Equal Opportunity and Affirmative Action via Counterfactual Predictions

Yixin Wang
Columbia University

Dhanya Sridhar
Columbia University

David M. Blei
Columbia University

Abstract

Machine learning (ML) can automate decision-making by learning to predict decisions from historical data. However, these predictors may inherit discriminatory policies from past decisions and reproduce unfair decisions. In this paper, we propose two algorithms that adjust fitted ML predictors to make them fair. We focus on two legal notions of fairness: (a) providing equal opportunities (EO) to individuals regardless of sensitive attributes and (b) repairing historical disadvantages through affirmative action (AA). More technically, we produce fair EO and AA predictors by positing a causal model and considering counterfactual decisions. We prove that the resulting predictors are theoretically optimal in predictive performance while satisfying fairness. We evaluate the algorithms, and the trade-offs between accuracy and fairness, on datasets about admissions, income, credit, and recidivism.

1 Introduction

Machine learning (ML) methods can automate costly decisions by learning from historical data. For example, ML algorithms can assess the risk of recidivism to decide bail releases or predict admissions to determine which college applicants should be further reviewed (Waters & Miikkulainen, 2014).

Consider an admissions committee that decides which applicants to accept to a university. Given historical data about the applicants and the admission decisions, an ML algorithm can learn to predict who will be admitted and who will not. The ML predictor might accurately simulate the committee's decisions, but it might also inherit some of the undesirable properties of the committee. If the committee was systematically and unfairly biased then its ML replacement will be biased as well.

In this paper, we develop algorithms for learning ML predictors that are both accurate and fair. Our methods maintain as much fidelity as possible to the decisions of the committee, but they adjust the predictor to ensure that the algorithmic decisions are fair.

We focus on two legal notions of fairness: equal opportunities (EO) and affirmative action (AA). An EO decision provides the same opportunities to similar candidates with different backgrounds (Barocas & Selbst, 2016). In admissions, for example, an EO decision admits applicants based only on their qualifications; applicants with the same merit, such as a test score, should have the same chance of admission regardless of their race or sex.

Some applicants might have lower test scores only because they belong to a historically disadvantaged demographic group, one that has less access to opportunities. The legal notion of affirmative action (AA) acknowledges these historical disadvantages (Anderson, 2002). For example, AA will correct for situations where male applicants have easier access to extra test preparation. (Notice that, by definition, an AA decision is not an EO decision because AA decisions provide advantages based on group membership.)

Using these two notions of fairness, we develop algorithms that adjust ML predictors to be fair predictors. The first produces algorithmic decisions that ensure equal opportunities for disadvantaged groups; the second produces decisions that exercise affirmative action to repair existing disparities. We prove the algorithms are theoretically optimal—they provide as good predictions as possible while still being fair.

In detail, we use ideas from causal machine learning (Pearl, 2009; Peters et al., 2017). First we posit a causal model of the historical decision-making process. In that model, the decision can be affected by all the attributes, both sensitive and non-sensitive, and the sensitive attributes can also affect the non-sensitive ones (e.g., sex can affect test scores). Then we consider algorithmic decisions as causal variables, ones whose values are functions of the attributes. Once framed this way, we can ask counterfactual questions about an algorithm, e.g., "what would the algorithm decide for this applicant if she had achieved the same test score but was a male?" We operationalize EO and AA fairness as probabilistic properties of such counterfactual decisions. (AA-fairness is the same as counterfactual fairness, proposed in Kusner et al. (2017).) Finally, we derive a method that adjusts a classical ML decision-maker to be EO-fair, and that adjusts an EO decision-maker to exercise AA.

We study our algorithms on simulated admissions data and on three public datasets, about income, credit, and recidivism. We examine the gap in accuracy between classical prediction and EO/AA-fair prediction, and we investigate the degree of unfairness that classical prediction would exhibit. Compared to other approaches to EO/AA-fairness, ours provide higher accuracy while remaining fair.

Related Work. Several threads of related work draw on causal models to formalize fairness. Kilbertus et al. (2017) relate paths and variables in causal models to violations of EO and AA. Zhang & Bareinboim (2018) decompose existing fairness measures into discrimination along different paths in the causal graph, and propose fair decision algorithms that satisfy the path-specific criteria. Kusner et al. (2017) introduce counterfactual fairness and an algorithm to satisfy it. Our work also uses causal graphs to formalize the EO and AA criteria, but our fair algorithms are designed to minimally correct fitted ML predictors. Like FairLearning Kusner et al. (2017), our AA-fair algorithm also satisfies counterfactual fairness, but our algorithm has higher fidelity to the historical decisions.

Other research has proposed a variety of statistical definitions of fairness for supervised learning (Hashimoto et al., 2018; Chen et al., 2019; Kleinberg et al., 2016; Chouldechova, 2017; Zafar et al., 2017). Hardt et al. (2016) and Kallus & Zhou (2018) show, however, that these approaches cannot distinguish between different mechanisms of discrimination and they are not robust to changes in the data distribution. By building on causal machine learning, our work separates direct discrimination from indirect bias, and it is robust to changes in the data distribution.

Another line of work focuses on data preprocessing. Zemel et al. (2013) learn intermediate representations that remove information about the sensitive attributes; Calders & Verwer (2010) and Kamiran & Calders (2012) reweight training data to control false positive/negative rates. In contrast, our EO-fair and AA-fair algorithms adjust ML predictions without changing the training data.

Wang et al. (2019) modify black-box ML predictors to correct for historical disadvantages, as in AA. They define adjusted distributions over attributes for the disadvantaged group and solve optimal transport problems to repair predictors. Our predictors also minimally adjust ML predictors but we derive adjustments using causality.

Finally, Dwork et al. (2012) define individual fairness, which formalizes EO without causality. Given a distance metric between individuals, it requires that similar individuals receive similar decisions. Our EO-fair algorithm recovers this requirement without relying on an explicit distance metric, which can be difficult to construct. Dwork et al. (2012) also specify group-level AA fairness as independence between the sensitive attribute and decision (demographic parity). We exercise AA for each individual while also satisfying demographic parity.

Contributions. We develop optimal EO and AA algorithms. These algorithms modify existing ML predictors to produce EO-fair and AA-fair decisions, and we prove these decisions maximally recover the ML predictors under the fairness constraints. While existing approaches must omit descendants of the sensitive attributes to construct AA-fair decisions (Kusner et al., 2017; Kilbertus et al., 2017), our AA algorithm uses all available attributes and is still AA-fair. Empirically, we show that our EO and AA algorithms produce better decisions than existing approaches that satisfy the same fairness criteria.

ID	Sex	Test	Admit	$\hat{y}^{\text{ml}}$	$\hat{y}^{\text{eo}}$	$\hat{y}^{\text{aa}}$
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: We simulate an unfair admissions process that violates both EO and AA. This table shows the training data (1-5000) and three new applicants (A, B, C).

