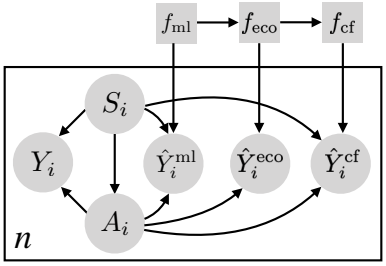


Figure 2: The causal model reflects domain knowledge and assumptions about an admissions decision process: protected attributes (S), such as gender, can affect both the admissions decision (Y) and the remaining attributes (A), such as test scores. Algorithmic decisions ($\hat{Y}^{\text{ml}}, \hat{Y}^{\text{eco}}, \hat{Y}^{\text{cf}}$) are variables in the causal model, produced by fitted algorithms ($f_{\text{ml}}, f_{\text{eco}}, f_{\text{cf}}$) and we can inspect their biases. Classical ML reproduces the historical data it was trained on, inheriting the unfairness of a past admissions committee. ECO decisions adjust ML, eliminating any problematic dependence on the protected attribute. CF decisions further adjust ECO decisions to exercise affirmative action based on the protected attributes.

One of the goals of CF is to ensure that, all else being equal (such as effort or aptitude), an applicant would receive the same decision had they not been in a disadvantaged group. This goal is different from that of ECO because it accounts for how the protected attribute affects the other attributes, how an applicant's gender affects their test score. Moreover, ECO prohibits direct impacts of protected attributes on decisions but allows an indirect influence through unprotected attributes. In contrast, CF prohibits both types of influences, direct or indirect.

Mathematically, counterfactuals help capture this idea. Consider an applicant with test score a and gender s. Had they belonged to a different group s', they may have achieved a different test score $A(s') \,|\, \{S = s, A = a\}$. Notice this counterfactual score is conditional on the attributes of the applicant, including their (factual) score. The variable $A(\text{male}) \,|\, \{S = \text{female}, A = \text{high score}\}$ can capture that a female applicant with a high test score will have an even higher test score had she been male.

The CF criterion further uses a nested counterfactual. The variable $Y(s', A(s')) \,|\, \{S = s, A = a\}$ involves an intervention on the protected attribute along with a resulting change in the unprotected attribute. An CF-fair decision is one where this counterfactual does not change the distribution of the decision.

Definition 2 (Counterfactual fairness (Kusner et al., 2017))**.** *A decision Y exercises counterfactual fairness (CF) if, for all possible values of s, s' and a,*

$$Y(s, A(s)) \,|\, \{S = s, A = a\} \stackrel{d}{=} Y(s', A(s')) \,|\, \{S = s, A = a\}.$$

Algorithmic decisions that satisfy CF are not affected by (hypothetical) interventions on protected attributes, including accounting for how those interventions might subsequently change the other attributes. They use the protected attribute to correct for historical disadvantages.

3 Constructing fair algorithmic decisions

Defs. 1 and 2 provide criteria for fairness in terms of distributions of counterfactual decisions. Using the causal model of the decision-making process, these criteria help evaluate the fairness of its outcomes. The criteria can be used to evaluate decisions produced by a human process, such as an admissions committee, or produced by an algorithm, such as a fitted ML model.

Denote an algorithmic decision as $\hat{Y}$, a causal variable that depends on the attributes $\{s, a\}$. It comes from a fitted probabilistic decision-maker, $\hat{Y} \sim f(s, a)$, where $f(\cdot)$ is a probability density function. For example, a binary admissions decision $\hat{Y} \sim f(s, a)$ is drawn from a Bernoulli that depends on the attributes (e.g., a logistic regression). By considering algorithmic decisions as variables in the causal model, we can infer algorithmic counterfactuals $\hat{Y}(s, a)$ and confirm whether they satisfy the ECO and CF criteria.

Now consider a classical ML model that is fit to emulate historical admissions data. If the historical decisions did not exercise ECO or CF then neither will the decisions produced by ML. Below we develop fair ML decision-makers, algorithms that adjust the decisions made by classical ML to be ECO-fair or CF-fair, i.e., to satisfy the ECO and CF criteria. We will show that these adjusted decisions are as accurate as possible relative to the historical data, while still producing ECO-fair or CF-fair decisions.

Machine learning (ml) decisions. An ML decision-maker uses historical data to accurately predict the decision Y from the protected attribute S and unprotected attribute A. Admissions involves binary decisions, and so logistic regression is a common choice. The decision-maker $f_{\text{ml}}(s, a)$ draws the decision from a Bernoulli, $\hat{Y}^{\text{ml}} \mid \{S = s, A = a\} \sim \text{Bern}(\sigma(\beta_A \cdot a + \beta_S \cdot s + \beta_0))$, where $\sigma(\cdot)$ is the logistic function and the coefficients are fit to maximize the observed data likelihood. When included in the causal model, f_{ml} and $\hat{Y}^{\text{ml}}$ are illustrated in Fig. 2.

