Database Repair Meets Algorithmic Fairness

Babak Salimi, Bill Howe, Dan Suciu
University of Washington

ABSTRACT

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflect discrimination, suggesting a database repair problem. Existing treatments of fairness rely on statistical correlations that can be fooled by anomalies, such as Simpson's paradox. Proposals for causality-based definitions of fairness can correctly model some of these situations, but they rely on background knowledge of the underlying causal models. In this paper, we formalize the situation as a database repair problem, proving sufficient conditions for fair classifiers in terms of admissible variables as opposed to a complete causal model. We show that these conditions correctly capture subtle fairness violations. We then use these conditions as the basis for database repair algorithms that provide provable fairness guarantees about classifiers trained on their training labels. We demonstrate the effectiveness of our proposed techniques with experimental results.

1. INTRODUCTION

Fairness is increasingly recognized as a critical component of machine learning (ML) systems. These systems are now routinely used to make decisions that affect people's lives [7], with the aim of reducing costs, reducing errors, and improving objectivity. While this is a positive trend, there is also enormous potential for harm. The functionality of ML systems are defined by their parameters as dictated by the data used for training them. More often than not, the available data reflects societal inequities and historical biases, and, as a consequence, the models trained on such data will therefore reinforce and legitimize discrimination and opacity.

There has been a steady stream of reports of discriminatory ML systems, due to biased data, across many different domains. In 2014, a team of machine learning experts from Amazon Inc. began work on an automated system to review job applicants' resumes. According to a recent Reuters article [8], the experimental system gave job candidates scores ranging from one to five and was trained on 10 years of recruiting data from Amazon. However, by 2015 the team realized that the system showed a significant gender bias towards male candidates over females due to historical discrimination in the training data. Amazon edited the system to make it gender agnostic, but there was no guarantee that discrimination did not occur through other means, and the project was totally abandoned in 2017.

In another example, in 2016, a team of journalists from ProPublica analysed COMPAS, one of the many widely used commercial risk assessment algorithms for predicting recidivism, and revealed that it overpredicts recidivism for African-Americans and underpredicts it for Caucasians [20]. In the context of predicting recidivism (which is itself a questionable application!), fairness issue arise because these systems are trained using data on arrested individuals, as opposed to data on individuals who commit crime. Because of historical racial biases in arrest data, probabilities produced by these systems are racially biased as well.

Mitigating Bias. These examples underpin the importance of understanding and accounting for historical bias in data. A naïve (and ineffective) approach sometimes used in practice is to simply omit the protected attribute (say, race or gender) when training the classifier. However, since the protected attribute is frequently represented implicitly by some combination of proxy variables, the classifier still learns the discrimination present in training data. For example, zip code tends to predict race due to a history of segregation [13, 34]; answers to personality tests identify people with disabilities [37]; and keywords can reveal gender on a resume [8]. As a result, a classifier trained without regard to the protected attribute not only fails to remove discrimination, but it can complicate the detection and mitigation of discrimination downstream via existing techniques [29, 6, 5, 18, 17, 24, 36], such as those we describe next.

The two main approaches to reduce or eliminate sources of discrimination are summarized in Fig. 1. The most popular is in-processing, where the ML algorithm itself is modified to account for fairness during the training time; this approach must be reimplemented for every ML application. The alternative is to process either the training data (pre-processing) or the output of the classifier itself (post-processing). We advocate for the pre-processing strategy, which can be designed to be agnostic to the choice of ML algorithm and instead interprets the problem as a database repair task.

Fairness Definitions. One needs a quantitative measure of discrimination in order to remove it. A large number of fairness definitions have been proposed, which we broadly categorize in Fig. 1. The best-known measures are based on statistical (i.e., *associative*) relationships between the protected attribute and the outcome. For example, *demographic parity* requires that, for all groups of the protected attribute, the overall probability of a positive prediction of an outcome should be the same. However, it has been shown that associative definitions of fairness can be mutually exclusive [5] and fail to distinguish between discriminatory, non-discriminatory, and spurious association between a protected attribute and the outcome of an algorithm [17, 24, 9]. The following example highlights the pitfalls of associative fairness:

EXAMPLE 1.1. *In 1973, UC Berkeley was sued for discrimination against females in graduate school admissions when it was found that 34.6% of females were admitted in 1973 as opposed to 44.3% of males, hence demographic parity was violated. However, analysis revealed that the effect occurred because females tended to apply to departments with lower overall acceptance rates [30]. When broken down by department, a slight bias* toward *female applicants was observed, a result that did not constitute evidence for gender-based discrimination.*

Such situations have recently motivated a search for a more principled measure of fairness and discrimination based on *causality* [17, 24, 18, 29, 31]. These approaches assume access to background knowledge on the underlying causal models that usually visualised as directed graphs, consisting of nodes (representing variables) and directed edges between the nodes (representing potential causal relations). These approaches, then, measure discrimination as the causal influence of the protected attribute on the outcome of an algorithm, through certain causal paths that deemed to be socially unacceptable. For instance, in Example 1.1, the direct causal influence of gender on admission decisions as well as its indirect effect through applicants' hobbies might be considered as discriminatory. In terms of causal models, the former is expressed by prohibiting the directed edge from gender to admission decision, and the latter is expressed by prohibiting any directed path from gender to hiring decision that is intercepted by applicant's hobbies. However, causal approaches to fairness assume access to a complete causal model, and no existing proposals describe comprehensive systems for pre-processing data to mitigate causal discrimination.

Fairness via Database Repair. This paper describes a new approach to removing discrimination by *repairing the training data*. Our proposal is based on the following key observations: 1) In causal models, a missing arrow between two variables X and Y encodes the assumption that there exists a set of variables $\mathbf{Z}$ such that X and Y are statistically independent given $\mathbf{Z}$; denoted as the conditional independence statement $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$. Consequently, causal fairness constraints (expressed as requirements about the absence of certain causal paths from protected attributes to an outcome) can be compiled into conditional independence statements. Therefore, to enforce causal fairness, we can intervene on the data and enforce the corresponding conditional independence statements instead of intervening on the causal models over which we have no control. 2) There is a clear connection between conditional independence statements and well-studied integrity constraints in data management such as Multivalued Dependencies (MVDs) [1]. Our paper leverages these connections to frame algorithmic fairness as a database repair problem for Multivalued Dependencies. The problem of database repair has been studied for various types of constraints, for example the complexity of repairing for functional dependencies (FD) has been completely solved in [21]. However, the problem of database repairs for MVDs has received less attention and is still open. Recently, the problem of mining MVDs from data is studied in [16].

	Statistical	Causal
In-processing (Modify Algorithm)	[15, 41, 3, 24, 17]	[24, 17, 29]
Pre/post-processing (Modify input/output Data)	[10, 4, 12, 39]	CAPUCHIN (this paper)

Figure 1: Fairness metrics and enforcement methods.

Capuchin. Our system, CAPUCHIN, accepts a dataset consisting of a protected attribute (e.g., gender, race, etc.), an outcome attribute (e.g., college admissions, loan application, or hiring decisions), and a set of *admissible variables* through which it is permissible for the protected attribute to influence the outcome. For instance, the applicant's choice of department in Example 1.1 may be considered as admissible despite being correlated with gender. The system repairs the input data by inserting or removing tuples to remove the influence of the protected attribute on the outcome through any directed causal paths that includes inadmissible attributes, by means of enforcing the corresponding MVDs. That is, the repaired training data can be seen as a *sample from a counterfactual fair world*.

Unlike previous measures of fairness based on causality [24, 17, 29], which require the presence of the underlying causal model, our definition is based solely on the notion of *intervention* [25] and can be guaranteed even in the absence of causal models. The user needs only distinguish admissible and inadmissible attributes; we prove that this information is sufficient to mitigate discrimination.

We use this *interventional* approach to derive in Sec. 3.1 a new fairness definition, called *justifiable fairness*. Justifiable fairness subsumes and improves on several previous definitions and can correctly distinguish fairness violations and non-violations that would otherwise be hidden by statistical coincidences, such as Simpson's paradox. We prove next, in Sec. 3.2, that, if the training data satisfies a simple saturated conditional independence, then any reasonable algorithm trained on it will be fair.