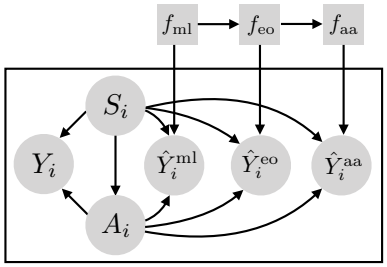


Figure 2: Classical ML, EO and AA decisions are causal variables. EO predictors upgrade classical ML ones, and AA adjust EO ones.

2 Assessing fairness with counterfactuals

As a running example, consider automating the admissions process at a university. Using a dataset of past admissions, the goal is to algorithmically compute the admissions decision for new applicants. The dataset contains n applicants and the committee's decision about whether to admit them. Each applicant has demographic attributes, such as sex, religion, or ethnicity, that the university has deemed sensitive. Each applicant also has other attributes, such as a standardized test score or grade point average, that are not sensitive.

For simplicity, suppose each applicant has two attributes, their sex and their score on a standardized test. The sex of the applicant is a sensitive attribute; the test score is not. (Our algorithms extend to multiple attributes but this simple case illustrates the concepts.) Fig. 1 shows an example of such data.

We will use a causal model of this decision-making process to define two notions of fairness.

A causal model of the decision. First posit a *causal* model of the admissions process, such as the causal graph in Fig. 2. It assumes that an applicant's sex S and test score A affect the admissions decision Y. For example, the admissions committee might unfairly prefer to admit males. The model also assumes the applicant's sex can affect the test score. For example, female applicants might receive less exposure to test preparation, a disparity that results in lower scores.[1] (This simple model illustrates the core ideas; § 3 discusses extensions to more complicated causal graphs.)

The data in Fig. 1 matches the assumptions. It comes from the following structural model,

$$\begin{aligned} s_i &\sim \text{Bernoulli}(0.5), \\ a_i \,|\, s_i &= \max(0, \min(\lambda \cdot s_i + 100 \cdot \varepsilon, 100)); \quad \varepsilon \sim \text{Uniform}[0,1] \\ y_i \,|\, s_i, a_i &\sim \text{Bernoulli}(\sigma(-1.0 + \beta_a \cdot a + \beta_s \cdot s)). \end{aligned} \tag{1}$$

(We threshold the test score to mimic real-world test scores that are usually bounded.) We set $\beta_a = 2.0$ and $\beta_s = 1.0$ so that for the same test score the committee is more likely to admit males than females. Moreover, we set $\lambda = 2.0$ to model that females perform less well on the test.

With data like Fig. 1 and a causal model, we can estimate the parameters of the functions that govern each variable in Fig. 2. (This estimation makes the usual assumptions about causal estimation, i.e., no open backdoor paths (Pearl, 2009).)

The causal model implies counterfactuals, how the distribution of a variable changes when we intervene on others. For example, we can consider the counterfactual decision of a female applicant had she been a male. [2] 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 sex to be $S_i = s$.

[1] We acknowledge these assumptions might be debatable; it is up to the user of our method to posit a sensible causal model of the historical decision process.

[2] Even when sensitive attributes are immutable, we can define hypothetical interventions to articulate the counterfactuals (Greiner & Rubin, 2011). In this example, "intervening on sex" means to make a person born of a different sex.

Marginal distributions of counterfactuals provide an alternative to the *do* notation, $P(Y_i(s,a)) = P(Y_i\,;\,\text{do}(S_i = s, A_i = a))$. But counterfactual distributions can also be more complex: for example, $P(Y_i(\text{female}) \,|\, Y_i = \text{Yes}, S_i = \text{male})$ is the probability of applicant i's counterfactual admission if he was female, given that he is (factually) male and was admitted. See Ch. 7 of Pearl (2009) for a treatment of how causal models imply distributions of counterfactuals.

We use counterfactual distributions to assess the fairness of decisions, either those made by the committee or those produced by an algorithm. We next describe the counterfactual criteria of equal opportunities and affirmative action.

Equal opportunities and its counterfactual criterion. A decision satisfies EO if changing the sensitive attribute S does not change the distribution of the decision Y. Consider an applicant with sex s and test score a. An EO decision gives the same probability of admission even when changing the sex to s' (but keeping the test score at a). In Fig. 1, candidates A and B have the same test score but different sex; they should receive the same probability of admission.

Definition 1 (The EO criterion)**.** *A decision Y satisfies equal opportunities (*EO*) in the sensitive attribute S if*

$$Y(s,a) \,|\, \{S = s, A = a\} \overset{d}{=} Y(s',a) \,|\, \{S = s, A = a\} \tag{2}$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, where $\mathcal{S}$ and $\mathcal{A}$ are the domains of S and A.[3]

The EO criterion differs from counterfactual fairness (Kusner et al., 2017). EO focuses on the counterfactual prediction had the applicant had the same non-sensitive attribute A but a different sensitive attribute S; counterfactual fairness does not hold the attribute A fixed.

The EO criterion also differs from equalized odds (Hardt et al., 2016). Equalized odds requires the sensitive attribute S to depend on the decision Y only via its ability to predict desirable downstream outcomes (e.g., college success). Equalized odds enforces statistical properties while EO assesses counterfactuals. The EO criterion assumes that the sensitive attribute S does not causally affect the decision Y. (We prove this in Appendix A.)

Affirmative action and its counterfactual criterion. Affirmative action (AA) aims to correct for the historical disadvantages that may be caused by sensitive attributes. For example, female applicants may have had fewer opportunities for extracurricular test preparation, which leads 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.

The goal of AA is to ensure that, all else being equal (e.g., effort or aptitude), an applicant would receive the same decision had they not been in a disadvantaged group. This is different from EO because it accounts for how the sensitive attribute affects the other attributes, for example how an applicant's sex affects their test score.

The AA criterion requires that changing the sensitive attribute S, along with the resulting change in the attribute A, does not change the distribution of the decision Y. Consider an applicant with test score a and sex s. Had they belonged to a different group s', they may have achieved a different test score $A(s')$. The AA criterion requires the counterfactual decision with the alternate sex $Y(s', A(s'))$ to have the same distribution as the decision with the current sex $Y(s, A(s))$. It was first proposed by Kusner et al. (2017), who dubbed it *counterfactual fairness*.

Definition 2 (The AA criterion, counterfactual fairness (Kusner et al., 2017))**.** *A decision Y satisfies affirmative action (*AA*) if*

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

for all $s' \in \mathcal{S}$. The random variable $A(s')$ is the counterfactual of A under intervention of the sensitive attribute $S = s'$.

[3]Two random variables are equal in distribution $X_1 \overset{d}{=} X_2$ if they have the same distribution function: $P(X_1 \leq x) = P(X_2 \leq x)$ for all x.

In Eq. 3 the counterfactual attribute $A(s')$ depends on the attributes of the applicant, $\{S = s, A = a\}$. The variable $A(\text{male}) \,|\, \{S = \text{female}, A = \text{high score}\}$ captures, for example, that female applicants with high test scores will have even higher test scores had they been male. Assuming the causal model in Fig. 2, the AA criterion implies demographic parity (Dwork et al., 2012; Kusner et al., 2017), which requires the decision be independent of the sensitive attribute. We note that a decision cannot be both EO-fair and AA-fair.

