

# Counterfactual Fairness

Matt J. Kusner<sup>\*12</sup> Joshua R. Loftus<sup>\*13</sup> Chris Russell<sup>\*14</sup> Ricardo Silva<sup>15</sup>

## Abstract

Machine learning has matured to the point to where it is now being considered to automate decisions in loan lending, employee hiring, and predictive policing. In many of these scenarios however, previous decisions have been made that are unfairly biased against certain subpopulations (e.g., those of a particular race, gender, or sexual orientation). Because this past data is often biased, machine learning predictors must account for this to avoid perpetuating discriminatory practices (or incidentally making new ones). In this paper, we develop a framework for modeling fairness in any dataset using tools from counterfactual inference. We propose a definition called *counterfactual fairness* that captures the intuition that a decision is fair towards an individual if it gives the same predictions in (a) the observed world and (b) a world where the individual had always belonged to a different demographic group, other background causes of the outcome being equal. We demonstrate our framework on two real-world problems: fair prediction of law school success, and fair modeling of an individual’s criminality in policing data.

## 1. Introduction

Machine learning has spread to fields as diverse as credit scoring (Khandani et al., 2010), crime prediction (Brennan et al., 2009), and loan assessment (Mahoney & Mohe, 2007). As machine learning enters these new areas it is necessary for the modeler to think beyond the simple objective of maximizing prediction accuracy, and to consider the societal impact of their work.

For many of these applications, it is crucial to ask if the predictions of a model are *fair*. For instance, imagine a bank wishes to predict if an individual should be given a loan

to buy a house. The bank wishes to use historical repayment data, alongside individual data. If they simply learn a model that predicts whether the loan will be paid back, it may unjustly favor applicants of particular subgroups, due to past and present prejudices. The Obama Administration released a report describing this which urged data scientists to analyze “how technologies can deliberately or inadvertently perpetuate, exacerbate, or mask discrimination”.<sup>1</sup>

As a result, there has been immense interest in designing algorithms that make fair predictions (Bolukbasi et al., 2016; Calders & Verwer, 2010; Dwork et al., 2012; Grgic-Hlaca et al., 2016; Hardt et al., 2016; Joseph et al., 2016; Kamiran & Calders, 2009; 2012; Kamishima et al., 2011; Kleinberg et al., 2016; Louizos et al., 2015; Zafar et al., 2015; 2016; Zemel et al., 2013; Zliobaite, 2015). In large part, the initial work on fairness in machine learning has focused on formalizing fairness into quantitative definitions and using them to solve a discrimination problem in a certain dataset. Unfortunately, for a practitioner, law-maker, judge, or anyone else who is interested in implementing algorithms that control for discrimination, it can be difficult to decide which definition of fairness to choose for the task at hand. Indeed, we demonstrate that depending on the relationship between a sensitive attribute and the data, certain definitions of fairness can actually *increase discrimination*.

We describe how techniques from causal inference can be effective tools for designing fair algorithms and argue, as in (DeDeo, 2014), that it is essential to properly address causality. Specifically, we leverage the causal framework of Pearl et al. (2009) to model the relationship between sensitive attributes and data. Our contributions are as follows:

1. We model questions of fairness within a causal framework. This allows us to directly model *how* unfairness affects the data at hand.
2. We introduce *counterfactual fairness*, which enforces that a distribution over possible predictions for an individual should remain unchanged, in a world where an individual’s sensitive attribute had been different from birth.
3. We analyze how enforcing existing definitions of fair-

<sup>\*</sup>Equal contribution, author order decided randomly <sup>1</sup>Alan Turing Institute <sup>2</sup>University of Warwick <sup>3</sup>University of Cambridge <sup>4</sup>University of Edinburgh <sup>5</sup>University College London. Correspondence to: <mkusner@turing.ac.uk>, <jloftus@turing.ac.uk>, <crussell@turing.ac.uk>, <ricardo.silva@ucl.ac.uk>.

<sup>1</sup><https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights>

ness for different data may correspond or be in conflict with counterfactual fairness. In particular, we show that depending on the underlying state of the world some definitions of fairness may be inappropriate.

4. We devise techniques for learning predictors that are counterfactually fair and demonstrate their use in several examples.

## 2. Fairness

Our goal in this paper is to design automated algorithms that make fair predictions across various demographic groups. This unfairness can arise in several ways:

**Historically biased distributions:** Individuals with different protected attributes  $A$  may have many different attributes due to current and historic biases (e.g., racial inequality caused by things like colonialism, slavery, a history of discrimination in hiring and housing etc.).

**Selection unfairness:** The training data could contain selection bias. For instance, if we are using a dataset describing who paid loans back in full in order to train a loan prediction algorithm, it may be that loans were unfairly distributed. Since we can't see whether people will pay back a loan if they didn't receive one, our algorithms may be biased by this sampling.

**Prediction unfairness:** The learned classifier could use either protected attributes such as race or correlated attributes as features, and learn a biased predictor.

There has been a wealth of recent work towards fair algorithms. These include fairness through unawareness (Grgic-Hlaca et al., 2016), demographic parity/disparate impact (Zafar et al., 2015), individual fairness (Dwork et al., 2012; Joseph et al., 2016; Louizos et al., 2015; Zemel et al., 2013), and equality of opportunity (Hardt et al., 2016; Zafar et al., 2016).

**Definition 1** (Fairness Through Unawareness (FTU)). *An algorithm is fair so long as any sensitive attributes  $A$  are not explicitly used in the decision-making process. Any mapping  $\hat{Y} : X \rightarrow Y$  that excludes  $A$  (or other unfair attributes, see Grgic-Hlaca et al. (2016)) satisfies this.*

Initially proposed as a baseline method, the approach has found favor recently with more general approaches such as Grgic-Hlaca et al. (2016). The approach has a compelling simplicity, and constructs a predictor  $\hat{Y}$  based on a feature vector  $X$  that excludes  $A$ , and in the case of Grgic-Hlaca et al. (2016) other attributes labeled as unfair.