The ML decision $\hat{Y}^{\text{ml}}$ will accurately mimic the historical data. But it will also replicate harmful discriminatory practices. If the committee did not give equal opportunities to applicants of different genders then the decision $\hat{Y}^{\text{ml}}$ will violate ECO. If the committee did not correct for the disparate impact of gender on the test score, then $\hat{Y}^{\text{ml}}$ will not exercise CF.

Return to the illustrative simulation (Fig. 1). We fit a logistic regression to the training data, which finds coefficients close to the mechanism that generated the data. Consequently, when used to form algorithmic decisions, it replicates the unfair committee. Consider female applicant A ($s = \text{f}, a = 85$) and male applicant B ($s = \text{m}, a = 85$) and the classical ML decisions for each. For A, her probability of being admitted is 67%. For B, his probability of admission is 84%. Despite identical scores, the female applicant is 17% less likely to get in.

Algorithmic decisions that satisfy eco. We use the ML decision-maker f_{ml} to produce an ECO-fair decision-maker f_{eco}, one whose decisions satisfy ECO. Consider an ML decision-maker $f_{\text{ml}}(s, a)$ and an applicant with attributes $\{s_{\text{new}}, a_{\text{new}}\}$. Her ECO decision is $\hat{Y}^{\text{eco}}(s_{\text{new}}, a_{\text{new}}) \sim f_{\text{eco}}(a_{\text{new}})$, where

$$f_{\text{eco}}(a_{\text{new}}) = \int f_{\text{ml}}(s, a_{\text{new}}) p(s) \, \text{d}s. \tag{1}$$

The ECO decision probability holds the unprotected attribute a_{new} fixed and takes a weighted average of the ML decision-maker for the different values of the protected attribute s. The weights are determined by the proportions of each group in the whole population. Note that the ML decision-maker is fixed; it does not need to be retrained.

Return to Fig. 1. We use the fitted logistic regression $f_{\text{ml}}(s, a)$ to produce the ECO-fair decision-maker. This data has equal numbers of women and men so the weighted average is $f_{\text{eco}}(s, a) = 0.5 f_{\text{ml}}(\text{male}, a) + 0.5 f_{\text{ml}}(\text{female}, a)$. Using ECO-fair decisions, applicants A and B both have a 77% probability of admission.

The ECO decision $\hat{Y}^{\text{eco}} \sim f_{\text{eco}}(a_{\text{new}})$ satisfies the ECO criterion: applicants with the same score will have the same chance of admissions regardless of their gender. The intuition is that the ECO decision preserves the causal relationship between the test score A and the ML decision $\hat{Y}^{\text{ml}}$, but it ignores the possible effect of the protected attribute S. In the *do* notation for interventions (Pearl, 2009), what this means is that

$P(\hat{Y}^{\text{eco}}; \text{do}(s, a)) = P(\hat{Y}^{\text{ml}}; \text{do}(a))$ for all s. The weighted average is the adjustment formula (Pearl, 2009), which calculates $P(\hat{Y}^{\text{ml}}; \text{do}(a))$.

An alternative method to satisfy ECO is fairness through unawareness (FTU) (Kusner et al., 2017). FTU satisfies ECO by fitting an ML model from the unprotected attribute A to the decision Y, and completely omitting the protected attribute S; FTU decisions are also ECO-fair. While both FTU and the ECO decision-makers satisfy the ECO criterion, FTU can have undesirable consequences if an unprotected attribute correlates with the protected attribute. If the past admissions committee discriminated based on gender, then FTU could mimic this discrimination with any other attribute that correlated with gender. In contrast, the ECO decision-maker would not inherit this discrimination pattern. (We illustrate this fact in Fig. 7.) Finally, Thm. 1 shows the theoretical optimality of the ECO decision-makers among all those that satisfy ECO.

Theorem 1 (ECO-fairness and optimality of ECO decisions)**.** *The decision $\hat{Y}^{\text{eco}} \sim f_{\text{eco}}(s, a)$ is ECO-fair. Moreover, among all ECO-fair decisions, $\hat{Y}^{\text{ECO}}$ maximally recovers the ML decision $\hat{Y}^{\text{ml}}$,*

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

where $\mathcal{Y}^{ECO}$ is the set of ECO-fair decisions and the expectation is taken over $P(S)P(A)$. (Proof in App. A.)

Algorithmic decisions that exercise cf. We now show how to adjust the ECO-fair decision-maker to satisfy counterfactual fairness.

