Counterfactual Fairness

Anonymous Author(s)
Affiliation
Address
email

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) in 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 ao real-world problem of fair prediction of law school success.

1 Contribution

Machine learning has spread to fields as diverse as credit scoring [17], crime prediction [5], and loan assessment [21]. 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". [1]

As a result, there has been immense interest in designing algorithms that make fair predictions [4, 6, 8, 9, 11, 13–16, 18, 20, 29–32]. 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 *protected attribute* and the data, certain definitions of fairness can actually *increase discrimination*. In perhaps the most closely related work, Johnson et al. [12] give similar arguments but from a non-causal perspective.

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

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

In this paper, we introduce the first explicitly formal causal model approach for casting problems of fair predictions with explicit counterfactual assumptions. Specifically, we leverage the causal framework of Pearl et al. [25] to model the relationship between protected attributes and data. We describe how techniques from causal inference can be effective tools for designing fair algorithms and argue, as in DeDeo [7], that it is essential to properly address causality.

In Section 2, we provide a summary of basic concepts in fairness and causal modeling. In Section 3, we provide the formal definition of *counterfactual fairness*, which enforces that a distribution over possible predictions for an individual should remain unchanged, in a world where an individual's protected attributes had been different in a causal sense. In Section 4, we describe an algorithm to implement this definition, while distinguishing it from existing approaches. In Section 5, we illustrate the algorithm with a case of fair assessment of law school success.

2 Background

This section provides a basic account of two separate areas of research in machine learning, which are formally unified in this paper. We suggest Berk et al. [1] and Pearl et al. [24] as references for further reading.

Throughout this paper, we will use the following notation. Let A denote the set of *protected attributes* of an individual, variables that must not be discriminated against in a formal sense defined differently by each notion of fairness discussed. The decision of whether an attribute is protected or not is taken as a primitive in any given problem, regardless of the definition of fairness adopted. Moreover, let X denote the other observable attributes of any particular individual, U the set of latent attributes, and let Y denote the outcome to be predicted, which itself might be contaminated with historical biases. Finally, $\hat{Y}$ is the *predictor*, a random variable that depends on A, X and U, and which is produced by a machine learning algorithm as a prediction of Y.

2.1 Fairness

The goal of fairness in machine learning is to design automated algorithms that make fair predictions across various demographic groups. This unfairness can arise in several ways. Consider for instance historically biased distributions, where individuals with different protected attributes A may have different attributes X such as their current level of wealth, which is then used for credit scoring. Disparities may be due to discriminative measures in hiring practices and perpetuated by reduced financial support for the education of minorities due to bad credit ratings.

There has been a wealth of recent work towards fair algorithms. These include fairness through unawareness [9], individual fairness [8, 13, 20, 31], demographic parity/disparate impact [29], and equality of opportunity [11, 30].

Definition 1 (Fairness Through Unawareness (FTU))**.** *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

Any mapping $\hat{Y} : X \to Y$ that excludes A satisfies this. Initially proposed as a baseline, the approach has found favor recently with more general approaches such as Grgic-Hlaca et al. [9]. Despite its compelling simplicity, FTU has a clear shortcoming as elements of X can contain discriminatory information analogous to A that may not be obvious at first. The need for expert knowledge in assessing the relationship between A and X was highlighted in the work on individual fairness:

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

As described in [8], the notion of similarity must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling. This can also be contrasted again population level criteria such as

Definition 3 (Demographic Parity (DP))**.** *A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$.*

Definition 4 (Equality of Opportunity (EO))**.** *A predictor $\hat{Y}$ satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.*

These criteria can be incompatible in general, as discussed by Kleinberg et al. [18] and Berk et al. [1]. Following the motivation of IF, we propose that knowledge about individual differences should be taken into consideration, even if strong assumptions are necessary. Moreover, it is not immediately clear for any of these approaches in which ways historical biases can be tackled. We approach such issues from an explicit causal modeling perspective.

2.2 Causal Models and Counterfactuals

We follow Pearl [23], and define a causal model as a triple (U, V, F) of sets such that

- U is a set of latent **background** variables,which are factors not caused by any variable in the set V of **observable** variables;
- F is a set of functions $\{f_1, \ldots, f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \backslash \{V_i\}$ and $U_{pa_i} \subseteq U$. Such equations are also known as **structural equations** [2].