**Definition 2** (Individual Fairness (IF)). *An algorithm is fair if it gives similar predictions to similar individuals.*

*Formally, if individuals  $i$  and  $j$  are similar apart from their protected attributes  $A_i, A_j$  then*

$$\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)}).$$

This approach can be understood loosely as a continuous analog of FTU. As described in (Dwork et al., 2012), the notion of similarity must be carefully chosen and this notion of fairness will not correct for the historical biases described above.

**Definition 3** (Demographic Parity (DP)). *An algorithm is fair if its predictions are independent of the sensitive attributes  $A$  across the population. A prediction  $\hat{Y}$  satisfies this definition if,*

$$P(\hat{Y}|A=0) = P(\hat{Y}|A=1).$$

**Definition 4** (Equal Opportunity (EO)). *An algorithm is fair if it is equally accurate for each value of the sensitive attribute  $A$ . A prediction  $\hat{Y}$  satisfies this if,*

$$P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1).$$

While these definitions address the notion of algorithmic fairness, they guarantee that historic biases in the data are preserved. As shown by Kleinberg et al. (2016), EO and DP are mutually exclusive notions of fairness.

## 3. Causal Models and Counterfactuals

We follow the framework of Pearl (2000), and define a causal model as a triple  $(U, V, F)$  of sets such that

- $U$  is a set of latent **background** variables<sup>2</sup>, which are generated by factors outside of our control, and in general do not depend on any protected attributes  $A$  (unless this is explicitly specified);
- $V$  is a set of **endogenous** variables, where each member is determined by other variables in  $U \cup V$ ;
- $F$  is a set of functions  $\{f_1, \dots, f_n\}$ , one for each  $V_i \in V$ , such that  $V_i = f_i(pa_i, U_{pa_i})$ ,  $pa_i \subseteq V \setminus \{V_i\}$  and  $U_{pa_i} \subseteq U$ . Such equations are also known as **structural equations** (Bollen, 1989).

The notation “ $pa_i$ ” refers to the “parents” of  $V_i$  and is motivated by the assumption that the model factorizes according to a directed acyclic graph (DAG). That is, we can define a directed graph  $\mathcal{G} = (U \cup V, \mathcal{E})$  where each node is an element of  $U \cup V$ , and each edge from some  $Z \subseteq U \cup V$

<sup>2</sup>These are sometimes called **exogenous variables**, but the fact that members of  $U$  might depend on each other is not relevant to what follows.

to  $V_i$  indicates that  $Z \in pa_i \cup U_{pa_i}$ . By construction,  $\mathcal{G}$  is acyclic.

The model is causal in that, given a distribution  $p(U)$  over the background variables  $U$ , you can derive the distribution of a subset  $Z \subseteq V$  following an **intervention** on the complementary subset  $V \setminus Z$ . Here, an **intervention** on the variable  $V_i$  of value  $v$  refers to the substitution of equation  $V_i = f_i(pa_i, U_{pa_i})$  with the equation  $V_i = v$ . This captures the idea of an agent, external to the system, modifying it by forcefully assigning value  $v$  to  $V_i$ . This occurs in a randomized controlled trials where the value of  $V_i$  is overridden by a treatment setting it to  $v$ , a value chosen at random, and thus independent of any other causes.

In contrast with the independence constraints given by a DAG, the full specification of  $F$  requires much stronger assumptions but also leads to much stronger claims. In particular, it allows for the calculation of **counterfactual** quantities. In brief, consider the following counterfactual statement, “the value of  $Y$  if  $Z$  had taken value  $z$ ”, for two endogenous variables  $Z$  and  $Y$  in a causal model. By assumption, the state of any endogenous variable is fully determined by the background variables and structural equations. The counterfactual is modeled as the solution for  $Y$  for a given  $U = u$  where the equations for  $Z$  are replaced with  $Z = z$ . We denote it by  $Y_{Z \leftarrow z}(u)$  (Pearl, 2000), and sometimes as  $Y_z$  if the context of the notation is clear.

Counterfactual inference, as specified by a causal model  $(U, V, F)$  given evidence  $W$ , is the computation of probabilities  $P(Y_{Z \leftarrow z}(U) \mid W = w)$ , where  $W$ ,  $Z$  and  $Y$  are subsets of  $V$ . Inference proceeds in three steps, as explained in more detail in Chapter 4 of Pearl et al. (2016):

1. Abduction: for a given prior on  $U$ , compute the posterior distribution of  $U$  given the evidence  $W = w$ ;
2. Action: substitute the equations for  $Z$  with the interventional values  $z$ , resulting in the modified set of equations  $F_z$ ;
3. Prediction: compute the implied distribution on the remaining elements of  $V$  using  $F_z$  and the posterior  $P(U \mid W = w)$ .

#### 4. Counterfactual Fairness

Given a causal model  $(U, V, F)$ , let  $A \subseteq V$  be a set of protected attributes,  $\hat{Y} \subseteq V$  a variable which we will be the basis for any decision making, and  $W$  the set of complementary measurements such that  $W = V \setminus (A \cup \{\hat{Y}\})$ .

**Definition 5** (Counterfactual fairness). *We say  $\hat{Y}$  is **counterfactually fair** if under any context uniquely defined by*

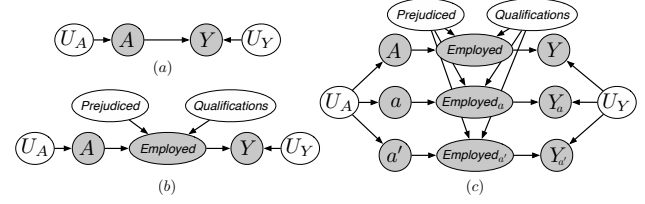


Figure 1. (a) The graph corresponding to a causal model with  $A$  being the protected attribute and  $Y$  some outcome of interest, with background variables assumed to be independent. (b) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (latent) background variables **Prejudiced** (if the person offering the job is prejudiced) and **Qualifications** (a measure of the individual’s qualifications). (c) A twin network representation of this system (Pearl, 2000) under two different counterfactual levels for  $A$ . This is created by copying nodes descending from  $A$ , which inherit unaffected parents from the factual world.

evidence  $W = w$  and sensitive  $A = a$ ,

$$\begin{aligned} P(\hat{Y}_{A \leftarrow a}(U) = y \mid W = w, A = a) = \\ P(\hat{Y}_{A \leftarrow a'}(U) = y \mid W = w, A = a), \end{aligned} \quad (1)$$

for all  $y$  and for any value  $a'$  attainable by  $A$ .