Consider the ECO-fair decision-maker $f_{\text{eco}}(a)$ and an applicant with attributes $\{s_{\text{new}}, a_{\text{new}}\}$. Her CF-fair decision is $\hat{Y}^{\text{cf}} \sim f_{\text{cf}}(s_{\text{new}}, a_{\text{new}})$, where

$$f_{\text{cf}}(s_{\text{new}}, a_{\text{new}})) = \iint f_{\text{eco}}(a(s))\, p(a(s) \mid s_{\text{new}}, a_{\text{new}}) p(s)\, \text{d}a(s)\, \text{d}s. \tag{2}$$

Thus we form an CF decision $\hat{Y}^{\text{cf}}$ by drawing from a mixture: (a) sample a gender from the population distribution $s \sim p(s)$; (b) sample a test score from its counterfactual distribution $a(s) \sim p(a(s) \mid s_{\text{new}}, a_{\text{new}})$; (c) sample the ECO-fair decision for that counterfactual test score $\hat{y}^{\text{cf}} \sim f_{\text{eco}}(a(s))$. We repeat this sampling over different values of s and $a(s)$ to form the CF decision $\hat{Y}^{\text{cf}}$.

One subtlety of $\hat{Y}^{\text{cf}}$ is step (b). It draws the counterfactual test score under an intervened gender, but conditions on the observed gender and test score. This requires *abduction* (Pearl, 2009), where unobserved noise variables are inferred conditioned on observed variables. For example, consider a model where females have fewer opportunities for test preparation. If we observe a female applicant with a high test score, the unobserved noise might capture that this applicant is particularly gifted. under this value of noise, her counterfactual test score will be even higher.

Put differently, the CF-fair decision-maker uses the ECO-fair decision-maker $f_{\text{eco}}(a)$, which only depends on the test score (and averages over the gender). However, the CF-fair decision-maker also corrects for the affect on the test score due to the gender of the applicant. It replaces the current test score a_{new} with the adjusted score $a(s) \mid \{s_{\text{new}}, a_{\text{new}}\}$ under $s \sim p(s)$. With this corrected score, it produces an CF-fair decision.

For the simulated data in Fig. 1, we calculate $f_{\text{cf}}(s, a)$ from $f_{\text{eco}}(s, a)$. This decision-maker requires abduction, calculating the test score each applicant would have achieved had their gender been different. Consider applicant C. Her ECO-fair probability of acceptance is 69%, but when exercising CF it increases to 70%. This adjustment corrects for the (simulated) systemic difficulty of females to receive test preparation and, consequently, higher scores.

The CF-fair decision-maker $f_{\text{cf}}(s, a)$ accounts for historical disadvantage because it adjusts the applicant's test score (and their resulting admissions decision) to their counterfactual test scores under intervention on gender. Further, each element is computable from the dataset. (We discuss how to calculate these decisions below.) Note that CF-fair decisions can be formed from any ECO-fair decision-maker.

The CF-fair decision-maker depends on the protected attribute, and thus the CF-fair decision $\hat{Y}^{\text{cf}}$ is not ECO-fair. But among all CF-fair decisions, it is closest to the ECO-fair decision; we prove this fact in Thm. 2.

Theorem 2 (CF-fairness and optimality of the CF decisions). *The decision $\hat{Y}^{\text{cf}} \sim f_{\text{cf}}(s, a)$ is CF-fair. Among all CF decisions, the CF decision-maker minimally modifies the marginal distribution of Y^{eco},*

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

where $\mathcal{Y}^{\text{cf}}$ are all CF decisions and the expectation is over $P(S)P(A)$. It also preserves the marginal distribution of the ECO decision-maker, $P(\hat{Y}^{\text{cf}}) = P(\hat{Y}^{\text{eco}})$. (Proof in App. B.)

Thm. 2 also says that CF-fair decisions preserve the marginal distribution of the ECO-fair decisions. If the ECO decision-maker admits 20% of the applicants then the CF decision-maker will also admit 20%. (The CF decision-maker will likely admit a different set of applicants from the ECO decision-maker.) That it preserves the marginal distribution makes the CF-fair decision-maker applicable as a decision policy, such as when there a fixed budget for admissions.

Finally, CF-fair decisions also satisfy demographic parity, a group-level statistical criteria that has been used as a measure of CF (Dwork et al., 2012). Demographic parity requires equal decision distributions for the advantaged and disadvantaged groups of the protected attribute. (It does not involve a causal model.) Kusner et al. (2017) shows that, assuming the causal model in Fig. 2, decisions that satisfy counterfactual CF also satisfy demographic parity.

Why are the eco and cf decisions optimal? Thms. 1 and 2 establish the optimality of the the ECO and CF decisions; they rely on the following observation. The potential for unfairness arises when individuals have different values in their protected attributes. In the admissions example, applicants have different genders (protected attribute), which may causally affect their admission decisions (outcome) and lead to unfairness. If all applicants had the same gender then any predictor would be "fair" in that there is no possibility for discrimination.

This observation leads to the theorems. The ECO and CF decision-makers predict in a fictitious world where all applicants had the same gender S'. They take an existing predictor and ask: what would its counterfactual prediction be if each applicant had the gender S'? The ECO decision (Eq. 1) averages over a distribution $P(S')$. Thm. 1 says that the ECO decisions are closest to the ML decisions when S' is distributed as the gender distribution in the data. Thm. 2 says that the CF decision-maker is closest to (i.e. "minimally modifies") the ECO decision-maker.