Disparate treatment and impact. We have described two counterfactual criteria for fairness. One view on fairness is to seek to avoid "disparate treatment" or "disparate impact." Applicants face disparate treatment when they are rejected by the biased admissions committee because of their sex. Applicants face disparate impact when they are systematically disadvantaged because they do not have the same opportunities for test preparation. From this perspective on fairness, an EO-fair decision eliminates disparate treatment; an AA-fair decision addresses disparate impact.

3 Constructing fair algorithmic decisions

Definitions 1 and 2 define fairness in terms of distributions of counterfactual decisions. Using the causal model of the decision-making process, these definitions help evaluate the fairness of its outcomes. The criteria can assess any decisions, including those produced by a real-world process, such as an admissions committee, or those produced by an algorithm, such as a fitted ML model.

Consider a classical ML model that was fit to emulate historical admissions data; if the historical decisions are not EO or AA then neither will be the ML decisions. To this end, we develop fair ML predictors, those that are accurate with respect to historical data and produce EO-fair or AA-fair decisions. The key idea is to treat the algorithmic decision as another causal variable in the model and then to analyze its fairness properties in terms of the corresponding counterfactuals. We will show how to adjust decisions made by an ML model to make them EO-fair or AA-fair.

Formally, we use the language of *decisions* and *predictors*. Denote an algorithmic decision as $\hat{Y}$, a causal variable that depends on the attributes $\{s, a\}$. It comes from a fitted probabilistic predictor, $\hat{Y} \sim f(s, a)$, where $f(\cdot)$ is a probability density function. For example, an 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 in the causal model, we can infer algorithmic counterfactuals $\hat{Y}(s, a)$ and assess whether they satisfy the fairness criteria of § 2.

ML decisions. Consider a classical machine learning predictor, one that is fit to accurately predict the decision Y from S and A. In the admissions example, a binary decision, logistic regression is a common choice. The predictor $f_{\text{ml}}(s, a)$ draws the decision from a Bernoulli,

$$\hat{Y}^{\text{ml}} \,|\, \{S = s, A = a\} \sim \text{Bern}(\sigma(\beta_A \cdot a + \beta_S \cdot s + \beta_0)),$$

and the coefficients are fit to maximize the likelihood of the observed data. When incorporated into 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. However, it will also replicate harmful discriminatory practices. The decision $\hat{Y}^{\text{ml}}$ might violate EO and not give equal opportunities to applicants of different sex. Nor will $\hat{Y}^{\text{ml}}$ exercise AA. If policy mandates that admissions correct for the disparate impact of sex on the test score, then the ML decision cannot be used.

Return to the illustrative simulation in 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. To see this, 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 their identical test scores, the female applicant is 17% less likely to get in.

ML decisions that satisfy equal opportunities. We use the ML predictor f_{ml} to produce an EO predictor f_{eo}, one whose decisions satisfy the EO criterion. Consider an ML predictor $f_{\text{ml}}(s, a)$ and an applicant with attributes $\{s_{\text{new}}, a_{\text{new}}\}$. Their EO decision is $\hat{Y}^{\text{eo}}(s_{\text{new}}, a_{\text{new}}) \sim f_{\text{eo}}(a_{\text{new}})$, where

$$f_{\text{eo}}(a_{\text{new}}) = \int f_{\text{ml}}(s, a_{\text{new}}) p(s) \, \mathrm{d}s. \tag{4}$$

The distribution $p(s)$ is the population distribution of the sensitive attribute.

The EO decision probability holds the non-sensitive attribute a_{new} fixed, and takes a weighted average of the ML predictor for the sensitive attribute s. Note all of its ingredients are computable: the distribution $p(s)$ is estimated from the data and $f_{\text{ml}}(s, a)$ is the fitted ML predictor.

The EO decision $\hat{Y}^{\text{eo}} \sim f_{\text{eo}}(a_{\text{new}})$ satisfies Definition 1: applicants with the same test score will have the same chance of admissions regardless of their sex. Further, it is the most accurate among all EO-fair decisions. We prove that it is EO-fair and minimally corrects $\hat{Y}^{\text{ml}}$..

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

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

where $\mathcal{Y}^{EO}$ is the set of EO-fair decisions. (The proof is in Appendix A.)

The intuition is that the EO 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 sensitive attribute S. In the *do* notation, what this means is that $P(\hat{Y}^{\text{eo}}; \text{do}(s, a)) = P(\hat{Y}^{\text{ml}}; \text{do}(a))$ for all s. Eq. 4 is the adjustment formula in Pearl (2009), which calculates $P(\hat{Y}^{\text{ml}}; \text{do}(a))$.

Again return to Fig. 1. We use the fitted logistic regression $f_{\text{ml}}(s, a)$ to produce the EO predictor. This data has equal numbers of women and men, so the weighted average is

$$f_{\text{eo}}(s, a) = 0.5 f_{\text{ml}}(\text{male}, a) + 0.5 f_{\text{ml}}(\text{female}, a).$$

Using EO-fair decisions, applicants A and B both have a 77% probability of admission.

Why not use fairness through unawareness (FTU) (Kusner et al., 2017)? FTU satisfies EO by fitting an ML model from the non-sensitive attribute A to the decision Y. FTU decisions are also EO-fair; the sensitive attribute is never used. However, Theorem 1 states that the EO decisions of Eq. 4 are more accurate than FTU while still being EO-fair. We show this fact empirically in § 4.

ML decisions that exercise affirmative action. How can we construct an algorithmic decision that exercises affirmative action (AA)? We now show how to correct the EO-fair predictor to account for historical disadvantages that stem from the sensitive attribute S.

Consider the EO-fair predictor $f_{\text{eo}}(a)$ and an applicant with attributes $\{s_{\text{new}}, a_{\text{new}}\}$. Their AA-fair decision is $\hat{Y}^{\text{aa}} \sim f_{\text{aa}}(s_{\text{new}}, a_{\text{new}})$, where

$$f_{\text{aa}}(s_{\text{new}}, a_{\text{new}})) = \int \int f_{\text{eo}}(a(s)) \; p(a(s) \mid s_{\text{new}}, a_{\text{new}}) \; p(s) \, \mathrm{d}a(s) \, \mathrm{d}s. \tag{5}$$

The distribution $p(s)$ is the population distribution of the sensitive attribute; the distribution $p(a(s) \mid s_{\text{new}}, a_{\text{new}})$ is the counterfactual distribution of the non-sensitive attribute a, under intervention of s and conditional on the observed values a_{new} and s_{new}.

This predictor $f_{\text{aa}}(s, a)$ exercises affirmative action because it corrects the applicant's test score (and their resulting admissions decision) to their counterfactual test scores under intervention of the sex. Further, each element is computable from the dataset. (See below.) Note that AA-fair decisions can be formed from any EO-fair predictor.

One way to form an AA-fair decision is to draw from the mixture defined in Eq. 5. Consider the admissions example and an applicant $\{s_{\text{new}}, a_{\text{new}}\}$: (a) sample a sex 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 EO-fair decision for that counterfactual test score $\hat{y}^{\text{aa}} \sim f_{\text{eo}}(a(s))$.