This captures the idea that any decision based on the conditional distribution of  $\hat{Y}$  would be the same despite  $A$  being different, given the full implications of  $A$  having always been different. We can also see  $\hat{Y}$  as satisfying “counterfactual exchangeability” under this model.

An associated concept of causal fairness appears as Example 4.4.4 in Pearl et al. (2016). There, the authors condition instead on  $W$ ,  $A$ , and the observed realization of  $\hat{Y}$ , and calculate the probability of the counterfactual realization differing from the factual<sup>3</sup>. This example conflates the recorded decision  $\hat{Y}$  with the information  $Y$  on which we should ideally base our decision making, a difference which we maintain. Our framing makes the connection to other existing machine learning methods more explicit, as we discuss in Section 5. Evidence used to determine the state of background variables  $U$  should come from  $A$  and  $W$  alone, as in many setups we wish to predict some  $Y$  as  $\hat{Y}$ , when  $Y$  is unavailable at any point in our inference.

We also emphasize that counterfactual fairness is an individual-level definition. This is substantially different from comparing different units that happen to share the same “treatment” and coincide on values of  $X$ , as discussed in Section 4.3.1 of (Pearl et al., 2016). Here, differences in the value of  $X$  must be caused by variations on  $A$  only.

<sup>3</sup>The result is an expression called the “the probability of sufficiency” for  $A$ , capturing the notion that switching  $A$  to a different value would be sufficient to change  $\hat{Y}$  with some probability.

#### 4.1. Implications

As discussed by Halpern (2016), it is unproductive to debate if a particular counterfactual definition is the “correct” one to satisfy socially constructed concepts such as blame and responsibility. The same applies to fairness. Instead, we discuss the implications of definition (5) and some choices that arise in its application.

First, we wish to make explicit the difference between  $\hat{Y}$ , the predictor we use for fair decisions, and  $Y$ , the related state generated by an unfair world. For instance,  $Y$  could be an indicator of whether a client defaults on a loan, while  $\hat{Y}$  is the actual decision of giving the loan. Consider the DAG  $A \rightarrow Y$  for a causal model where  $V = \{A, Y\}$ , and in Figure 4(a) the DAG with explicit inclusion of set  $U$  of independent background variables. Assume  $Y$  is an objectively ideal measure used in decision making, such as a binary indicator that the individual defaults on a loan. In this setup, the mechanism  $f_Y(A, U)$  is causally unfair, with the arrow  $A \rightarrow Y$  being the result of a world that punishes individuals in a way that is out of their control. Figure 4(b) shows a more fine-grained model, where the path is mediated by a measure of whether the person is employed, which is itself caused by two background factors: one representing whether the person hiring is prejudiced, and the other the employee’s qualifications. In this world,  $A$  is a cause of defaulting, even if mediated by other variables. The counterfactual fairness principle however forbids us from using  $Y$ : using the twin network of Pearl (2000), we see in Figure 4(c) that  $Y_a$  and  $Y_{a'}$  need not be identically distributed given the background variables. For example, if the function determining employment  $f_E(A, P, Q) = I_{(Q>0, P=0 \text{ or } A \neq a)}$  then an individual with sufficient qualifications and prejudiced potential employer may have a different counterfactual employment value for  $A = a$  compared to  $A = a'$ , and a different chance of default.

In contrast, any function of variables not descendants of  $A$  can be used as a basis for fair decision making. This means, that any variable  $\hat{Y}$  defined by  $\hat{Y} = g(U)$  will be counterfactually fair for any function  $g(\cdot)$ . Hence, given a causal model, the functional defined by the function  $g(\cdot)$  minimizing some predictive error for  $Y$  will satisfy the criterion. If  $\hat{Y}$  must be randomized, it suffices that the stochastic component of it is independent of any descendant of  $A$ .

There is a subtlety to address here: by abduction,  $U$  will typically depend on  $A$ , and hence so will  $\hat{Y}$  when marginalizing over  $U$ . This seems to disagree with the intuition that our fair variable should be not be caused by  $A$ . However, this is a comparison *across individuals*, not within an individual, as discussed by Section 4.3.1 of (Pearl et al., 2016). More intuitively, consider the simple case where  $U$  is fully determined by  $A$  and  $X$  (which occurs in some important

special cases). In this scenario, we proceed just as if we have *measured*  $U$  from the beginning rather than performing abduction. We then generate  $\hat{Y}$  from  $g(U)$ , so  $U$  is the cause of  $\hat{Y}$  and not  $A$ .

Note that we can build counterfactually fair predictive models for some  $\hat{Y}$  even if the structural equations that generated  $Y$  are unfair. The idea is that we are learning a projection of  $Y$  into an alternate world where it would be fair, which we may think of as a “closest world” defined by our class of models and the causal structure of the world<sup>4</sup>.

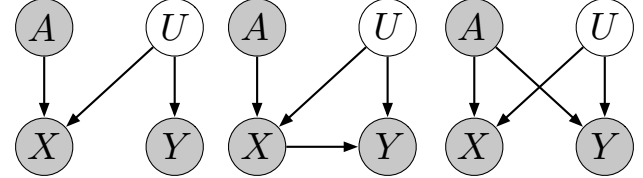


Figure 2. Three causal models for different real-world fair prediction scenarios. See section 4 for discussion.

#### 4.2. Examples

To give an intuition for counterfactual fairness we will consider three fair prediction scenarios: **insurance pricing**; **crime prediction**; **college admissions**. Each of these correspond to one of the three causal graphs in Figure 2.