The decisions are probabilistic, so closeness in the two theorems is measured by the average KL distance. The average is over a target population where the protected S and unprotected A are independent; this mimics an ideal setting that fair learning algorithms target, i.e. where no applicants are disadvantaged.

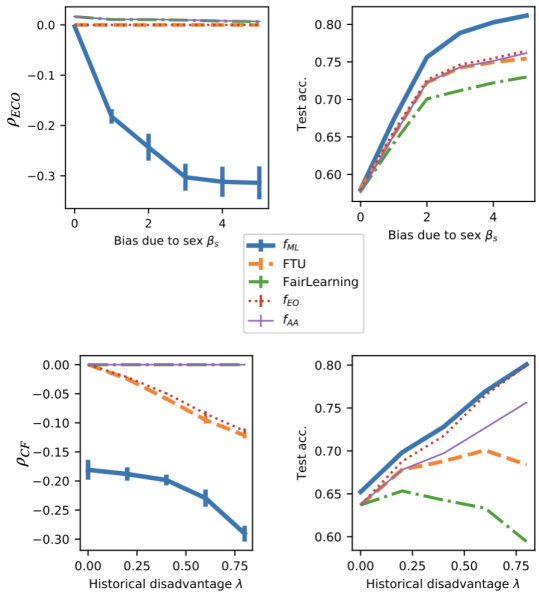


Figure 3: Measures of decision-quality in the simulated admissions data (error bars indicate ± 1 sd). The ECO decision-maker f_{ECO} is the most accurate one that achieves ECO-fairness; the CF decision-maker f_{cf} is the most accurate one that satisfies CF-fairness. The classical ML decision-maker f_{ml} is most accurate overall but violates both fairness measures. We vary the historical disadvantage β_s from 0.0 to +5.0 and vary direct gender discrimination λ from 0.0 to +0.8.

Calculating eco and cf decision-makers. The ECO and CF decision-makers can be calculated from data. The ECO decision-maker uses the fitted ML decision-maker f_{ml}, but marginalizes out the protected attribute. The CF decision-maker uses the ECO decision-maker f_{eco}, including an abduction step (Pearl, 2009) to calculate the counterfactual value of the attribute $a(s')$.

We focus on settings where the counterfactual decisions are identifiable from observational data. We assume: (1) The protected attributes S and unprotected attributes A follow the causal graph in Fig. 2. That is, there is no unobserved confounding between S and A, between A and Y, and between S and Y. (2) The structural equation model for S, A, Y needs to be correct so that structural equation can be identified from observational data. These two conditions enable the identification of counterfactual decisions.

Given the necessary assumptions for identifiability, the counterfactual calculation requires modeling the effect of the protected attribute S (gender) on the other attribute A (test score). With linear models, this calculation involves combining the residual on the observed data with a prediction about the counterfactual setting. In the data of Fig. 1, the abduction infers what applicant C's test would have been had she been male, given her test score (as a female) was 65. The algorithm can also be applied to nonlinear models; see App. C for details.

As a concrete example, Alg. 1 provides the algorithm for calculating ECO and CF decision-makers (Eqs. 1 and 2) from a fitted ML decision-maker in the case of additive-error models, e.g. $a = g(s) + \epsilon$ for some function g and some random variable ϵ. (We discuss the identification of the ECO and CF decision-makers and prove the correctness of Alg. 1 in App. C.)

The FairLearning (Kusner et al., 2017) algorithm, which also satisfies CF-fairness performs the same abduction step as f_{cf} to compute residuals $\varepsilon_i = a_i - \mathbb{E}[A \mid S = s_i]$ but then fits a decision-maker using only the residuals and non-descendants of the protected attributes. For example, in admissions, after calculating the residual it does not include the test score in its decision-maker. In contrast, Alg. 1 uses all available attributes and still is CF-fair. The ECO decision-maker in Alg. 1 also makes use of all attributes; it does not omit protected attributes as done in FTU (Kusner et al., 2017). Empirically, Alg. 1 provides more accurate decisions than these existing ECO and CF algorithms.

Multiple protected and unprotected attributes. Alg. 1 can be generalized to settings with multiple protected and unprotected attributes. Note that both S and A can be a vector, so we can collect all protected attributes into the S variable and all unprotected attributes into the A variable. As long as the resulting attributes still follow the causal graph in Fig. 2, Alg. 1 is still correct. That said, Alg. 1 may not apply in other cases, e.g. when the following three conditions simultaneously hold: (1) the multiple unprotected attributes are chained; (2) we consider interventions on only a subset of the unprotected attributes; (3) we focus on a subpopulation by conditioning on a descendant of the intervened unprotected attributes. In this case, the corresponding counterfactual is not identifiable.

Algorithm 1 The ECO and CF decision makers (for additive-error models).