One subtlety of $\hat{Y}^{\text{aa}}$ is in step (b): it draws the counterfactual test score a under an intervened sex s, but conditional on the observed sex and test score. This is an *abduction*—if a female applicant s_{new} has a high test score a_{new}, then her counterfactual test score $a(\text{male})$ will be even higher, and the likelihood of admission goes up beyond the EO-fair decision (Pearl, 2009).

Put differently, the AA-fair predictor uses the EO-fair predictor $f_{\text{eo}}(a)$, which only depends on the test score (and averages over the sex). However, the AA-fair predictor also corrects for the bias in the test score due to the sex 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 AA-fair decision.

The AA-fair predictor depends on the sensitive attribute, and thus the AA-fair decision $\hat{Y}^{\text{aa}}$ is not EO-fair: its likelihood changes based on the sex of the applicant. But among all AA-fair decisions, it is closest to the EO-fair decision. The following theorem establishes the theoretical properties of $\hat{Y}^{\text{aa}}$.

Theorem 2 (AA-fairness and optimality of the AA decisions)**.** *The decision $\hat{Y}^{\text{aa}} \sim f_{\text{eo}}(s, a)$ is AA-fair. Moreover, among all AA decisions, the AA predictor minimally modifies the marginal distribution of Y^{eo},*

$$\hat{Y}^{\text{aa}} = \underset{Y^{\text{aa}} \in \mathcal{Y}^{\text{aa}}}{\arg\min}\, \text{KL}(P(\hat{Y}^{\text{eo}}(S, A)) || P(Y^{\text{aa}}(S, A))),$$

where $\mathcal{Y}^{\text{aa}}$ are all AA decisions. It also preserves the marginal distribution of the EO predictor, $P(\hat{Y}^{\text{aa}}) = P(\hat{Y}^{\text{eo}})$. (The proof is in Appendix B.)

In Theorem 2, we also prove that AA-fair decisions preserve the marginal distribution of the EO-fair decisions. This property makes the AA predictor applicable as a decision policy. If the EO-fair predictor admits 20% of the applicants, the AA-fair predictor will also admit 20%. We need not worry that we admit more applicants than we can.

To finish the running example in Fig. 1, we exercise AA and calculate $f_{\text{aa}}(s, a)$ from $f_{\text{eo}}(s, a)$. This is the predictor that requires abduction, calculating the test score that the applicant would have gotten had their sex been different. Consider applicant C. Her EO-fair probability of acceptance is 69%, but when exercising AA it increases to 70%. This adjustment corrects for the (simulated) systemic difficulty of females to receive test preparation and, consequently, higher scores.

Calculating EO and AA predictors. Alg. 1 provides the algorithm for calculating EO and AA predictors. (We prove its correctness in Appendix C.) The EO predictor adjusts the ML predictor f_{ml}; it uses the ML predictor, but marginalizes out the sensitive attribute. The AA predictor adjusts the EO predictor f_{eo}; it uses an abduction step (Pearl, 2009) to impute $a(s')$ from the model $g(\cdot)$ and its residual $a - g(s)$. In admissions, for example, the abduction step infers what a female applicant's test score would have been if she had been male, given the score that she did achieve.

Alg. 1 differs from existing EO and AA algorithms in that it uses all available attributes. Recent works like Kusner et al. (2017) and Kilbertus et al. (2017) avoid using descendants of the sensitive attributes to construct AA-fair decisions. FairLearning (Kusner et al., 2017) performs the same abduction step as f_{aa} to compute residuals $\varepsilon_i = a_i - \mathbb{E}[A \mid S = s_i]$ but then fits a predictor using only the residuals and non-descendants of the sensitive attributes. For example, in admissions, after calculating the residual it does not include the test score in its predictor. In contrast, Alg. 1 uses all available attributes and still is AA-fair. The EO predictor in Alg. 1 also makes use of all attributes; it does not omit sensitive attributes as done in FTU (Kusner et al., 2017). In § 4 we demonstrate empirically that Alg. 1 forms more accurate decisions than these existing EO and AA algorithms.

Multiple sensitive attributes and general causal graphs. The admissions example involves one sensitive attribute, one non-sensitive attribute, and a binary decision. However, EO-fair and AA-fair decisions directly apply to cases with multiple sensitive (and non-sensitive) attributes, non-binary decisions, and more general causal structures.

Notably, AA-fair decisions can correct for bias when an individual belongs to an advantaged group in one sensitive attribute and a disadvantaged one in another. Consider both race and gender as sensitive attributes. If an applicant comes from an advantaged racial group but a disadvantaged gender group, the AA-fair decision would consider the counterfactual test score averaging over both racial and gender groups. It corrects for both the positive racial bias and the negative gender bias.

Algorithm 1: The EO and AA predictors.

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

Output: Predictors $\{f_{\text{ml}}(s,a), f_{\text{eo}}(a), f_{\text{aa}}(s,a)\}$

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

(2) The EO predictor $f_{\text{eo}}(a)$ draws from Eq. 4,

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

(3) The AA predictor $f_{\text{aa}}(s,a)$ draws from Eq. 5,

$$s' \sim p(s); \qquad a' = g(s') + (a - g(s)); \qquad \hat{y}^{\text{aa}} \sim f_{\text{eo}}(s', a).$$

The EO-fair and AA-fair decisions also extend to settings where the bias in some attributes need not be corrected. For example, in affirmative action we might correct for the effect of the applicant's sex on their test scores, but allow applicants of different sex to differ in their interests and other abilities. The AA-fair predictor can calculate counterfactual $A(s)$ on only those attributes we want to correct.

4 Empirical studies

We study algorithmic decisions on simulated and real datasets. We examine the fairness/accuracy trade-off of the EO and AA algorithms presented here, and compare to existing algorithms that target the same EO and AA criterion. (The supplement provides software that reproduces these studies.)

We find the following: 1) the ML decision $\hat{Y}^{\text{ml}}$ is accurate but unfair; 2) The EO decision $\hat{Y}^{\text{eo}}$ is less accurate than $\hat{Y}^{\text{ml}}$, but is EO-fair; 3) The AA decision $\hat{Y}^{\text{aa}}$ is less accurate than the EO-fair decision, but is AA-fair and achieves demographic parity; 4) The fairness through unawareness (FTU) decision $\hat{Y}^{\text{FTU}}$ is EO-fair but less accurate than $\hat{Y}^{\text{eo}}$; 5) The FairLearning decision $\hat{Y}^{\text{FL}}$ is AA-fair and achieves demographic parity, but less accurate than $\hat{Y}^{\text{aa}}$.

Methods and metrics. Each problem consists of training and test sets, potentially with multiple sensitive and non-sensitive attributes. We use Alg. 1 to fit ML, EO and AA predictors with training data. We also use FTU and FairLearning, both from Kusner et al. (2017). The FTU predictor follows classical ML but omits the sensitive attribute; FairLearning is described above.