Our core technical contribution consists of a new approach to repair training data in order to enforce the saturated conditional independence that guarantees fairness. In Sec. 4 we first define the problem formally and then present a new technique to reduce it to a multivalued functional dependency MVD [1]. Then, we introduce new techniques to repair a dataset for an MVD. In Sec. 5 we evaluate our algorithms on real data and show that they meet our goals.

2. PRELIMINARIES

This section reviewers the basic background on database repair, algorithmic fairness and causal inference, the building blocks of our paper.

We denote variables (i.e., dataset attributes) by upper-

case letters, X, Y, Z, V; their values with lowercase letters, x, y, z, v; and denote sets of variables or values using boldface ($\mathbf{X}$ or $\mathbf{x}$). The domain of a variable X is $Dom(X)$, and the domain of a set of variables is $Dom(\mathbf{X}) = \prod_{Y \in \mathbf{X}} Dom(Y)$. In this paper, all domains are discrete and finite; continuous domains are assumed to be binned, as is typical. A *database instance* D is a relation whose attributes we denote as $\mathbf{V}$. We assume set semantics (*i.e.*, no duplicates) unless otherwise stated, and we denote the cardinality of D as $n = |D|$. Given a partition $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} = \mathbf{V}$, we say that D satisfies the *multivalued dependency* (MVD) $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ if $D = \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{ZY}}(D)$.

Typically, training data for ML is a bag B. We convert it into a set D (by eliminating duplicates) and a probability distribution Pr, which accounts for multiplicities; We call D the support of Pr. We say that Pr is *uniform* if all tuples have the same probability. We say $\mathbf{X}$ and $\mathbf{Y}$ are *conditionally independent (CI)* given $\mathbf{Z}$, written $(\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y} | \mathbf{Z})$, or just $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})$, if $\Pr(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \Pr(\mathbf{x}|\mathbf{z})$ whenever $\Pr(\mathbf{y}, \mathbf{z}) > 0$. When $\mathbf{V} = \mathbf{XYZ}$, then the CI is said to be *saturated*. A uniform Pr satisfies a saturated CI iff its support D satisfies the MVD $\mathbf{Z} \twoheadrightarrow \mathbf{X}$. Training data usually does not have a uniform Pr, and in such cases the equivalence between the CI and MVD fails [38]. This issue can be addressed by converting a bag to a corresponding set; see [32] for details.

The *database repair problem* is the following: we are given a set of constraints Γ and a database instance D, and we need to perform a minimal set of updates on D such that the new database D' satisfies Γ [2].

2.1 Background on Algorithmic Fairness

Algorithmic fairness considers a *protected attribute* S, a *response variable* Y, and a prediction algorithm $A : Dom(\mathbf{X}) \to Dom(O)$, where $\mathbf{X} \subseteq \mathbf{V}$, and the prediction of A is denoted O (some references denote it $\hat{Y}$) and called *outcome*. For simplicity, we assume S classifies the population into protected $S = 1$ and privileged $S = 0$, for example, female and male. Fairness definitions can be classified as statistical or causal.

Statistical Fairness. This family of fairness definitions is based on statistical measures on the variables of interest; a summary is shown in Fig. 2. *Demographic Parity* (DP) [3, 14, 42, 35, 9], requires an algorithm to classify both the protected and the privileged group with the same probability. As we saw in Example 1.1, the lack of statistical parity cannot be considered as evidence for gender-based discrimination; this has motivated the introduction of *Conditional Statistical Parity* (CSP) [6], which controls for a set of admissible factors $\mathbf{A}$. Another popular measure used for predictive classification algorithms is *Equalized Odds* (EO), which requires that both protected and privileged groups to have the same false positive (FP) rate, and the same false negative (FN) rate. Finally, *Predictive Parity* (PP) requires that both protected and unprotected groups have the same predicted positive value (PPV) It has been shown that these measures are inconsistent [5].

Causal Fairness. Causal notions of fairness were motivated by the need to address difficulties generated by statistical fairness and assumes an underlying causal model [18, 17, 24, 29, 11]. We first discuss causal DAGs before reviewing causal fairness.