Input: Data $\mathcal{D} = \{(s_i, a_i, y_i)\}_{i=1}^n$, where s_i is protected, a_i is not, and y_i is the decision.

Output: Decision makers $\{f_{\text{ml}}(s,a), f_{\text{eco}}(a), f_{\text{cf}}(s,a)\}$

From the data $\mathcal{D}$, fit $f_{\text{ml}}(s,a)$, $p(s)$, and $g(s) = \mathbb{E}[A \mid S = s]$ (e.g., with regression).

1. The ML decision maker $f_{\text{ml}}(s,a)$ draws from

$$\hat{y}^{\text{ml}} \sim f_{\text{ml}}(s,a).$$

2. The ECO decision maker $f_{\text{eco}}(a)$ draws from Eq. 1,

$$s' \sim p(s); \qquad \hat{y}^{\text{eco}} \sim f_{\text{ml}}(s', a).$$

3. The CF decision maker $f_{\text{cf}}(s,a)$ draws from Eq. 2,

$$s' \sim p(s); \quad a' = g(s') + (a - g(s)); \quad \hat{y}^{\text{cf}} \sim f_{\text{eco}}(s', a').$$

4 Empirical studies

We study algorithmic decisions on both simulated and real datasets, examining the tradeoff between fairness and prediction quality. We compare the ECO and CF decision-makers to existing algorithmic decision-makers—fairness through unawareness (FTU) and FairLearning (Kusner et al., 2017)—that target the same ECO and CF fairness criteria. (The supplement provides software that reproduces the studies.)

We find the following. (1) As expected, the classical ML decision $\hat{Y}^{\text{ml}}$ is accurate but unfair. (2) The ECO decision $\hat{Y}^{\text{eco}}$ is less accurate than $\hat{Y}^{\text{ml}}$, but is ECO-fair; it is more accurate than FTU. (3) The CF decision $\hat{Y}^{\text{cf}}$ is less accurate than the ECO-fair decision, but is CF-fair and achieves demographic parity; it is more accurate than FairLearning.

Evaluating algorithmic decision-makers. For each decision-maker, we measure the ECO and CF fairness of its decisions, and evaluate its fidelity relative to the historical data.

One way to measure fairness is to compare distributions of counterfactual decisions. Suppose a protected attribute S with values s (for the advantaged group) and s' (for the disadvantaged group). As one metric, we consider the average violations of ECO and CF over the population as metrics: $\rho_{\text{ECO}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left[\hat{Y}(s,a_i)\,|\,a_i,s_i\right] - \mathbb{E}\left[\hat{Y}(s',a_i)\,|\,a_i,s_i\right]\right], \rho_{\text{CF}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left[\hat{Y}(s,a(s))\,|\,a_i,s_i\right] - \mathbb{E}\left[\hat{Y}(s',a(s'))\,|\,a_i,s_i\right]\right]$. The sum is over n samples in a held-out test set; expectations are with respect to the decision-maker.

The ECO metric contrasts the decision probabilities when changing the protected attribute but holding other attributes fixed. The CF metric contrasts the decision probabilities when changing the protected attribute S and allowed other attributes to counterfactually vary. When the ECO metric is 0.0, the decision-maker achieves ECO-fairness; when it is greater than 0.0, there is bias towards the disadvantaged group; when it is less than 0.0, the decision-maker favors the disadvantaged group. The same interpretation applies to CF.

Another measure of CF fairness is demographic parity Dwork et al. (2012). We measure demographic parity with the symmetric Kullback-Leibler (KL) divergence between prediction distributions for the decision $P(\hat{Y}\,|\,S=s)$ and $P(\hat{Y}\,|\,S=s')$; it is zero for a decision-maker that achieves demographic parity. (We evaluate the symmetric KL by binning the values of predictions.)

Finally, we measure the fidelity of the decision-maker to the historical data on which it was fit. The metric is the prediction score: for binary outcomes, it is the mean accuracy of out-of-sample predictions; for real-valued outcomes, it is the coefficient of determination R^2 of the predictions.

Simulated admissions. We first study simulated datasets about an unfair admissions committee, which is from the following structural model, $s_i \sim \text{Bernoulli}(0.5); a_i\,|\,s_i = \max(0, \min(\lambda \cdot s_i + 100 \cdot \varepsilon, 100)); \varepsilon \sim \text{Uniform}[0,1]; y_i\,|\,s_i, a_i \sim \text{Bernoulli}(\sigma(-1.0 + \beta_a \cdot a + \beta_s \cdot s))$.

(We threshold the test score to mimic real-world test scores that are usually bounded.) We fix the effect of test score on admissions to $\beta_a = 2.0$. We generate multiple datasets by varying the gender bias β_s and the historical disadvantage on test score λ.