The notation "pa_i" refers to the "parents" of V_i and is motivated by the assumption that the model factorizes as a directed graph, here assumed to be a directed acyclic graph (DAG). The model is causal in that, given a distribution $P(U)$ over the background variables U, we can derive the distribution of a subset $Z \subseteq V$ following an **intervention** on $V \setminus Z$. An intervention on variable V_i is the substitution of equation $V_i = f_i(pa_i, U_{pa_i})$ with the equation $V_i = v$ for some v. This captures the idea of an agent, external to the system, modifying it by forcefully assigning value v to V_i.

The specification of F is a strong assumption but 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 observable variables Z and Y. By assumption, the state of any observable 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 $Y_{Z \leftarrow z}(u)$ [23], 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. [24]: 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 $P(U \mid W = w)$.

3 Counterfactual Fairness

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

Definition 5 (Counterfactual fairness)**.** *Predictor $\hat{Y}$ is* **counterfactually fair** *if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

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

The appeal of counterfactual fairness is its close relationship to **actual causes** [10], or token causality: other things been equal, A should not be a cause of $\hat{Y}$ in any individual instance. In other words, changing A while holding everything else the same will not change the distribution of $\hat{Y}$[2]. We also emphasize that counterfactual fairness is an individual-level definition. This is substantially different from comparing different individuals that happen to share the same "treatment" $A = a$ and coincide on the values of X, as discussed in Section 4.3.1 of [24] and the Supplementary Material[3].

[2]Notice that we always assume counterfactuals to be well-defined by the model. For instance, "race" can be taken as a surrogate for "race perception".

[3]We strongly suggest that reviewers look at it.

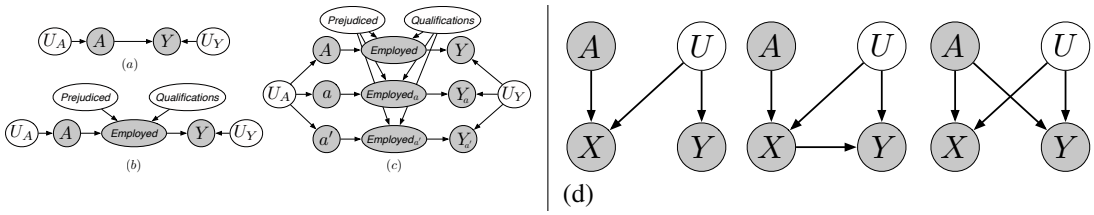

Figure 1: **Left:** (a) The graph corresponding to a causal model with A being the protected attribute and Y some outcome of interest, with background variables assumed to be independent. (b) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (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 [23] under two different counterfactual levels for A. This is created by copying nodes descending from A, which inherit unaffected parents from the factual world. (d) Two causal models for different real-world fair prediction scenarios. See Section 3 for discussion.

Differences between X_a and $X_{a'}$ must be caused by variations on A only. Notice also that this definition is agnostic with respect to how good a predictor $\hat{Y}$ is, which we discuss in Section 4.

Relation to individual fairness. IF is agnostic with respect to its notion of similarity metric, which is both a strength (generality) and a weakness (no unified way of defining similarity). Counterfactuals and similarities are related, as in the classical notion of distances between "worlds" corresponding to different counterfactuals [19]. Defining $\hat{Y}$ as a deterministic function of $W \subset A \cup X \cup U$, as in several of our examples to follow, then IF can be defined by treating equally two individuals with the same W in a way that is also counterfactually fair.

Relation to Pearl et al. [24]. In Example 4.4.4 of [24], the authors condition instead on X, A, and the observed realization of $\hat{Y}$, and calculate the probability of the counterfactual realization $\hat{Y}_{A \leftarrow a'}$ differing from the factual. This example conflates the predictor $\hat{Y}$ with the outcome Y, of which we remain agnostic in our definition but which is used in the construction of $\hat{Y}$ as in Section 4. Our framing makes the connection to machine learning more explicit.