**Scenario 1: The Red Car.** Imagine a car insurance company wishes to price insurance for car owners by predicting their accident rate  $Y$ . They assume there is an unobserved factor corresponding to aggressive driving  $U$ , that (a) causes drivers to be more likely have an accident, and (b) causes individuals to prefer red cars (the observed variable  $X$ ). Moreover, individuals belonging to a certain race  $A$  are more likely to drive red cars. However, these individuals are no more likely to be aggressive or to get in accidents than any one else. We show this in Figure 2 (Left).

Thus, using the red car feature  $X$  to predict accident likelihood  $Y$  would seem to be an unfair prediction because it may charge individuals of a certain race more than others, even though no race is more likely to have an accident. Counterfactual fairness agrees with this notion.

**Lemma 1.** Consider the structure in Figure 2 (left). There exist model classes and loss functions where fitting a predictor to  $X$  only is not counterfactually fair, while the same algorithm will give a fair predictor using both  $A$  and  $X$ .

*Proof.* As in the definition, we will consider the popula-

<sup>4</sup>The notion of “closest world” is pervasive in the literature of counterfactual inference under different meanings (Halpern, 2016; Pearl, 2000). Here, the cost function used to map fair variables to unfair outcomes also plays a role, but this concerns a problem dependent utility function that would be present anyway in the unfair prediction problem, and is orthogonal to the causal assumptions.



tion case, where the joint distribution is known. Consider the case where the equations described by the model in Figure 2 (Left) are deterministic and linear:

$$X = \alpha A + \beta U, \quad Y = \gamma U$$

and the variance of  $U$  is  $v_U$ , the variance of  $A$  is  $v_A$ , and we assume all coefficients are non-zero. The predictor  $\hat{Y}(X)$  defined by least-squares regression of  $Y$  on *only*  $X$  is given by  $\hat{Y}(X) \equiv \lambda X$ , where  $\lambda = \text{Cov}(X, Y) / \text{Var}(X) = \beta\gamma v_U / (\alpha^2 v_A + \beta^2 v_U) \neq 0$ .

We can test whether a predictor  $\hat{Y}$  is counterfactually fair using the procedure described in Section 3: (i) Compute  $U$  given observations of  $X, Y, A$ ; (ii) Substitute the equations involving  $A$  with an interventional value  $a'$ ; (iii) Compute the variables  $X, Y$  with the interventional value  $a'$ . It is clear here that  $\hat{Y}_a(U) = \lambda(\alpha a + \beta U) \neq \hat{Y}_{a'}(U)$ . This predictor is not counterfactually fair. Thus, in this case fairness through unawareness actually perpetuates unfairness.

Consider instead doing least-squares regression of  $Y$  on  $X$  and  $A$ . Note that  $\hat{Y}(X, A) \equiv \lambda_X X + \lambda_A A$  where  $\lambda_X, \lambda_A$  can be derived as follows:

$$\begin{aligned} \begin{pmatrix} \lambda_X \\ \lambda_A \end{pmatrix} &= \begin{pmatrix} \text{Var}(X) & \text{Cov}(A, X) \\ \text{Cov}(X, A) & \text{Var}(A) \end{pmatrix}^{-1} \begin{pmatrix} \text{Cov}(X, Y) \\ \text{Cov}(A, Y) \end{pmatrix} \\ &= \frac{1}{\beta^2 v_U v_A} \begin{pmatrix} v_A & -\alpha v_A \\ -\alpha v_A & \alpha^2 v_A + \beta^2 v_U \end{pmatrix} \begin{pmatrix} \beta\gamma v_U \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\gamma}{\beta} \\ \frac{-\alpha\gamma}{\beta} \end{pmatrix} \end{aligned} \quad (2)$$

Now imagine we have observed  $A = a$ . This implies that  $X = \alpha a + \beta U$  and our predictor is  $\hat{Y}(X, a) = \frac{\gamma}{\beta}(\alpha a + \beta U) + \frac{-\alpha\gamma}{\beta}a = \gamma U$ . Thus, if we substitute  $a$  with a counterfactual  $a'$  (the action step described in Section 3) the predictor  $\hat{Y}(X, A)$  is unchanged! This is because our predictor is constructed in such a way that any change in  $X$  caused by a change in  $A$  is cancelled out by the  $\lambda_A$ . Thus this predictor is counterfactually fair.  $\square$

Note that if Figure 2 (Left) is the true model for the real world then  $\hat{Y}(X, A)$  will also satisfy demographic parity and equality of opportunity as  $Y$  will be unaffected by  $A$ .

The above lemma holds in a more general case for the structure given in Figure 2 (Left): any non-constant estimator that depends only on  $X$  is not counterfactually fair as changing  $A$  always alters  $X$ . We also point out that the method used in the proof is a special case of a general method to building a predictor based on information deduced about  $U$  that will be described in the next section. We note that, outside of this particular causal model in Figure 2 (Left), the predictor  $\hat{Y}(X, A)$  is not counterfactually fair, as described in the following scenarios.

**Scenario 2: High Crime Regions.** A local police precinct wants to know how likely a given house is to be broken into,  $Y$ . This likelihood depends on many unobserved factors ( $U$ ) but also upon the neighborhood the house lies in ( $X$ ). However, different ethnic groups are more likely to live in particular neighborhoods, and so neighborhood and break-in rates are often correlated with the race  $A$  of the house occupier. This can be seen in Figure 2 (Center). Unlike the previous case, a predictor  $\hat{Y}$  trained using  $X$  and  $A$  is not counterfactually fair. The only change from Scenario 1 is that now  $Y$  depends on  $X$  as follows:  $Y = \gamma U + \theta X$ . Now if we solve for  $\lambda_X, \lambda_A$  it can be shown that  $\hat{Y}(X, a) = (\gamma - \frac{\alpha^2 \theta v_A}{\beta v_U})U + \alpha \theta a$ . As this predictor depends on the values of  $A$ ,  $\hat{Y}(X, a) \neq \hat{Y}(X, a')$  and thus  $\hat{Y}(X, A)$  is not counterfactually fair.