Fairness Metric	Description
Demographic Parity (DP) [9, 35]	$S \perp\!\!\!\perp O$
Conditional Statistical parity [6]	$S \perp\!\!\!\perp O \mid \mathbf{A}$
Equalized Odds (EO) [12, 40]	$S \perp\!\!\!\perp O \mid Y$
Predictive Parity (PP)[5, 35, 5, 12]	$S \perp\!\!\!\perp Y \mid O$

Figure 2: Common statistical definitions of fairness.

2.2 Background on Causal DAGs

Causal DAG. A *causal DAG* G over set of variables $\mathbf{V}$ is a directed acyclic graph that models the functional interaction between variables in $\mathbf{V}$. Each node X represents a variable in $\mathbf{V}$ that is functionally determined by: (a) its parents $\mathbf{Pa}(X)$ in the DAG, and (b) some set of *exogenous* factors that need not appear in the DAG, as long as they are mutually independent. This functional interpretation leads to the same decomposition of the joint probability distribution of $\mathbf{V}$ that characterizes Bayesian networks [25]:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X | \mathbf{Pa}(X)) \tag{1}$$

***d*-Separation and Faithfulness.** A common inference question in a causal DAG is how to determine whether a CI $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})$ holds. A sufficient criterion is given by the notion of d-separation, a syntactic condition $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ that can be checked directly on the graph. Pr and G are called *Markov compatible* if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ implies $(\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y} | \mathbf{Z})$; if the converse implication holds, then we say that Pr is *faithful* to G. If G is a causal DAG and Pr is given by Eq.(1), then they are Markov compatible [26].

Counterfactuals and `do` Operator. A *counterfactual* is an intervention where we actively modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. Pearl [25] described the *do* operator that allows this effect to be computed on a causal DAG, denoted $\Pr(Y|do(X = x))$. To compute this value, we assume that X is determined by a constant function $X = x$ instead of a function provided by the causal DAG. This assumption corresponds to a modified graph with all edges into $\mathbf{X}$ removed, and values of $\mathbf{X}$ are set to $\mathbf{x}$. The Bayesian rule Eq.(1) for the modified graph defines $\Pr(Y|do(\mathbf{X} = \mathbf{x}))$; the exact expression is in [25, Theorem 3.2.2]. We proved and illustrated the following in [32]:

THEOREM 2.1. *Given a causal DAG G and a set of variables $\mathbf{X} \subseteq \mathbf{V}$, suppose $\mathbf{X} = \{X_0, X_1 \ldots, X_m\}$ are ordered such that X_i is a non-descendant of X_{i+1} in G. The effect of a set of interventions $do(\mathbf{X} = \mathbf{x})$ is given by the following* extended adjustment formula*:*

$$\Pr(y|do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z} \in Dom(\mathbf{Z})} \Pr(y|\mathbf{x}, \mathbf{z}) \left(\prod_{i=0}^{m} \Pr\left(\mathbf{pa}(X_i) \middle| \bigcup_{j=0}^{i-1} \mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j \right) \right) \tag{2}$$

where $\mathbf{Z} = \bigcup_{X \in \mathbf{X}} \mathbf{Pa}(X)$ *and* $j \geq 0$.

2.3 Causal Fairness

Counterfactual Fairness. Kusner et al. [18, 19] (see also the discussion in [22]) define a classifier as *counterfactually fair* if the protected attribute of an individual is not a cause of the outcome of the classifier for that individual, i.e., had the protected attributes of the individual been different, and other things being equal, the outcome of the predictor

would have remained the same. However, it is known that individual-level counterfactuals can not be estimated from data in general [26].

Proxy Fairness. To avoid individual-level counterfactuals, a common approach is to study population-level counterfactuals or interventional distributions that capture the effect of interventions at the population level rather than an individual level [26, 27, 28]. Kilbertus et. al. [17] defined proxy fairness as follows:

$$P(\tilde{Y} = 1|do(\mathbf{P} = \mathbf{p})) = P(\tilde{Y} = 1|do(\mathbf{P} = \mathbf{p}')) \quad (3)$$

for any $\mathbf{p}, \mathbf{p}' \in Dom(\mathbf{P})$, where $\mathbf{P}$ consists of proxies to a sensitive variable S (and might include S). Intuitively, a classifier satisfies proxy fairness in Eq 3, if the distribution of $\tilde{Y}$ under two interventional regimes in which $\mathbf{P}$ set to $\mathbf{p}$ and $\mathbf{p}'$ is the same. Thus, proxy fairness is not an individual-level notion. The next example shows that proxy fairness fails to capture group-level discrimination in general.