3.1 Implications

One simple but important implication of the definition of counterfactual fairness is the following:

Lemma 1. *Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then $\hat{Y}$ will be counterfactual fair if it is a function of the non-descendants of A.*

Proof. Let W be any non-descendant of A in $\mathcal{G}$. Then $W_{A \leftarrow a}(U)$ and $W_{A \leftarrow a'}(U)$ have the same distribution by the three inferential steps in Section 2.2. Hence, the distribution of any function $\hat{Y}$ of the non-descendants of A is invariant with respect to the counterfactual values of A. □

This does not exclude using a descendant W of A as a possible input to $\hat{Y}$. However, this will only be possible in the case where the overall functional dependence of W on A disappears, which will not happen in general. The practical implication is that $\hat{Y}$ can use all and only the non-descendants of A.

Ancestral closure of protected attributes. Suppose that a parent of a member of A is not in A. Counterfactual fairness allows for the use of it in the definition of $\hat{Y}$. If this seems counterintuitive, then we argue that the fault should be at the postulated set of protected attributes rather than with the definition of counterfactual fairness, and that typically we should expect set A to be closed under ancestral relationships given by the causal graph. For instance, if *Race* is a protected attribute, and *Father's race* is a parent of *Race*, then it should also be in A.

Dealing with historical biases. The explicit difference between $\hat{Y}$ and Y allows us to tackle historical biases. For instance, let Y 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 $A \rightarrow Y$, shown in Figure 1(a) with the explicit inclusion of set U of independent background variables. Y is the objectively ideal measure

for decision making, the binary indicator of the event that the individual defaults on a loan. If A is postulated to be a protected attribute, then the predictor $\hat{Y} = Y = f_Y(A, U)$ is not counterfactually fair, with the arrow $A \to Y$ being (for instance) the result of a world that punishes individuals in a way that is out of their control. Figure 1(b) shows a finer-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[4]. The counterfactual fairness principle however forbids us from using Y: using the twin network of Pearl [23], we see in Figure 1(c) that Y_a and $Y_{a'}$ need not be identically distributed given the background variables.

In contrast, any function of variables not descendants of A can be used 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, as proposed in Section 4.1. We are essentially learning a projection of Y into the space of fair decisions, removing historical biases as a by-product.

3.2 Further Examples

To give further intuition for counterfactual fairness, we will consider two real-world fair prediction scenarios: **insurance pricing** and **crime prediction**. Each of these correspond to one of the two causal graphs in Figure 1(d). The Supplementary Material provides a more mathematical discussion of these examples with more detailed insights.

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 1(d) (*Left*). Thus, using the red car feature X to predict accident rate 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, as X cannot be added to $\hat{Y}$ but U can. Interestingly, we can show (Supplementary Material) that in a linear model, regression Y on A and X is equivalent to regressing on U, so off-the-shelf regression here is counterfactually fair. Regressing Y on X alone obeys the FTA criterion but is not counterfactually fair, so conditioning on A *adds* fairness.

Scenario 2: High Crime Regions. A local police precinct wants to know Y, whether a given house is to be broken into in any given day. The probability of $Y = 1$ 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 1(d) (*Right*). The only change from Scenario 1 is that now Y depends on X directly. If all structural equations are linear, then U is a linear function of A and X, and so, indirectly, a counterfactually fair $\hat{Y}$ can be expressed as a linear function of A and X. However, this is different from assuming that $\hat{Y}$ can be *any* linear combination of A and X. As a matter of fact, the solution for the unrestricted least-squares regression of Y on A and X cannot be written as a function of U only, as shown in the Supplementary Material.

The lesson from the last example is that in general we need a multistage procedure in which we derive latent variables U, and based on them we minimize some loss with respect to Y. This is the core of the algorithm discussed next.

4 Implementing Counterfactual Fairness

As discussed in the previous Section, we need to relate $\hat{Y}$ to Y if the predictor is to be useful, and that we will restrict to $\hat{Y}$ to be a (parameterized) function of the non-descendants of A in the causal

[4]For example, if the function determining employment $f_E(A, P, Q) \equiv I_{(Q>0, P=0 \text{ or } A \neq a)}$ 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.

graph. An algorithm is introduced in Section 4.1, followed by a discussion of the assumptions that can be used to express counterfactuals.