**Scenario 3: University Success.** A university wants to know if students will be successful post-graduation  $Y$ . They have information such as: grade point average (GPA), advanced placement (AP) exams results, and other academic features  $X$ . The university believes however, that an individual's gender  $A$  may influence these features and their post-graduation success  $Y$  due to social discrimination. They also believe that independently, an individual's latent talent  $U$  causes  $X$  and  $Y$ . We show this in Figure 2 (Right). We can again ask, is the predictor  $\hat{Y}(X, A)$  counterfactually fair? In this case, the difference between this and Scenario 1 is that  $Y$  is a function of  $U$  and  $A$  as follows:  $Y = \gamma U + \eta A$ . We can again solve for  $\lambda_X, \lambda_A$  and show that  $\hat{Y}(X, a) = (\gamma - \frac{\alpha \eta v_A}{\beta v_U})U + \eta a$ . Again  $\hat{Y}(X, A)$  is a function of  $A$  so it cannot be counterfactually fair.

## 5. Methods and Assessment

Given that the unaware and full information models are not counterfactually fair, how can we design predictors that are? In general given a causal model, a counterfactually fair classifier  $\hat{Y}$  is one that is a function of *any*  $U$  and *any* variables  $X$  which are not descendants of  $A$ . As defined, these variables are independent of  $A$  and thus any change in  $A$  cannot change  $\hat{Y}$ . In this section we describe techniques for constructing latent variables  $U$  and a predictor  $\hat{Y}$ .

Before delving into details, we point out two important observations. First, if a strict subset of  $U$  is used, the causal model need not be fully specified: equation  $V_i = f_i(p_{a_i}, U_{p_{a_i}})$  can be substituted by a conditional probability  $p(V_i | p_{a_i}, U'_{p_{a_i}})$ , where  $U'_{p_{a_i}} \subset U_{p_{a_i}}$  and  $p(V_i | p_{a_i}, U'_{p_{a_i}}) = \int f_i(p_{a_i}, U_{p_{a_i}}) dU''_{p_{a_i}}$ , where  $U''_{p_{a_i}} \equiv U_{p_{a_i}} \setminus U'_{p_{a_i}}$ . This marginalization has implications in modeling discussed in the next section.

Second, any random variable generated independently is trivially counterfactually fair. However, we desire that

$\hat{Y}$  is a *good* predictor, not simply a coin toss. That is,  $\hat{Y}$  is typically a parameterized function  $g_\theta(U, X)$  where  $\theta$  is learned by minimizing the empiric expected loss  $E[l(Y, g_\theta(U, X)) | X, A]$ . For instance,  $l(Y, g_\theta(U, X)) = (Y - g_\theta(U, X))^2$ , or the log-loss for Bernoulli classification. In practice, the distribution of  $A \cup X \cup \{Y\}$  can be the empirical distribution as given by some training data, while  $p(U | X, A)$  comes from the estimated causal model fit to the same training data. Any predictor can be used to learn  $g_\theta(U, X)$  including random forests and neural networks.

### 5.1. Limitations and a Guide to Model Building

Causal modeling requires untestable assumptions. Experimental data can sometimes be used to infer causal connections, but counterfactual modeling requires functional decompositions between background and endogenous variables. Such decompositions are not uniquely identifiable with experimental data. As in several matters of law and regulation, fairness at an individual level is a counterfactual quantity and some level of assumptions are unavoidable. As a guide for building fair predictive models, we categorize assumptions by three levels of increasing strength.

- Level 1 Given a causal DAG, build  $\hat{Y}$  using as covariates only the observable variables not descendants of the protected attributes  $A$ . This requires information about the DAG, but no assumptions about structural equations or priors over background variables.
- Level 2 Level 1 ignores much information, particularly if the protected attributes are typical attributes such as race or sex, which are parents of many other variables. To include information from descendants of  $A$ , we postulate background latent variables that act as causes of observable variables, based on explicit domain knowledge and learning algorithms<sup>5</sup>. Information from  $X$  will propagate to the latent variables by conditioning.
- Level 3 In Level 2, the model factorizes as a general DAG, and each node follows a non-degenerate distribution given observed and latent variables. In this level, we remove all randomness from the conditional distributions obtaining a full decomposition  $(U, V, F)$  of the model. For instance, the distribution  $p(V_i | V_1, \dots, V_{i-1})$  can be treated as an additive error model,  $V_i = f_i(V_1, \dots, V_{i-1}) + e_i$  (Peters et al., 2014). The error term  $e_i$  then becomes an input to  $\hat{Y}$  after conditioning on the observed variables. This maximizes the information extracted by the fair predictor  $\hat{Y}$ .

<sup>5</sup>In some domains, it is actually common to build a model entirely around latent constructs with few or no observable parents nor connections among observed variables (Bollen, 1989).

### 5.2. Special cases

Consider the graph  $A \rightarrow X \rightarrow Y$ . In general, if  $\hat{Y}$  is a function of  $X$  only, then  $\hat{Y}$  need not obey demographic parity, i.e.

$$P(\hat{Y} | A = a) \neq P(\hat{Y} | A = a').$$

If we postulate a structural equation  $X = \alpha A + e_X$ , then given  $A$  and  $X$  we can deduce  $e_X$ . If  $\hat{Y}$  is a function of  $e_X$  only and, by assumption,  $e_X$  is independent of  $A$ , then the assumptions imply that  $\hat{Y}$  will satisfy demographic parity, and that can be falsified. By way of contrast, if  $e_X$  is not uniquely identifiable from the structural equation and  $(A, X)$ , then the distribution of  $\hat{Y}$  depends on the value of  $A$  as we marginalize  $e_X$ , and demographic parity will not follow. This leads to the following:

**Lemma 2.** *If all background variables  $U' \subseteq U$  in the definition of  $\hat{Y}$  are determined from  $A$  and evidence  $W$ , and all observable variables in the definition of  $\hat{Y}$  are independent of  $A$  given  $U'$ , then  $\hat{Y}$  satisfies demographic parity.*