Fig. 3 show how ECO-fairness and prediction quality trade off as the bias β_s increases. Only the ECO decision-makers f_{ECO} and FTU achieve ECO fairness. Although both decision-makers are less accurate than classical ML, the ECO decision-maker is more accurate than FTU. Fig. 3 shows how CF-fairness and prediction quality trade off as the historical disadvantage λ increases. Only the f_{cf} and FairLearning decision-makers achieve CF-fairness. Among these CF-fair decision-makers, the CF decision-maker of this paper is more accurate.

	Metrics ($\times 10^2$) on Adult			
	ρ_{ECO}	ρ_{CF}	**KL**	**Prediction**
f_{ml}	2.5	16.2	15.0	78.6
FTU	**0**	14.8	12.2	77.3
f_{eco}	**0**	14.1	12.6	**77.4**
FL	-14.8	**0**	5.3	75.1
f_{cf}	-9.1	**0**	**1.5**	**77.1**

Figure 4: In the Adult dataset, algorithmic decision makers decide which individuals are loanworthy based on their income and we examine their fairness relative to gender. The ECO and CF decision makers are fair towards females while remaining the most accurate. Both the ECO decision maker f_{eco} and FTU achieve ECO-fairness as measured by the ECO-metric. The CF decision maker f_{cf} and FairLearning (FL) (Kusner et al., 2017) are CF-fair and achieve demographic parity (close-to-zero KL). The ML decision maker f_{ml} is the most accurate overall but the ECO decision maker is the most accurate among the ECO-fair decision makers. We report mean values across individuals. ECO and CF metric standard deviations are ≤ 0.1 and ≤ 0.11, respectively.

Case studies. We study ECO and CF decision-makers on a simulated admissions dataset and three real datasets of sensitive decisions about people. The adult income data (Dua & Graff, 2017a) and the German credit data (Dua & Graff, 2017b) contain data about people and decisions about which are loan worthy.[1] ProPublica's COMPAS data contains information about criminal defendants and decisions about their

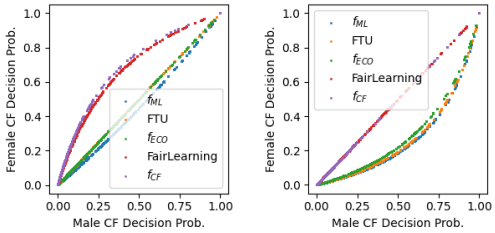
[1]Readers who are familiar with these datasets know that they contain scores, as opposed to decisions. We use these scores to create "decisions" in order to evaluate the methods of this paper on binary decisions.

recidivism score. Each dataset contains protected and unprotected attributes, and in the studies we adjust classical ML decision-makers to achieve fairness relative to the protected attributes. We focus on the adult income data in this section and defer the results of the German credit data and the COMPAS data to App. F.

Fig. 4 summarize the fairness and prediction quality from the adult income data. In the adult income data, race and gender are the protected attributes. We discuss the fairness metrics ρ_{ECO} and ρ_{CF} relative to gender and report the mean (and s.d.) of all the metrics. The findings are consistent with the simulation studies and generalize across the studies.

Although the classical ML decision-maker is the most accurate, its decisions are biased against the disadvantaged group. Positive values in the EO-metric ρ_{ECO} reveal direct discrimination in classical ML decisions; males and white individuals receive a higher probability of being decided as loan-worthy or lower recidivism score than their equivalent female or non-white counterparts. Positive values in the AA-metric ρ_{CF} indicate that the classical ML decision-maker does not exercise CF. It provides lower decision probabilities to females or non-white individuals than their male or white counterparts whose attribute values also reflect the historical advantage.

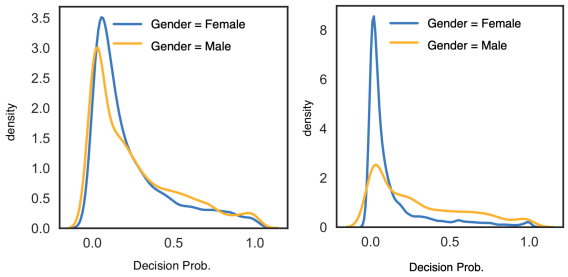
The ECO and FTU decisions are ECO-fair (ρ_{ECO} is zero) and the ECO decisions are more accurate. The ECO-fair decision-makers provide equal decision probabilities to individuals that are equal in all other attributes, regardless of gender or race. Neither counterfactual CF nor FairLearning achieve ECO-fairness, but this is by design. These decisions produce negative values for the ECO metric, showing an advantage for the disadvantaged group. The CF-fair decision-makers correct for historical biases, which leads to different decisions for the disadvantaged and advantaged groups.



(a) Assessing ECO **(b)** Assessing CF

Figure 5: Comparing individual decision probabilities under (hypothetical) "intervention" on gender (in the adult income data). (a) To study ECO, the other attributes remain at their original values and we plot $p(\hat{y}(s,a) \mid a,s)$ against $p(\hat{y}(s',a) \mid a,s)$. The f_{ECO} decision maker provides ECO to females; it produces decision probabilities that align along the diagonal. (b) To study CF, the remaining attributes vary with the intervention on gender and we plot $p(\hat{y}(s,a) \mid a,s)$ against $p(\hat{y}(s',a(s')) \mid a,s)$. The decision maker f_{cf} exercises CF; it provides equal decision probabilities to females as their male counterparts after correcting for the historical disadvantage they face. In contrast, the classical ML decision maker f_{ml} demonstrates both direct and indirect discrimination against females; it provides lower decision probabilities to females compared to their male counterparts in both comparisons.