4.1 Algorithm

Let $\hat{Y} \equiv g_\theta(U, X_{\not\succ A})$ be a predictor parameterized by θ, such as a logistic regression or a neural network, and where $X_{\not\succ A} \subseteq X$ are non-descendants of A. Given a loss function $l(\cdot, \cdot)$ such as squared loss of log-likelihood, and training data $\mathcal{D} \equiv \{(A^{(i)}, X^{(i)}, Y^{(i)})\}$ for $i = 1, 2, \ldots, n$, we define $L(\theta) \equiv \sum_{i=1}^{n} \mathbb{E}[l(y^{(i)}, g_\theta(U^{(i)}, x^{(i)}_{\not\succ A}) \mid x^{(i)}, a^{(i)}]/n$ as the empirical loss to be minimized with respect to θ. Each expectation is with respect to random variable $U^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$ where $P_{\mathcal{M}}(U \mid x, a)$ is the conditional distribution of the background variables as given by a causal model $\mathcal{M}$ that is available by assumption. If this expectation cannot be calculated analytically, Markov chain Monte Carlo (MCMC) can be used to approximate it, resulting in the following algorithm.

1: **procedure** FAIRLEARNING($\mathcal{D}, \mathcal{M}$) ▷ Learned parameters $\hat{\theta}$
2: For each data point $i \in \mathcal{D}$, sample m MCMC samples $U_1^{(i)}, \ldots, U_m^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
3: Let $\mathcal{D}'$ be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in $\mathcal{D}$ is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}$.
4: $\hat{\theta} \leftarrow \text{argmin}_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(U^{(i')}, x^{(i')}_{\not\succ A}))$.
5: **end procedure**

At prediction time, we report $\tilde{Y} \equiv \mathbb{E}[\hat{Y}(U^\star, x^\star_{\not\succ A}) \mid x^\star, a^\star]$ for a new data point $(a^\star, x^\star)$.

Deconvolution perspective. The algorithm can be understood as a deconvolution approach that, given observables $A \cup X$, extracts its latent sources and pipeline them into a predictive model. We advocate that *counterfactual assumptions must underlie all approaches that claim to extract the sources of variation of the data as "fair" latent components*. As an example, Louizos et al. [20] start from the DAG $A \rightarrow X \leftarrow U$ to extract $P(U \mid X, A)$. As U and A are not independent given X in this representation, a type of penalization is enforced to create a posterior $P_{fair}(U \mid A, X)$ that is close to the model posterior $P(U \mid A, X)$ while satisfying $P_{fair}(U \mid A = a, X) \approx P_{fair}(U \mid A = a', X)$. But *this is neither necessary nor sufficient for counterfactual fairness*. The model for X given A and U must be justified by a causal mechanism, and that being the case, $P(U \mid A, X)$ requires no postprocessing. As a matter of fact, model $\mathcal{M}$ can be learned by penalizing empirical dependence measures between U and A given X (e.g. Mooij et al. [22]), but this concerns $\mathcal{M}$ and not $\hat{Y}$, and is motivated by explicit assumptions about structural equations, as described next.

4.2 Designing the Input Causal Model

Model $\mathcal{M}$ must be provided to algorithm FAIRLEARNING. Causal models always require strong assumptions, but counterfactuals in particular are typically presented in terms of structural equations which are in general unfalsifiable. We point out that we do not need to specify a fully deterministic model, and structural equations can be relaxed as conditional distributions. In particular, the concept of counterfactual fairness holds under three levels of assumptions of increasing strength:

Level 1. Build $\hat{Y}$ using only the observable non-descendants of A. This only requires partial causal ordering and no further causal assumptions, but in many problems there will be few, if any, observables which are not descendants of demographic factors.

Level 2. We postulate background latent variables that act as non-deterministic causes of observable variables, based on explicit domain knowledge and learning algorithms[5]. Information about X is passed to $\hat{Y}$ via $P(U \mid x, a)$.

Level 3. We postulate a fully deterministic model with latent variables. For instance, the distribution $P(V_i \mid V_1, \ldots, V_{i-1})$ can be treated as an additive error model, $V_i = f_i(V_1, \ldots, V_{i-1}) + e_i$ [26]. The error term e_i then becomes an input to $\hat{Y}$ as calculated from the observed variables. This maximizes the information extracted by the fair predictor $\hat{Y}$.

[5] 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 [2].

5 Illustration: Law School Success

We illustrate our approach on a practical problem that requires fairness, the *prediction of success in law school*. A second problem, *separating actual and perceived criminality in police stops*, is described in the Supplementary Material. Following closely the usual framework for assessing causal models in the machine learning literature, the goal of this experiment is to quantify how our algorithm behaves with finite sample sizes assuming ground truth compatible with a synthetic model.