Thus, counterfactual fairness can be thought of as a counterfactual analog of demographic parity. We advocate that counterfactual assumptions should underlie all approaches that separate the sources of variation of the data into “fair” and “unfair” components. As an example, Louizos et al. (2015) explains the variability in  $X$  from  $A$  and an independent source  $U$  following the DAG  $A \rightarrow X \leftarrow U$ . As  $U$  and  $A$  are not independent given  $X$  in this representation, a type of “posterior regularization” (Ganchev et al., 2010) is enforced such that a posterior  $p_{fair}(U | A, X)$  is close to the model posterior  $p(U | A, X)$  while satisfying  $p_{fair}(U | A = a, X) \approx p_{fair}(U | A = a', X)$ . But this is neither necessary nor sufficient for counterfactual fairness if the model for  $X$  given  $A$  and  $U$  is not justified by a causal mechanism. If it is,  $p(U | A, X)$  is justified as distribution which we can use to marginalize  $U$  in  $p(\hat{Y}(U) | A, X)$ , without requiring regularization. Methods which estimate the relationship between  $A$ ,  $U$  and  $X$  based on penalizing dependence measures between an estimated  $U$  and  $A$  are relevant in estimating a causal model (e.g. Mooij et al. (2009)), but these are motivated by  $U$  being is deterministically inferred from  $A$  and  $X$  by construction. It is unclear in Louizos et al. (2015) how the ideal label  $Y$  is causally connected to  $U$  and  $A$ , and the semantics of the “unfair” components of  $Y$  are not detailed.

## 6. Experiments

We test our approach on two practical problems that require fairness, the first is *prediction of success in law school* and the second is *separating actual and perceived criminality in police stops*. For each problem we construct causal models, and make explicit how unfairness may affect observed and unobserved variables in the world. Given these

models we derive counterfactually fair predictors, and predict latent variables such as a person’s ‘criminality’ (which may be useful for predicting crime) as well as their ‘perceived criminality’ (which may be due to prejudices based on race and sex). We analyze empirically how counterfactually fair the unaware and full predictors are, assuming knowledge of the correct causal model, and compare the prediction accuracies of all models. Finally we judge how well our counterfactually fair ‘criminality’ score satisfies demographic parity.

### 6.1. Law school success

The Law School Admission Council conducted a survey across 163 law schools in the United States (Wightman, 1998). It contains information on 21,790 law students such as their entrance exam scores (LSAT), their grade-point average (GPA) collected prior to law school, and their first year average grade (FYA).

Given this data, a school may wish to predict if an applicant will have a high FYA. The school would also like to make sure these predictions are not biased by an individual’s race and sex. However, the LSAT, GPA, and FYA scores, may be biased due to social factors. We compare our framework with two unfair baselines: 1. **Full**: the standard technique of using all features, including sensitive features such as race and sex to make predictions; 2. **Unaware**: fairness through unawareness, where we do not use race and sex as features. For comparison, we generate predictors  $\hat{Y}$  for all models using logistic regression.

**Fair prediction.** As described in Section 5.1, there are three ways in which we can model a counterfactually fair predictor of FYA. Level 1 uses any features which are not descendants of race and sex for prediction. Level 2 models latent ‘fair’ variables which are parents of observed variables. These variables are independent of both race and sex. Level 3 models the data using an additive error model, and uses the independent error terms to make predictions. These models make increasingly strong assumptions corresponding to increased predictive power. We split the dataset 80/20 into a train/test set, preserving label balance, to evaluate the models.

As we believe LSAT, GPA, and FYA are all biased by race and sex, we cannot use any observed features to construct a counterfactually fair predictor as described in Level 1.

In Level 2, we postulate that a latent variable: a student’s **knowledge** (K), affects GPA, LSAT, and FYA scores. The causal graph corresponding to this model is shown in Fig-

Table 1. Prediction results using logistic regression. Note that we must sacrifice a small amount of accuracy to ensuring counterfactually fair prediction (Fair K, Fair Add), versus the models that use unfair features: GPA, LSAT, race, sex (Full, Unaware).

	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918

ure 3, (**Level 2**). This is a short-hand for the distributions:

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G)$$

$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S))$$

$$\text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1)$$

$$K \sim \mathcal{N}(0, 1)$$

We perform inference on this model using an observed training set to estimate the posterior distribution of  $K$ . We use the probabilistic programming language Stan (Stan Development Team, 2016) to learn  $K$ . We call the predictor constructed using  $K$ , **Fair K**.

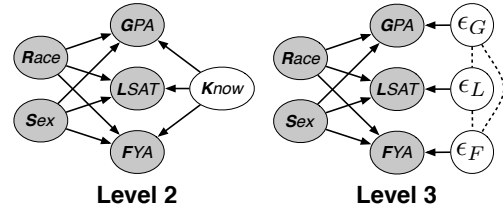


Figure 3. A causal model for the problem of predicting law school success fairly.

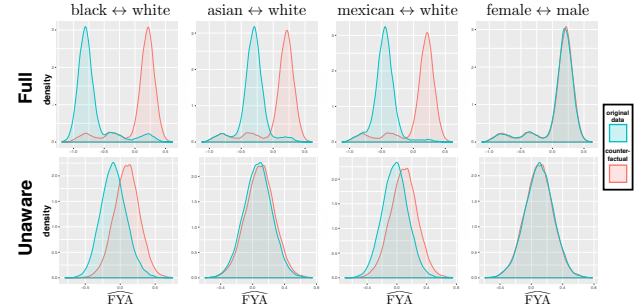


Figure 4. Density plots of predicted  $\text{FYA}_a$  and  $\text{FYA}_{a'}$ .