The CF and FairLearning decisions both achieve CF-fairness (ρ_{CF} is zero) and demographic parity (KL is zero). They succeed in providing equal decision probabilities to females or non-white individuals after considering their counterfactual attribute values had they been male or white. Among these CF-fair decision-makers, f_{cf} is more accurate.



(a) CF decision maker **(b)** ML decision maker

Figure 6: The adult income data: distributions of decision probabilities. (a) The CF-fair decision maker f_{cf} produces equal decision probabilities for females and males, achieving demographic parity with respect to gender. (b) The classical ML decision maker f_{ml} does not satisfy demographic parity.

Measuring decisions in the adult income data. We study the algorithmic decisions in more depth. Using the adult income data, we examine the metrics of Fig. 4. (The results below exhibit similar patterns in all three datasets.)

We first discuss ECO-fairness for gender. Fig. 5a reveals the same patterns of ECO-fairness that Fig. 4 showed. Both ECO and FTU decision probabilities align with the diagonal; they are ECO-fair in providing equal decision probabilities to equally qualified individuals regardless of gender. None of the other algorithmic decision makers is ECO-fair. Fig. 5a shows that classical ML decisions are biased against females, producing lower decision probabilities for them than

their male counterparts. In contrast, the CF-fair decision makers f_{cf} and FairLearning are less ECO-fair than f_{ml} but provide an advantage to females after accounting for historical disadvantages. Further, CF and FairLearning are less ECO-fair than classical ML.

We next discuss CF-fairness for gender. Fig. 5b compares the decision probabilities received by individuals if they were females to those they receive if they were males, with the resulting changes to their other attributes. An algorithmic decision maker is CF-fair if the decision probabilities align with the diagonal. The decision makers f_{cf} and FairLearning produce decision probabilities that align with the diagonal and are thus CF fair. The remaining decision makers do not exercise CF; ECO decisions provide a greater degree of CF-fairness than FTU decisions while classical ML decisions are the most unfair in exercising CF.

We next inspect demographic parity. Fig. 6 illustrates demographic parity by plotting the decision probabilities for both male and female individuals. The left panel comes from a classical ML decision maker; the right panel comes from f_{cf}. While the classical ML decision maker produces noticeably different decision probabilities for the two groups, the decision probabilities produced by the CF decision maker f_{cf} are nearly identical.

We now turn to the predictive performance. Fig. 4 reports the predictive performance of all methods. Though it is not fair by either criterion, classical ML produces the "best" predictions, closest to the historical dataset. Among the ECO-fair methods, ECO produces better predictions than FTU. This corroborates Thm. 1. Moreover, ECO's prediction scores are close to classical ML, so its fairness comes with little cost. Among the CF-fair methods, CF predicts better than FairLearning; this corroborates Thm. 2. But since it alters the decision probability of individuals from disadvantaged groups, it predicts less well than the ML methods.

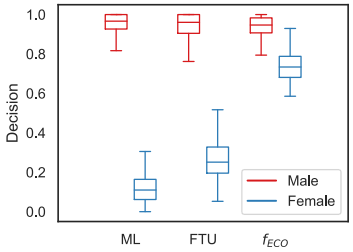


Figure 7: FTU vs. f_{ECO} ; The ECO decision avoids inheriting discrimination patterns when unprotected attributes are highly correlated with the protected attribute. While both ECO fair, FTU decisions remain highly correlated with gender while the ECO decision has a substantially lower correlation.

Finally, we contrast FTU and the ECO decision-maker. While both decision-makers are ECO-fair, they can have different behaviors when some unprotected attribute is highly correlated with the protected ones. As an example, we simulate an additional unprotected attribute that has a 0.9 correlation with gender. We further subsample the adult dataset such that the correlation between the protected attribute and the decision is 0.7. We then compare the FTU decision and the ECO. While both ECO-fair, Fig. 7 shows that the FTU decisions remain highly correlated with gender while the ECO decision has a substantially lower correlation. The ECO decision avoids inheriting discrimination patterns when unprotected attributes are highly correlated with the protected attribute. Such a high correlation between protected and unprotected attributes do not appear in the adult dataset; the prediction accuracy of FTU and f_{ECO} are similar. That said, Fig. 7 delineates a setting where the ECO decision-maker may be desired.

5 Discussion

We develop fair ML algorithms that modify fitted ML predictors to make them fair. We prove that the resulting predictors are ECO-fair or CF-fair, and they otherwise maximally recover the fitted ML predictor.