Problem definition: Law school success

The Law School Admission Council conducted a survey across 163 law schools in the United States [28]. 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 4.2, 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 Figure 2, (**Level 2**). This is a short-hand for the distributions:

$$\begin{aligned} \text{GPA} &\sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), & \text{FYA} &\sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1), \\ \text{LSAT} &\sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), & \text{K} &\sim \mathcal{N}(0, 1) \end{aligned}$$

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 [27] to learn K. We call the predictor constructed using K, **Fair** K.

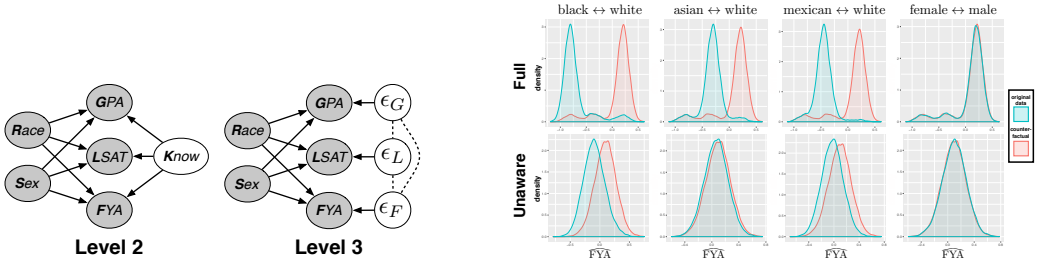

Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

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

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

in Figure 2, (**Level 3**), and is expressed by:

$$\begin{aligned}
\text{GPA} &= b_G + w_G^R R + w_G^S S + \epsilon_G, \;\; \epsilon_G \sim p(\epsilon_G) \\
\text{LSAT} &= b_L + w_L^R R + w_L^S S + \epsilon_L, \;\; \epsilon_L \sim p(\epsilon_L) \\
\text{FYA} &= b_F + w_F^R R + w_F^S S + \epsilon_F, \;\; \epsilon_F \sim p(\epsilon_F)
\end{aligned}$$

We estimate the error terms ϵ_G, ϵ_L by first fitting two models that each use race and sex to individually predict GPA 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 ϵ_G, ϵ_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 2 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.

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 2, (**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 2 shows this, where each row corresponds to a baseline predictor and each column corresponds to the couterfactual 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 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 fairness modeling problems.

References

[1] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207v1*, 2017. 2, 3

[2] Bollen, K. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989. 3, 6

[3] Bollen, K. and (eds.), J. Long. *Testing Structural Equation Models*. SAGE Publications, 1993. 11

[4] 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. 1

[5] 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. 1

[6] Calders, Toon and Verwer, Sicco. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010. 1

[7] DeDeo, Simon. Wrong side of the tracks: Big data and protected categories. *arXiv preprint arXiv:1412.4643*, 2014. 2

[8] 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. 1, 2

[9] 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. 1, 2

[10] Halpern, J. *Actual Causality*. MIT Press, 2016. 3

[11] 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. 1, 2

[12] Johnson, Kory D, Foster, Dean P, and Stine, Robert A. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*, 2016. 1

[13] Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie, Neel, Seth, and Roth, Aaron. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016. 1, 2

[14] 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.

[15] Kamiran, Faisal and Calders, Toon. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[16] 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. 1

[17] 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. 1

[18] Kleinberg, Jon, Mullainathan, Sendhil, and Raghavan, Manish. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. 1, 3

[19] Lewis, D. *Counterfactuals*. Harvard University Press, 1973. 4

[20] Louizos, Christos, Swersky, Kevin, Li, Yujia, Welling, Max, and Zemel, Richard. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015. 1, 2, 6

[21] Mahoney, John F and Mohen, James M. Method and system for loan origination and underwriting, October 23 2007. US Patent 7,287,008. 1

[22] 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. 6

[23] Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. 3, 4, 5

[24] Pearl, J., Glymour, M., and Jewell, N. *Causal Inference in Statistics: a Primer*. Wiley, 2016. 2, 3, 4, 11

[25] Pearl, Judea et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. 2

[26] 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`. 6

[27] Stan Development Team. Rstan: the r interface to stan, 2016. R package version 2.14.1. 7

[28] Wightman, Linda F. Lsac national longitudinal bar passage study. lsac research report series. 1998. 7