In Level 3, we model GPA, LSAT, and FYA as continuous variables with additive error terms independent of race and sex (that may in turn be correlated with one-another). This model is shown in Figure 3, (**Level 3**), and is expressed by:

$$\text{GPA} = b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G)$$

$$\text{LSAT} = b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L)$$

$$\text{FYA} = b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)$$

We estimate the error terms  $\epsilon_G, \epsilon_L$  by first fitting two models that each use race and sex to individually predict GPA

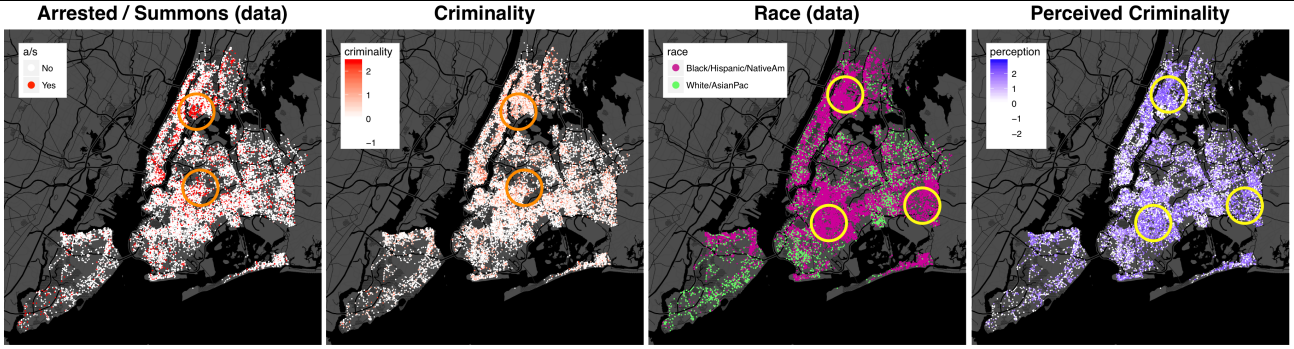


Figure 5. Understanding criminality. The above maps show the decomposition of stop and search data in New York into factors based on perceived criminality (a race dependent variable) and latent criminality (a race neutral measure). See section 6.2.

and LSAT. We then compute the residuals of each model (e.g.,  $\epsilon_G = \text{GPA} - \hat{Y}_{\text{GPA}}(R, S)$ ). We use these residual estimates of  $\epsilon_G, \epsilon_L$  to predict FYA. We call this *Fair Add*.

**Accuracy.** We compare the RMSE achieved by logistic regression for each of the models on the test set in Table 1. The **Full** model achieves the lowest RMSE as it uses race and sex to more accurately reconstruct FYA. Note that in this case, this model is not fair even if the data was generated by one of the models shown in Figure 3 as it corresponds to Scenario 3. The (also unfair) **Unaware** model still uses the unfair variables GPA and LSAT, but because it does not use race and sex it cannot match the RMSE of the **Full** model. As our models satisfy counterfactual fairness, they trade off some accuracy. Our first model **Fair K** uses weaker assumptions and thus the RMSE is highest. Using the Level 3 assumptions, as in **Fair Add** we produce a counterfactually fair model that trades lower RMSE for slightly weaker assumptions.

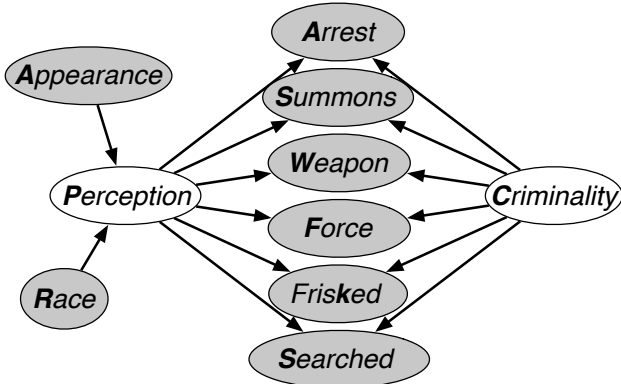


Figure 6. A causal model for the stop and frisk dataset.

**Counterfactual fairness.** We would like to empirically test whether the baseline methods are counterfactually fair. To do so we will assume the true model of the world is given by Figure 3, (**Level 2**). We can fit the parameters of this model using the observed data and evaluate counterfactual fairness by sampling from it. Specifically, we will generate samples from the model given either the observed race and sex, or *counterfactual* race and sex variables. We

will fit models to both the original and counterfactual sampled data and plot how the distribution of predicted FYA changes for both baseline models. Figure 6.1 shows this, where each row corresponds to a baseline predictor and each column corresponds to the counterfactual change. In each plot, the blue distribution is density of predicted FYA for the original data and the red distribution is this density for the counterfactual data. If a model is counterfactually fair we would expect these distributions to lie exactly on top of each other. Instead, we note that the **Full** model exhibits counterfactual unfairness for all counterfactuals except sex. We see a similar trend for the **Unaware** model, although it is closer to being counterfactually fair. To see why these models seem to be fair w.r.t. to sex we can look at weights of the DAG which generates the counterfactual data. Specifically the DAG weights from (male,female) to GPA are (0.93,1.06) and from (male,female) to LSAT are (1.1,1.1). Thus, these models are fair w.r.t. to sex simply because of a very weak causal link between sex and GPA/LSAT.

## 6.2. True vs. Perceived Criminality

Since 2002, the New York Police Department (NYPD) has recorded information about every time a police officer has stopped someone. The officer records information such as if the person was searched or frisked, their appearance, etc. We consider the data collected on males stopped during 2014 which constitutes 38,609 records.

**Model.** We model this stop-and-frisk data using the graph in Figure 6. Specifically, we posit main causes for the observations: *Arrest* (if an individual was arrested), *Summons* (an individual was called to a court-summons), *Weapon* (an individual was found to be carrying a weapon), *Force* (some sort of force was used during the stop), *Frisked*, and *Searched*. The first cause of these observations is some measure of an individual’s latent *Criminality*, which we do not observe. We believe there is an additional cause, an individual’s perceived criminality, *Perception*, also unobserved. This second factor is introduced as we believe that these observations may be biased based on



an officer’s perception of whether an individual is likely a criminal or not. This perception is affected by an individual’s *Appearance* and their *Race*. In this sense *Criminality* is counterfactually fair, while *Perception* models how race affects each of the other observed variables.