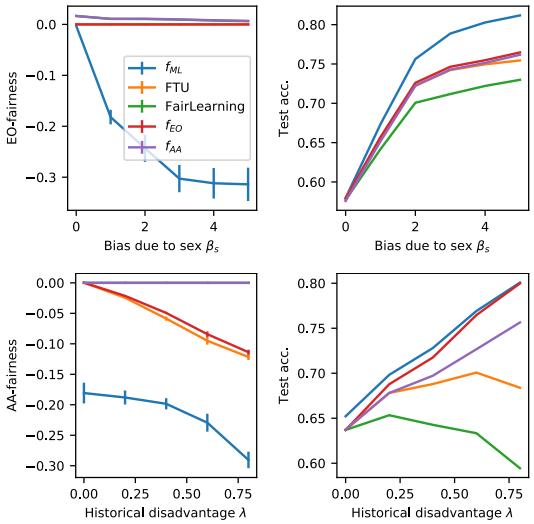
With these predictors, we make algorithmic decisions $\hat{y}$ about the n test set units. We evaluate the accuracy of these decisions relative to the ground truth and we assess their fairness. For a sensitive attribute with values "adv" (for the advantaged group) and "dis" (for the disadvantaged group), we assess fairness with the following metrics:

$$\text{EO metric} = \frac{1}{n}\sum_{i=1}^{n} \left[\mathbb{E}\left[\hat{y}(\text{adv}, a_i) \mid a_i, s_i\right] - \mathbb{E}\left[\hat{y}(\text{dis}, a_i) \mid a_i, s_i\right]\right],$$

$$\text{AA metric} = \frac{1}{n}\sum_{i=1}^{n} \left[\mathbb{E}\left[\hat{y}(\text{adv}, a(\text{adv})) \mid a_i, s_i\right] - \mathbb{E}\left[\hat{y}(\text{dis}, a(\text{dis})) \mid a_i, s_i\right]\right].$$

When the EO metric is close to 0.0, we achieve EO-fairness. Less than 0.0 indicates a bias towards the disadvantaged group. More than 0.0 indicates a bias towards the advantaged group. The same interpretation applies to the AA metric.

Simulation studies. We first study datasets that simulate the unfair admissions committee from Eq. 1. We fix the effect of test score on admissions to $\beta_a = 2.0$. We generate multiple datasets by varying the committee's bias due to sex β_s and the historical disadvantage on test score λ.



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

Figure 3: Both the EO predictor f_{eo} and FTU are EO-fair. The AA predictor f_{aa} and FairLearning (FL) (Kusner et al., 2017) are AA-fair and achieve demographic parity (close-to-zero KL). The ML predictor f_{ml} predicts best overall but the EO predictor predicts best among the EO-fair predictors. We report mean values across individuals. EO and AA metric standard deviations are ≤ 0.1 and ≤ 0.11, respectively.

Figure 4: Varying β_s from 0.0 to $+5.0$; varying λ from 0.0 to $+0.8$; error bars indicate +/- 1 sd. f_{ml} is not fair but predicts best; f_{eo} is the best EO-fair predictor; f_{aa} is the best AA-fair predictor.)

Fig. 4 show how EO-fairness and predictive accuracy trade off as the bias β_s increases. Only the EO predictor and FTU achieve EO fairness. Although both are less accurate than classical ML, the EO predictor achieves greater accuracy than FTU. Fig. 4 show how AA-fairness and predictive accuracy trade off as the historical disadvantage λ increases. Only AA and FairLearning achieve AA-fairness. Among these AA-fair predictors, the AA predictor is more accurate.

Case studies. We study the fair algorithms on three real datasets: 1) Adult income for predicting whether individuals' income is higher than \$50K, 2) ProPublica's COMPAS for predicting recidivism scores 3) German credit for predicting whether individuals have good or bad credit. For Adult and COMPAS, both gender and race are sensitive attributes. For German credit, gender and marital status are sensitive attributes. Decisions are binary except for real-valued COMPAS recidivism scores.

Fig. 3 reports results on the Adult dataset. (In Appendix D, we report results on the COMPAS and German credit datasets, and provide all the details needed to reproduce our studies.) The table reports accuracy, EO, and AA. Further, to measure demographic parity, it reports the KL divergence between decisions $p(\hat{y}(s,a))$ and $p(\hat{y}(s',a))$; when there is demographic parity, the KL is close to zero.

The findings are consistent with the simulation studies. The EO and FTU predictors achieve EO-fairness and, among these, the EO predictor is more accurate. The AA predictor and FairLearning achieve AA-fairness and demographic parity; but the AA predictor is more accurate because it uses all of the attributes. While the ML predictor has the best accuracy, it is less EO-fair and AA-fair.

5 Discussion

We develop fair ML algorithms that modify fitted ML predictors to make them fair. We prove that the resulting predictors are EO-fair or AA-fair, and that they otherwise maximally recover the fitted ML predictor. We show empirically that the EO and AA predictors produce better predictions than existing algorithms for satisfying the same fairness criteria.

The EO and AA predictors both aim to recover the ML outcome as much as possible. However, they can also readily extend to recover the ground truth outcome subject to availability of data. For example, instead of predicting the admissions decision accurately, we can aim the EO and AA predictors for the ground truth outcome, i.e. first year GPA. They will lead to fair and accurate predictions of whether the applicants' succeed in college.

References

Anderson, E. S. (2002). Integration, affirmative action, and strict scrutiny. *NYUL Rev.*, 77, 1195.

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

Calders, T. & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, (pp. 277–292).

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

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, (pp. 153–163).

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

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. In *Neural Information Processing Ssystems*.

Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*.

Kallus, N. & Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*.

Kamiran, F. & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, (pp. 1–33).

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

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

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

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

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

Wang, H., Ustun, B., & Calmon, F. P. (2019). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *arXiv preprint arXiv:1901.10501*.

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

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*.

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

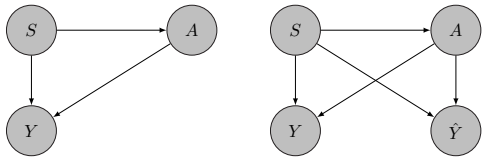
Appendix

A Proof of Theorem 1

Before proving the EO-fairness and optimality of the EO predictor, we first establish a lemma about EO-fairness. It says the EO-fairness of a decision $\hat{Y}$ is equivalent to no causal arrow from S_i to $\hat{Y}$.

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

Consider a decision $\hat{Y}(s, a)$ in the causal graphical model Fig. 5b. The reason behind Lemma 3 is that $\hat{Y}(s, a) \perp S, A$ in Fig. 5b (Pearl, 2009). So EO reduces to $\hat{Y}(s, a)$ being constant in s.



(a) **(b)**

Figure 5: The causal graph for the sensitive attribute S, the attribute A, the past decisions Y, and the predicted decisions $\hat{Y}$.

Lemma 3 shows that the decision $\hat{Y}$ is EO-fair as long as $\hat{Y}$ does not predict using the sensitive attribute.