[29] Zafar, Muhammad Bilal, Valera, Isabel, Rodriguez, Manuel Gomez, and Gummadi, Krishna P. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2015. 1, 2

[30] 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. 2

[31] Zemel, Richard S, Wu, Yu, Swersky, Kevin, Pitassi, Toniann, and Dwork, Cynthia. Learning fair representations. *ICML (3)*, 28:325–333, 2013. 2

[32] Zliobaite, Indre. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015. 1

S1 Population Level vs Individual Level Causal Effects

As discussed in Section 3, counterfactual fairness is an individual-level definition. This is fundamentally different from comparing different units that happen to share the same "treatment" $A = a$ and coincide on the values of X. To see in detail what this means, consider the following thought experiment.

Let us assess the causal effect of A on $\hat{Y}$ by controlling A at two levels, a and a'. In Pearl's notation, where "$do(A = a)$" expresses an intervention on A at level a, we have that

$$\mathbb{E}[\hat{Y} \mid do(A = a), X = x] - \mathbb{E}[\hat{Y} \mid do(A = a'), X = x], \tag{2}$$

is a measure of causal effect, sometimes called the average causal effect (ACE). It expresses the change that is expected when we intervene on A while observing the attribute set $X = x$, under two levels of treatment. If this effect is non-zero, A is considered to be a cause of $\hat{Y}$.

This raises a subtlety that needs to be addressed: in general, this effect will be non-zero *even if* $\hat{Y}$ *is counterfactually fair*. This may sound counter-intuitive: protected attributes such as race and gender are causes of our counterfactually fair decisions.

In fact, this is not a contradiction, as the ACE in Equation (2) is different from counterfactual effects. The ACE contrasts two independent exchangeable units of the population, and it is a perfectly valid way of performing decision analysis. However, the value of $X = x$ is affected by different background variables corresponding to different individuals. That is, the causal effect (2) contrasts two units that receive different treatments but which happen to coincide on $X = x$. To give a synthetic example, imagine the simple structural equation

$$X = A + U.$$

The ACE quantifies what happens among people with $U = x - a$ against people with $U' = x - a'$. If, for instance, $\hat{Y} = \lambda U$ for $\lambda \neq 0$, then the effect is $\lambda(a - a') \neq 0$.

Contrary to that, the counterfactual difference is zero. That is,

$$\mathbb{E}[\hat{Y}_{A \leftarrow a}(U) \mid A = a, X = x] - \mathbb{E}[\hat{Y}_{A \leftarrow a'}(U) \mid A = a, X = x] = \lambda U - \lambda U = 0.$$

In another perspective, we can interpret the above just as if we had *measured* U from the beginning rather than performing abduction. We then generate $\hat{Y}$ from some $g(U)$, so U is the within-unit cause of $\hat{Y}$ and not A.

If U cannot be deterministically derived from $\{A = a, X = x\}$, the reasoning is similar. By abduction, the distribution of U will typically depend on A, and hence so will $\hat{Y}$ when marginalizing over U. Again, this seems to disagree with the intuition that our predictor should be not be caused by A. However, this once again is a comparison *across individuals*, not within an individual.

It is this balance among (A, X, U) that explains, in the examples of Section **??**, why some predictors are counterfactually fair even though they are functions of the same variables $\{A, X\}$ used by unfair predictors: such functions must correspond to particular ways of balancing the observables that, by way of the causal assumptions, correspond to using non-descendants of A only.

More on conditioning and alternative definitions. As discussed in Example 4.4.4 of Pearl et al. [24], a different proposal for assessing fairness can be defined via the following concept:

Definition 6 (Probability of sufficiency)**.** *We define the probability of event $A = a$ being a* sufficient cause *for our decision $\hat{Y}$, contrasted against $A = a'$, as*

$$P(\hat{Y}_{A \leftarrow a'}(U) \neq y \mid X = x, A = a, \hat{Y} = y). \tag{3}$$

We can then, for instance, claim that $\hat{Y}$ is a fair predictor if this probability is below some pre-specified bound for all (x, a, a'). The shortcomings of this definition come from its original motivation: to *explain* the behavior of an *existing* decision protocol, where $\hat{Y}$ is the current practice and which in a unclear way is conflated with Y. The implication is that if $\hat{Y}$ is to be designed instead of being a natural measure of existing behaviour, then we are using $\hat{Y}$ itself as evidence for the background variables U. This does not make sense if $\hat{Y}$ is yet to be designed by us. If $\hat{Y}$ is to be interpreted as Y, then this does not provide a clear recipe on how to build $\hat{Y}$: while we can use Y to learn a causal model, we cannot use it to collect training data evidence for *U as the outcome Y will not be available to us at prediction time*. For this reason, we claim that while probability of sufficiency is useful as a way of assessing an existing decision making process, and it is not as natural as counterfactual fairness in the context of machine learning.