**Criminality and perception distributions.** After fitting this model to the data we can look at the distribution of *Criminality* and *Perception* across different races, shown as box plots in Figure 6. We see that the median criminality for each race is nearly identical, while the distributions are somewhat different, demonstrating that *Criminality* approaches demographic parity. The differences that do exist may be due to unobserved confounding variables that are affected by race or unmodeled noise in the data. On the right *Perception* varies considerably by race with white individuals having the lowest perceived criminality while black and black Hispanic individuals have the highest.

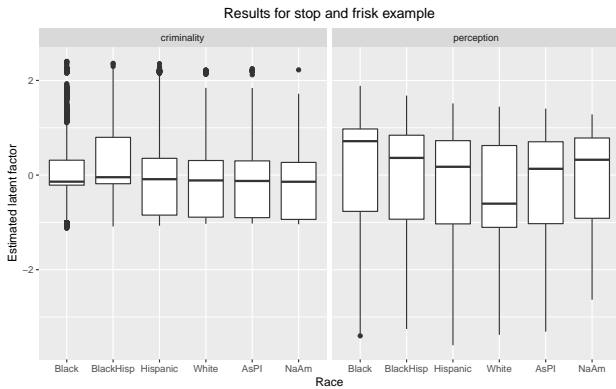


Figure 7. Distributions of estimated latent perception and criminality scores for the stop and frisk dataset.

**Visualization on a map of New York City.** Each of the stops can be mapped to longitude and latitude points for where the stop occurred<sup>6</sup>. Thus we can visualize *Criminality* and *Perception* alongside *Race* and the combination of *Arrest* and *Summons*, shown in Figure 5. Criminality seems to be a continuous approximation of arrest and summons as both plots show red in similar areas. However, the plots show that certain areas, while having a lot of arrests have low criminality scores such as south Bronx and west Queens (circled in orange). We can also compare the perceived criminality with a plot of race, where we have divided the races into Group A: black, black Hispanic, Hispanic, and Native American (shown in purple); and Group B: white and Asian/Pacific Islander (shown in green). Group A are all races that have positive weights on the connection from *Race* to *Perception* in the fitted model, while Group B all have negative weights. Thus being in Group A leads one to have a higher perceived criminality than being in Group B. This can be seen in the right-

most plot of Figure 5. Certain areas of town such as central Brooklyn, central Bronx, and southern Queens have very high criminality and almost all stops are by members of Group A (circled in yellow).

## 7. Conclusion

We have presented a new model of fairness we refer to as *counterfactual fairness*. It allows us to propose fair algorithms that, rather than simply ignoring protected attributes, are able to take into account the different social biases that may arise towards individuals of a particular race, gender, or sexuality and compensate for these biases effectively. We experimentally contrasted our approach with previous unfair approaches and show that our explicit causal models capture these social biases and make clear the implicit trade-off between prediction accuracy and fairness in an unfair world. We propose that fairness should be regulated by explicitly modeling the causal structure of the world. Criteria based purely on probabilistic independence cannot satisfy this and are unable to address *how* unfairness is occurring in the task at hand. By providing such causal tools for addressing fairness questions we hope we can provide practitioners with customized techniques for solving a wide array of fair modeling problems.

## References

- Bollen, K. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.
- Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, and Kalai, Adam T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.
- Brennan, Tim, Dieterich, William, and Ehret, Beate. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36 (1):21–40, 2009.
- Calders, Toon and Verwer, Sicco. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
- DeDeo, Simon. Wrong side of the tracks: Big data and protected categories. *arXiv preprint arXiv:1412.4643*, 2014.
- Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012.

<sup>6</sup><https://github.com/stablemarkets/StopAndFrisk>

- Ganchev, K., Graca, J., Gillenwater, J., and Taskar, B. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.
- Grgic-Hlaca, Nina, Zafar, Muhammad Bilal, Gummadi, Krishna P, and Weller, Adrian. The case for process fairness in learning: Feature selection for fair decision making. *NIPS Symposium on Machine Learning and the Law*, 2016.
- Halpern, J. *Actual Causality*. MIT Press, 2016.
- Hardt, Moritz, Price, Eric, Srebro, Nati, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.
- Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016.
- Kamiran, Faisal and Calders, Toon. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pp. 1–6. IEEE, 2009.
- Kamiran, Faisal and Calders, Toon. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.
- Kamishima, Toshihiro, Akaho, Shotaro, and Sakuma, Jun. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 643–650. IEEE, 2011.
- Khandani, Amir E, Kim, Adlar J, and Lo, Andrew W. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.
- Kleinberg, Jon, Mullainathan, Sendhil, and Raghavan, Manish. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- Louizos, Christos, Swersky, Kevin, Li, Yujia, Welling, Max, and Zemel, Richard. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.
- Mahoney, John F and Mohen, James M. Method and system for loan origination and underwriting, October 23 2007. US Patent 7,287,008.
- Mooij, J., Janzing, D., Peters, J., and Scholkopf, B. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 745–752, 2009.
- Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.
- Pearl, J., Glymour, M., and Jewell, N. *Causal Inference in Statistics: a Primer*. Wiley, 2016.
- Pearl, Judea et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.
- Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014. URL <http://jmlr.org/papers/v15/peters14a.html>.
- Stan Development Team. Rstan: the r interface to stan, 2016. R package version 2.14.1.
- Wightman, Linda F. Isac national longitudinal bar passage study. Isac research report series. 1998.
- Zafar, Muhammad Bilal, Valera, Isabel, Rodriguez, Manuel Gomez, and Gummadi, Krishna P. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2015.
- Zafar, Muhammad Bilal, Valera, Isabel, Rodriguez, Manuel Gomez, and Gummadi, Krishna P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *arXiv preprint arXiv:1610.08452*, 2016.
- Zemel, Richard S, Wu, Yu, Swersky, Kevin, Pitassi, Toniann, and Dwork, Cynthia. Learning fair representations. *ICML (3)*, 28:325–333, 2013.
- Zliobaite, Indre. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.