There are a few limitations of this work. One limitation is in the scope of fairness notions being studied. We focus on two counterfactual notions of fairness in this work; the proposed algorithms may only be fair with respect to these notions; they may not improve fairness with respect to other fairness notions. Develop post-hoc algorithms that simultaneously enforce multiple fairness notions is an interesting direction.

Another limitation lies in the assumptions required by counterfactual approaches to fairness. The methods in this paper rely on a correct causal model, the ability to estimate its parameters, and the ability to estimate the necessary counterfactuals. We recognize that these are strong requirements—positing and fitting causal models requires both domain expertise and statistical care. Moreover, even with the correct model in hand, issues of unobserved confounding can affect the ability to properly estimate its parameters or the counterfactuals. Developing fairness algorithms robust to violations of these assumptions is another worthwhile avenue of future work.

References

Artelt, A., Vaquet, V., et al. (2021). Evaluating robustness of counterfactual explanations. *arXiv preprint arXiv:2103.02354*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3–44.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21–40.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: assessing disparity when protected class is unobserved. *Fairness, Accountability and Transparency*.

Chiappa, S. (2019). Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 7801–7808.

Chouldechova, A. & Roth, A. (2019). The frontiers of fairness in machine learning. *Communications of the ACM*.

Corbett-Davies, S. & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

Coston, A., Mishler, A., Kennedy, E., & Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. *FAT**.

Dai, J., Upadhyay, S., Bach, S. H., & Lakkaraju, H. (2021). What will it take to generate fairness-preserving explanations? *arXiv preprint arXiv:2106.13346*.

De Lara, L., González-Sanz, A., Asher, N., & Loubes, J.-M. (2021). Transport-based counterfactual models. *arXiv preprint arXiv:2108.13025*.

Dua, D. & Graff, C. (2017a). UCI machine learning repository: Adult data set.

Dua, D. & Graff, C. (2017b). UCI machine learning repository: Statlog (german credit data) data set.

Dutta, S., Venkatesh, P., Mardziel, P., Datta, A., & Grover, P. (2021). Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10), 6675–6710.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Innovations in Theoretical Computer Science Conference*.

Foster, A., Vezér, Á., et al. (2021). Contrastive mixture of posteriors for counterfactual inference, data integration and fairness. *arXiv preprint arXiv:2106.08161*.

Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 577–590).

Greiner, D. J. & Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3), 775–785.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. *Neural Information Processing Ssystems*.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 353–362).

Kasirzadeh, A. & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 228–236).

Kilbertus, N., Ball, P., Kusner, M., Weller, A., & Silva, R. (2019). The sensitivity of counterfactual fairness to unmeasured confounding. *Uncertainty in Artificial Intelligence*.

Kilbertus, N., Carulla, M. R., et al. (2017). Avoiding discrimination through causal reasoning. *Neural Information Processing Systems*.

Kilbertus, N., Gascon, A., et al. (2018). Blind justice: Fairness with encrypted sensitive attributes. *International Conference on Machine Learning*.

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113, 1163.

Kusner, M., Russell, C., Loftus, J., & Silva, R. (2019). Making decisions that reduce discriminatory impact. *International Conference on Machine Learning*.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Neural Information Processing Systems*.

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

Lin, Z. "., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, 6(7).

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 349–358).

Mhasawade, V. & Chunara, R. (2021). Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 784–794).

Mishler, A. & Kennedy, E. (2021). Fade: Fair double ensemble learning for observable and counterfactual outcomes. *arXiv preprint arXiv:2109.00173*.

Mishler, A., Kennedy, E. H., & Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 386–400).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–163.

Nabi, R. & Shpitser, I. (2018). Fair inference on outcomes. *AAAI Conference on Artificial Intelligence*.

Parafita, A. & Vitria, J. (2021). Deep causal graphs for causal inference, black-box explainability and fairness. In *Artificial Intelligence Research and Development: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, volume 339 (pp. 415).: IOS Press.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pessach, D. & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3), 1–44.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Russell, C., Kusner, M., Loftus, J., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. *Neural Information Processing Systems*.

Speicher, T., Heidari, H., et al. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *Knowledge Discovery and Data Mining*.

Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021). Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34.

Waters, A. & Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *AI Magazine*, (pp.6̃4).

Wu, Y., Zhang, L., & Wu, X. (2019a). Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Wu, Y., Zhang, L., Wu, X., & Tong, H. (2019b). Pc-fairness: A unified framework for measuring causality-based fairness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 32: Curran Associates, Inc.

Zhang, J. & Bareinboim, E. (2018a). Equality of opportunity in classification: A causal approach. *Neural Information Processing Systems*.

Zhang, J. & Bareinboim, E. (2018b). Fairness in decision-making–the causal explanation formula. *AAAI Conference on Artificial Intelligence*.

Zhang, L., Wu, Y., & Wu, X. (2017). A causal framework for discovering and removing direct and indirect discrimination. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, (pp. 3929–3935).