Approximate fairness and model validation. The notion of probability of sufficiency raises the question on how to define approximate, or high probability, counterfactual fairness. This is an important question but for reasons of conciseness and focus we defer it entirely to future work. Before defining an approximation, it is important to first expose in detail what the exact definition is, which is the goal of this paper.

We also do not address the validation of the causal assumptions used by the input causal model of the FAIRLEARNING algorithm in Section 4.1. The reason is straightforward: this validation is an entirely self-contained step of the implementation of counterfactual fairness. An extensive literature already exists in this topic which the practitioner can refer to (a classic account for instance is [3]), and which can be used as-is in our context.

The experiments performed in Section 5 can be criticized by the fact that they rely on a model that obeys our assumptions, and "obviously" our approach should work better than alternatives. This criticism is not warranted: in machine learning, causal inference is typically assessed through simulations which assume that the true model lies in the family covered by the algorithm. Algorithms, including FAIRLEARNING, are justified in the population sense. How different competitors behave with finite sample sizes is the primary question to be studied in an empirical study of a new concept, where we control for the correctness of the assumptions. Although sensitivity analysis is important, there are many degrees of freedom on how this can be done. Robustness issues are better addressed by extensions focusing on approximate versions of the concept introduced here.

S2 Relation to Demographic Parity

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

$$P(\hat{Y} \mid A = a) \neq P(\hat{Y} \mid A = a'),$$

where, since $\hat{Y}$ is a function of X, the probabilities are obtained by marginalizing over $P(X \mid A = a)$ and $P(X \mid A = a')$, respectively.

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 marginally independent of A, then $\hat{Y}$ is marginally independent of A: this follows the interpretation given in the previous section, where we interpret e_X as "known" despite being mathematically deduced from the observation $(A = a, X = x)$. Therefore, 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 X, 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, as present in the Red Car example further discussed in the next section.

S3 Examples Revisited

In Section **??**, we discussed two examples. We reintroduce them here briefly, add a third example, and explain some consequences of their causal structure to the design of counterfactually fair predictors.

Scenario 1: The Red Car Revisited. In that scenario, the structure $A \to X \leftarrow U \to Y$ implies that $\hat{Y}$ should not use either X or A. On the other hand, it is acceptable to use U. It is interesting to realize, however, that since U is related to A and X, there will be some association between Y and $\{A, X\}$ as discussed in Section S1. In particular, if the structural equation for X is linear, then U is a linear function of A and X, and as such $\hat{Y}$ will also be a function of both A and X. This is not a problem, as it is still the case that the model implies that this is merely a functional dependence that disappears by conditioning on a postulated latent attribute U. Surprisingly, we must make $\hat{Y}$ a indirect function of A if we want a counterfactually fair predictor, as shown in the following Lemma.

Lemma 3. *Consider a linear model with the structure in Figure 1(d). Fitting a linear predictor to X* only *is not counterfactually fair, while the same algorithm will produce a fair predictor using* both *A and X.*

Proof. As in the definition, we will consider the population case, where the joint distribution is known. Consider the case where the equations described by the model in Figure 1(d) (*Left*) are deterministic and linear:

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

Denote the variance of U as v_U, the variance of A as v_A, and 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 $\lambda = Cov(X, Y)/Var(X) = \beta\gamma v_U/(\alpha^2 v_A + \beta^2 v_U) \neq 0$. This predictor follows the concept of fairness through unawareness.