Proof. The goal is to show that counterfactual fairness is equivalent to

$$\hat{Y}^{\text{eo}}(s, a) = \hat{Y}^{\text{eo}}(a) \tag{6}$$

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

We start with the definition of EO-fairness.

$$P(\hat{Y}(s', a) \,|\, S = s, A = a,) = P(\hat{Y}(s, a) \,|\, S = s, A = a) \tag{7}$$

$$\Leftrightarrow P(\hat{Y}(s, a')) = P(\hat{Y}(s, a)) \qquad \forall a, s, s' \tag{8}$$

$$\Leftrightarrow P(\hat{Y}(s', a)) = P(\hat{Y}(a)) \tag{9}$$

Eq. 8 is due to the observation that $\hat{Y}(s, a) \perp A, S$. Eq. 9 is true because

$$\hat{Y}(a) = \int \hat{Y}(s, a) p_S(s) \, \mathrm{d}s = \hat{Y}(s', a) \tag{10}$$

for any s'. The last equation is equivalent to no causal arrow from S to $\hat{Y}$.

□

Now we are ready to prove Theorem 1.

Proof. As a direct application of Lemma 3, we prove the first part of Theorem 1: $\hat{Y}^{\text{eo}}$ is EO-fair. The reason is that $\hat{Y}^{\text{eo}}(s_{\text{new}}, a_{\text{new}}) = \hat{Y}^{\text{eo}}(a_{\text{new}})$ as is defined in Eq. 4.

We then prove the second part of Theorem 1. It establishes the optimality of $\hat{Y}^{\text{eo}}$.

We start with rewriting the goal of the proof. We will show this goal is equivalent to the definition of the EO predictor.

$$\hat{Y}^{\mathrm{eo}} = \underset{Y^{\mathrm{eo}} \in \mathcal{Y}^{\mathrm{eo}}}{\arg\min}\, \mathbb{E}_{S,A}\left[\mathrm{KL}(P(\hat{Y}^{\mathrm{ml}}(S,A)) || P(Y^{\mathrm{eo}}(S,A))\right] \tag{11}$$

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

$$= \underset{Y^{\mathrm{eo}} \in \mathcal{Y}^{\mathrm{eo}}}{\arg\min}\, -\mathbb{E}_{A}\left[\int P(S) \int P(\hat{Y}^{\mathrm{ml}}(S,A)) \log P(Y^{\mathrm{eo}}(S,A)) \,\mathrm{d}y\, \mathrm{d}s\right] \tag{13}$$

$$= \underset{Y^{\mathrm{eo}} \in \mathcal{Y}^{\mathrm{eo}}}{\arg\min}\, -\mathbb{E}_{A}\left[\int P(S) \int P(\hat{Y}^{\mathrm{ml}}(S,A)) \log P(Y^{\mathrm{eo}}(A)) \,\mathrm{d}y\, \mathrm{d}s\right] \tag{14}$$

$$= \underset{Y^{\mathrm{eo}} \in \mathcal{Y}^{\mathrm{eo}}}{\arg\min}\, -\mathbb{E}_{A}\left[\int \left[\int P(S) P(\hat{Y}^{\mathrm{ml}}(S,A)) \,\mathrm{d}s\right] \log P(Y^{\mathrm{eo}}(A)) \,\mathrm{d}y\right] \tag{15}$$

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

$$= \underset{Y^{\mathrm{eo}} \in \mathcal{Y}^{\mathrm{eo}}}{\arg\min}\, \mathbb{E}_{A}\left[\mathrm{KL}(P(\hat{Y}^{\mathrm{ml}}\,;\, \mathrm{do(A = A)}) || \mathrm{P(Y^{eo}(A))}\right] \tag{17}$$

$$= P(\hat{Y}^{\mathrm{ml}}\,;\, \mathrm{do(a)}) \tag{18}$$

$$= \int p(\hat{y}^{\mathrm{ml}}(s,a)) p(s)\, \mathrm{d}s \tag{19}$$

Eq. 11 is the goal of the proof. Eq. 12 is due to the definition of the Kullback-Leibler (KL) divergence. Eq. 13 is due to $\hat{Y}^{\mathrm{acc}}$ being a given random variable. Eq. 14 is due to Lemma 3. Eq. 15 switches the integral subject to conditions of the dominated convergence theorem. Eq. 16 is due to Fig. 5b and the backdoor adjustment formula of Pearl (2009). Eq. 17 is due to the definition of the KL divergence and $\hat{Y}^{\mathrm{ml}}$ being given. Eq. 18 is because setting $P(Y^{\mathrm{eo}}(a)) = P(\hat{Y}^{\mathrm{ml}}\,;\, \mathrm{do(a)})$ has $\mathrm{KL}(P(\hat{Y}^{\mathrm{ml}}\,;\, \mathrm{do(A = a)}) || \mathrm{P(Y^{eo}(a))} = 0$ for all a. The expecation is hence also zero: $\mathbb{E}_A\left[\mathrm{KL}(P(\hat{Y}^{\mathrm{ml}}\,;\, \mathrm{do(A = A)}) || \mathrm{P(Y^{eo}(A))}\right] = 0$. Eq. 19 is due to the definition of the intervention distribution (Pearl, 2009).

This calculation implies $\hat{Y}^{\mathrm{eo}} = \int p(\hat{y}^{\mathrm{ml}}(s,a)) p(s)\, \mathrm{d}s$ minimizes the average KL between the ML decision and the EO decision $\mathbb{E}_{A,S}\left[\mathrm{KL}(P(\hat{Y}^{\mathrm{ml}}(S,A)) || P(\hat{Y}^{\mathrm{eo}}(S,A))\right]$. the EO predictor maximally recovers the ML decision. Put differently, the EO predictor minimally modifies the ML decision to achieve EO-fairness.

□

B Proof of Theorem 2

Proof. We first prove that the AA predictor is AA-fair.

Recall the definition of the AA predictor:

$$p(\hat{y}^{\mathrm{aa}}(s_{\mathrm{new}}, a_{\mathrm{new}})) = \int\int p(\hat{y}^{\mathrm{eo}}(a(s^0)))\; p(a(s^0) \mid s_{\mathrm{new}}, a_{\mathrm{new}})\; p(s^0)\, \mathrm{d}a(s^0)\, \mathrm{d}s^0. \tag{20}$$

Note the above equation is the same definition as in Eq. 5. We simply change s to s^0 for downstream notation convenience.

This implies

$$p(\hat{y}^{\mathrm{aa}}(s', a(s'))) = \int\int p(\hat{y}^{\mathrm{eo}}(a(s^0)))\; p(a(s^0) \mid s', a(s'))\; p(s^0)\, \mathrm{d}a(s^0)\, \mathrm{d}s^0. \tag{21}$$

Hence we have

$$p(\hat{y}^{\text{aa}}(s', a(s')) \mid S = s, A = a) \tag{22}$$

$$= \int \int p(\hat{y}^{\text{eo}}(a(s^0))) \; p(a(s^0) \mid s, a(s')') \; p(s^0) p(a(s') \mid s, a) \, \mathrm{d}a(s') \, \mathrm{d}a(s^0) \, \mathrm{d}s^0 \tag{23}$$

$$= \int \int p(\hat{y}^{\text{eo}}(a(s^0))) \; p(a(s^0) \mid s, a) \; p(s^0) \, \mathrm{d}a(s^0) \, \mathrm{d}s^0. \tag{24}$$

Eq. 24 is due to $\int p(a(s^0) \mid s', a(s')) p(a(s') \mid s, a) \, \mathrm{d}a(s') = p(a(s^0) \mid s, a)$ by the structural equation implied by the causal graph:

$$A \overset{a.s.}{=} f(S, \epsilon). \tag{25}$$

The intuition here is the observed s, a will provide the same information as $s', a(s')$, for the same person. They all contain the information about the background variable that affects A. We hence have $p(a(s^0) \mid s', a(s')) = p(a(s^0) \mid a(s'), s', a, s) = p(a(s^0) \mid a(s'), a, s)$.