We can test whether a predictor $\hat{Y}$ is counterfactually fair by using the procedure described in Section 2.2: *(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 λ_X, λ_A can be derived as follows:

$$\begin{aligned}\begin{pmatrix}\lambda_X \\ \lambda_A\end{pmatrix} &= \begin{pmatrix}Var(X) & Cov(A,X) \\ Cov(X,A) & Var(A)\end{pmatrix}^{-1}\begin{pmatrix}Cov(X,Y) \\ Cov(A,Y)\end{pmatrix} \\ &= \frac{1}{\beta^2 v_U v_A}\begin{pmatrix}v_A & -\alpha v_A \\ -\alpha v_A & \alpha^2 v_A + \beta^2 v_U\end{pmatrix}\begin{pmatrix}\beta\gamma v_U \\ 0\end{pmatrix} \\ &= \begin{pmatrix}\frac{\gamma}{\beta} \\ \frac{-\alpha\gamma}{\beta}\end{pmatrix} \end{aligned} \tag{4}$$

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 2.2) 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 λ_A. Thus this predictor is counterfactually fair. □

Note that if Figure 1(d) (*Left*) is the true model for the real world then $\hat{Y}(X, A)$ will also satisfy demographic parity and equality of opportunity as $\hat{Y}$ will be unaffected by A.

The above lemma holds in a more general case for the structure given in Figure 1(d) (*Left*): any non-constant estimator that depends only on X is not counterfactually fair as changing A always alters X.

We note that, outside of this particular causal model in Figure 1 (d)(*Left*), the predictor $\hat{Y}(X, A)$ is not counterfactually fair, as described in the following scenarios.

Scenario 2: High Crime Regions Revisited. The causal structure differs from the previous example by the extra edge $X \to Y$. For illustration purposes, assume again that the model is linear. 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 λ_X, λ_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 that are not explained by U, then $\hat{Y}(X, a) \neq \hat{Y}(X, a')$ and thus $\hat{Y}(X, A)$ is not counterfactually fair.

The following extra example complements the previous two examples.

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. The structure is similar to Figure 1(d) (*Left*), with the extra edge $A \to Y$. We can again ask, is the predictor $\hat{Y}(X, A)$ counterfactually fair? In this case, the different 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 λ_X, λ_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 not explained by U, so it cannot be counterfactually fair.

S4 Case Study: Criminality vs. Perceived Criminality

We test our approach on a problem of *separating actual and perceived criminality in police stops*. For this problem, we construct a causal model, and make explicit how unfairness may affect observed and unobserved variables in the world. Given the model 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 appearance). Finally we judge how well our counterfactually fair 'criminality' score satisfies demographic parity.

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

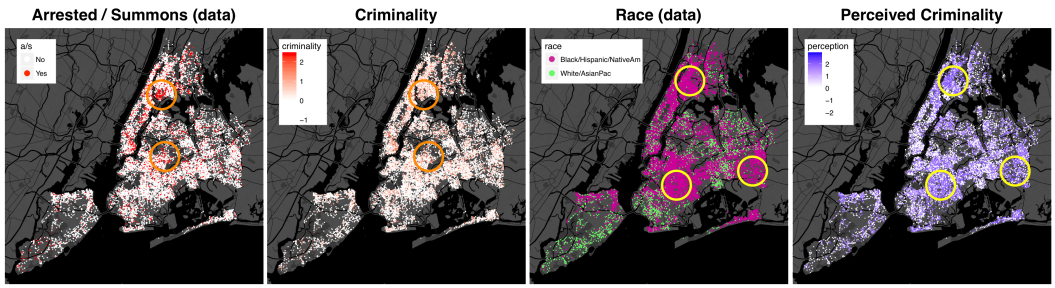

Figure 3: Understanding criminality. The above maps show the decomposition of stop and search data in New York into factors based on perceived criminality (a race dependent variable) and latent criminality (a race neutral measure). See section 6.

searched or frisked, was made or a summons issued, the data collected on males stopped during 2014 which constitutes 38,609 records. stopped as this accounts for more than 90% of the data. We fit a model which decomposes these records into two latent factors, one which depends on the race and appearance of the individual being stopped, labeled *Perceived Criminality*, and one which does not, labeled *Criminality*, and which could be used as a basis for counterfactually fair decisions. The full details of this experiment, including the DAG, are given in the supplementary materials. We now describe a spatial analysis of the estimated latent factors.

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[6]. Thus we can visualize *Criminality* and *Perception* alongside *Race* and the combination of *Arrest* and *Summons*, shown in Figure 3. 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 3. 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).

[6]https://github.com/stablemarkets/StopAndFrisk