Notice that the right hand side of Eq. 24 does not depend on s'. This implies $p(\hat{y}^{\text{aa}}(s', a(s')) \mid S = s, A = a) = p(\hat{y}^{\text{aa}}(a(s), s) \mid A = a, S = s)$. We simply repeat the same calculation for $p(\hat{y}^{\text{aa}}(s, a(s)) \mid S = s, A = a)$.

We hence have proved that the AA predictor is AA-fair:

$$P(\hat{Y}^{\text{aa}}(s', A(s')) \mid S = s, A = a) = P(\hat{Y}^{\text{aa}}(s, A(s)) \mid S = s, A = a). \tag{26}$$

It establishes the first part of Theorem 2.

We can also prove demographic parity for $\hat{Y}^{\text{aa}}$.

$$P(\hat{Y}^{aa}(S, A) \mid S) \tag{27}$$

$$= \int P(Y^{\text{eo}}(a')) P(A(s') = a' \mid S, A) P(s') \, \mathrm{d}a' \, \mathrm{d}s' P(A \mid S) \, \mathrm{d}A \tag{28}$$

$$= \int P(Y^{\text{eo}}(a')) P(A(s') = a' \mid S) P(s') \, \mathrm{d}a' \, \mathrm{d}s' \tag{29}$$

$$= \int P(Y^{\text{eo}}(a')) P(A(s') = a') P(s') \, \mathrm{d}a' \, \mathrm{d}s' \tag{30}$$

$$\tag{31}$$

The third equality is due to $A(s') \perp S$ by Pearl's twin network (Pearl, 2009).

We next compute the marginal distribution of $\hat{Y}^{aa}$.

$$P(\hat{Y}^{aa}(S, A)) \tag{32}$$

$$= \int \int P(Y^{\text{eo}}(a')) P(A(s') = a' \mid S, A) P(s') \, \mathrm{d}a' \, \mathrm{d}s' P(S, A) \, \mathrm{d}A \, \mathrm{d}S \tag{33}$$

$$= \int P(Y^{\text{eo}}(a')) P(A(s') = a') P(s') \, \mathrm{d}a' \, \mathrm{d}s' \tag{34}$$

$$\tag{35}$$

Therefore, we have

$$P(\hat{Y}^{aa}(S, A) \mid S) = P(\hat{Y}^{aa}(S, A)). \tag{36}$$

Hence,

$$\hat{Y}^{aa}(S, A) \perp S. \tag{37}$$

This establishes demographic parity.

We next prove the second part of Theorem 2. We show that Y^{aa} minimizes $\mathrm{KL}(P(Y^{\mathrm{eo}})||P(\hat{Y}^{\mathrm{aa}}))$. In fact, Y^{aa} recovers the marginal distribution of $\hat{Y}^{\mathrm{eo}}$.

$$
\begin{aligned}
P(\hat{Y}^{aa}) =& P(\hat{Y}^{aa}(S, A)) && (38)\\
=& \int P(Y^{\mathrm{eo}}(a'))P(A(s') = a')P(s')\,\mathrm{d}a'\,\mathrm{d}s' && (39)\\
=& \int P(Y^{\mathrm{eo}}(a'))P(A(s') = a')P(s')\,\mathrm{d}a'\,\mathrm{d}s' && (40)\\
=& \int P(Y^{\mathrm{eo}}(a'))P(A = a')\,\mathrm{d}a' && (41)\\
=& P(\hat{Y}^{\mathrm{eo}}(A)) && (42)\\
=& P(\hat{Y}^{\mathrm{eo}}) && (43)
\end{aligned}
$$

This gives $\mathrm{KL}(P(Y^{\mathrm{eo}})||P(\hat{Y}^{\mathrm{aa}})) = 0$. Hence, $\hat{Y}^{aa}$ minimizes this KL and preserves the marginal distribution of $\hat{Y}^{\mathrm{eo}}$.

□

C The correctness of Alg. 1

We leverage the backdoor adjustment formula in Pearl (2009) to compute $\hat{Y}^{\mathrm{eo}}$ from the past admissions records $\{(S_i, A_i, Y_i)\}_{i=1}^n$.

Proposition 4. *Assume the causal graph Fig. 5a. The* EO *predictor can be computed using the observed data* (S_i, A_i, Y_i)*:*

$$
P(\hat{Y}^{\mathrm{eo}}(s_{\mathrm{new}}, a_{\mathrm{new}})) = \int P(Y_i \,|\, S_i = s', A_i = a_{\mathrm{new}})P(S_i = s')\,\mathrm{d}s', \tag{44}
$$

if

$$
P(A_i \in \mathbf{A} \,|\, S_i = s) > 0 \tag{45}
$$

for all $P(\mathbf{A}) > 0$*,* $s \in \mathcal{S}$ *and* $\mathbf{A} \subset \mathcal{A}$*.*

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

Eq. 44 reiterates the fact that $\hat{Y}^{\mathrm{eo}}$ does not simply ignore the sensitive attribute S; estimating $P(Y_i \,|\, S_i, A_i)$ relies on S in the training data. However, $\hat{Y}^{\mathrm{eo}}(s_{\mathrm{new}}, a_{\mathrm{new}})$ does not rely on s_{new} in the test data.

We leverage the abduction-action-prediction approach (Pearl, 2009) to compute $\hat{Y}^{\mathrm{aa}}$.

Proposition 5. *Assume the causal graph Fig. 5a. Write*

$$
A \overset{a.s}{=} f_A(S, \epsilon_A) \tag{46}
$$

for some function f_A *and zero mean random variable* ϵ_A *satisfying* $S \perp \epsilon_A$*. The* AA *predictor can be computed using the observed data* (S_i, A_i, Y_i)*:*

$$
\begin{aligned}
& P(\hat{Y}^{\mathrm{aa}}(s, a))\\
=& \int P(Y^{\mathrm{eo}}(s, a')) \cdot P(A = a' \,|\, S = s', \epsilon_A) \cdot P(\epsilon_A \,|\, S = s, A = a) \cdot P(s')\,\mathrm{d}a'\,\mathrm{d}s' && (47)\\
\approx& \int P\left(Y^{\mathrm{eo}}\left(\mathbb{E}\left[A_i \,|\, S_i = s'\right] + \hat{\epsilon}_A\right), s\right) \cdot P(s')\,\mathrm{d}s' && (48)
\end{aligned}
$$

where $\hat{\epsilon}_A = a - \mathbb{E}\left[A_i \,|\, S_i = s\right]$.

	Metrics ($\times 10^{-2}$) on COMPAS			
	EO	**AA**	**KL**	**Prediction**
ML predictor f_{ml}	-68.1	-104.9	17.1	28.0
FTU	**0**	-52.9	0.5	**25.6**
EO predictor f_{eo}	**0**	-36.8	0.5	**25.6**
FairLearning	-41.2	**0**	**0.2**	**22.7**
AA predictor f_{aa}	-41.2	**0**	**0.2**	**22.7**

Table 1: Both the EO predictor f_{eo} and FTU are EO-fair; they achieve zero in the EO metric (lower is better). AA predictor f_{aa} and FairLearning Kusner et al. (2017) are AA-fair; they achieve zero in the AA metric (lower is better). AA predictor f_{aa} achieves demographic parity; it has close-to-zero KL divergence (Lower is better.) ML predictor f_{ml} predicts best; EO predicts best among the fair predictors (higher prediction scores are better.) We report mean values across individuals. EO and AA metric standard deviations are ≤ 0.42 and ≤ 0.6, respectively.

Proof. Eq. 47 is a direct consequence of Theorem 7.1.7 of Pearl (2009). Eq. 48 is due to a linear approximation of f_A. □

Alg. 1 can easily generalize to cases where sensitive attributes are unobserved. In these cases, we can leverage recent techniques in multiple causal inference like the approach from Wang & Blei (2018). These approaches construct substitutes for the unobserved sensitive attributes from the descendants of the sensitive attributes.

D Detailed results of the empirical studies

We study the fair algorithms on three real datasets:

- **Adult income dataset.**[4] The task is to predict whether someone has a yearly income higher than \$50K; the decision is binary. The sensitive 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 predict an individual's recidivism score; the decision is real-valued. The sensitive 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.**[5] The task is to predict whether an individual has good or bad credit; the decision is binary. The sensitive 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.

We report the detailed results of the COMPAS dataset and the German credit dataset in Table 1 and Table 2.

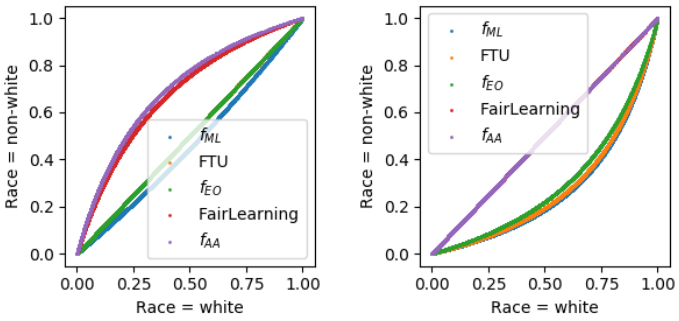
We further discuss EO-fairness and AA-fairness. Consider the EO-fairness against race. For the Adult dataset, Fig. 6a plots $\mathbb{E}[\hat{Y}(s,a) \,|\, S=s, A=a]$ against $\mathbb{E}[\hat{Y}(s',a) \,|\, S=s, A=a]$, where s and s' only differ by the individual's race being white or non-white. A method is EO-fair if the predictions align with the diagonal. Both the EO predictor and FTU align with the diagonal; they are EO-fair. None of classical ML, FairLearning, or the AA predictor are EO-fair, and counterfactual AA and FairLearning are less EO-fair than classical ML.

[4] https://www.kaggle.com/wenruliu/adult-income-dataset

[5] https://www.kaggle.com/uciml/german-credit

	Metrics ($\times 10^{-2}$) on German Credit			
	EO	**AA**	**KL**	**Prediction**
ML predictor f_{ml}	-5.4	-3.9	18.6	64.7
FTU	**0**	-2.0	15.6	63.4
EO predictor f_{eo}	**0**	1.5	13.2	**64.5**
FairLearning	-1.4	**0**	8.3	63.5
AA predictor f_{aa}	-1.3	**0**	**7.8**	**64.3**

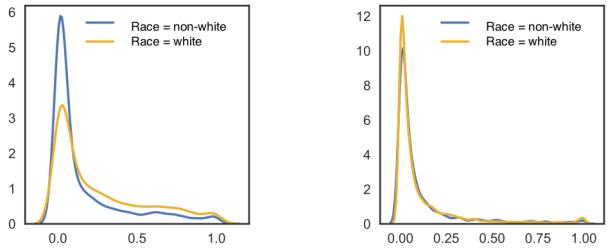
Table 2: Both EO predictor f_{eo} and FTU are EO-fair; they achieve zero in the EO metric (lower is better). AA predictor f_{aa} and FairLearning Kusner et al. (2017) are AA-fair; they achieve zero in the AA metric (lower is better). AA predictor f_{aa} achieves demographic parity; it has close-to-zero KL divergence (Lower is better.) ML predictor f_{ml} predicts best; EO predicts best among the fair predictors (higher prediction scores are better.) We report mean values across individuals. EO and AA metric standard deviations are ≤ 0.02 and ≤ 0.02, respectively.



(a) EO metric **(b)** AA metric

Figure 6: The EO predictor satisfies the EO criterion; the AA predictor satisfies the AA criterion. We compare counterfactual predictions for all individuals had they been white or non-white. (a) relates to the EO metric: it holds the attribute values fixed. (b) allows attribute values to change after intervening on race.

Now consider the AA-fairness against race. For the Adult dataset, Fig. 6b plots $\mathbb{E}[\hat{Y}(s, A(s)) \mid S = s, A = a]$ against $\mathbb{E}[\hat{Y}(s', A(s')) \mid S = s, A = a]$ when s and s' differ by whether the individual is white or non-white. If the predictions align with the diagonal, the decision is AA-fair. Both the AA predictor and FairLearning align with the diagonal; they are counterfactually fair. Counterfactual EO is less unfair than FTU. The classical ML predictor is most unfair. (These results generalize across datasets.)



(a) The ML predictor **(b)** The AA predictor

Figure 7: The decision distributions of white and non-white individuals. The AA predictor produces equal distributions; it achieves demographic parity.

Finally, we discuss demographic parity. Demographic parity is a group-level statistical measure that requires decisions to be independent of sensitive attributes. It has been used to measure affirmative action (Dwork et al., 2012). To evaluate demographic parity, we compare the prediction distributions between the groups of individuals with different sensitive attributes. The metric is the symmetric KL divergence between $P(\hat{Y} \mid S = s)$ and $P(\hat{Y} \mid S = s')$. For a predictor achieving demographic parity, the KL is zero. (We evaluate the symmetric KL by binning the values of predictions.)

Fig. 3 present the symmetric KL between the two gender groups in the data. Both the AA predictor and FairLearning have close-to-zero symmetric KL. The AA predictor generally has lower symmetric KL; it is closer to demographic parity. None of the other methods are close.

Now consider the demographic parity against race. Fig. 7 demonstrates the prediction distributions of white and non-white individuals. While classical ML has very different prediction distributions for the two groups, the prediction distributions of the AA predictor is nearly identical.

References

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

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

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

Wang, Y. & Blei, D. M. (2018). The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*.