Example 2.2. *To illustrate the difference between counterfactual and proxy fairness, consider the following college admission example. Both departments make decisions based on students' gender and qualifications, $O = f(G, D, Q)$, where, O stands for admission decision and G, D and Q are binary variables that respectively stands for applicants' gender, their choice of department and qualifications. The causal DAG is $G \to O, D \to O, Q \to O$. Let $D = U_D$ and $Q = U_Q$, where U_D and U_Q are exogenous factors that are independent and that are uniformly distributed, e.g., $P(U_Q = 1) = P(U_Q = 0) = \frac{1}{2}$. Further suppose $f(G, 'A', Q) = G \wedge Q$ and $f(G, 'B', Q) = (1-G) \wedge Q$, i.e., dep. A admits only qualified males and dep. B admits only qualified females. This admission process is proxy-fair, because $P(O = 1|do(G = 1)) = P(O = 1|do(G = 0)) = \frac{1}{2}$. On the other hand, it is clearly individually-unfair, in fact it is group-level unfair (for all applicants to the same department).*

Path-Specific Fairness. These definitions are based on graph properties of the causal graph, *e.g.*, prohibiting specific paths from the sensitive attribute to the outcome [24, 22]; however, identifying path-specific causality from data requires very strong assumptions and is often impractical.

3. DEFINING AND ENFORCING FAIRNESS

In this section we introduce a new definition of fairness, which, unlike proxy fairness [17], correctly captures group-level fairness, and unlike counterfactual fairness [18, 19] is based on the standard notion of intervention and hence is testable from the data. In the next section we will describe how to repair an unfair training dataset to enforce fairness.

3.1 Interventional Fairness

In this section we assume that the causal graph is given. The algorithm computes an output variable O from input variables $\mathbf{X}$ (Sec. 2.1). We begin with a definition describing when an outcome O is causally independent of the protected attribute S for any possible configuration of a given set of variables $\mathbf{K}$.

Definition 3.1 ($\mathbf{K}$-fair). *Fix a set of attributes $\mathbf{K} \subseteq \mathbf{V} - \{S, O\}$. We say an algorithm $\mathcal{A} : Dom(\mathbf{X}) \to Dom(O)$ is $\mathbf{K}$-fair w.r.t. a protected attribute S if, for any context $\mathbf{K} = \mathbf{k}$ and every outcome $O = o$, the following holds:*

$$\Pr(O = o|do(S = 0), do(\mathbf{K} = \mathbf{k})) = \Pr(O = o|do(S = 1), do(\mathbf{K} = \mathbf{k})) \quad (4)$$

We call an algorithm *interventionally fair* if it is $\mathbf{K}$-fair for every set $\mathbf{K}$. Unlike proxy fairness, this notion correctly captures group-level fairness, because it ensures that S does not affect O in *any configuration* of the system obtained by fixing other variables at some arbitrary values. Unlike counterfactual fairness, it does not attempt to capture fairness at the individual level, and therefore it uses the standard definition of intervention (the `do`-operator). In fact, we argue that interventional fairness is the strongest notion of fairness that is testable from data, yet correctly captures group-level fairness. We illustrate with an example (see also Ex 3.6).

Example 3.2. *In contrast to proxy fairness, interventional fairness correctly identifies the admission process in Ex. 2.2 as unfair at department-level. This is because the admission process fails to satisfy $\{D\}$-fairness since, $P(O = 1|do(G = 0), do(D = 'A')) = 0$ but $P(O = 1|do(G = 1), do(D = 'A')) = \frac{1}{2}$. Therefore, interventional fairness is a more fine-grained notion than proxy fairness. We note however that, interventional fairness does not guarantee individual fairness in general. To see this suppose the admission decisions in both departments are based on student's gender and an unobserved exogenous factor U_O that is uniformly distributed, i.e., $O = f(G, U_O)$, such that $f(G, 0) = G$ and $f(G, 1) = 1 - G$. Hence, the causal DAG is $G \to O$. Then the admission process is $\emptyset$-fair because, $P(O = 1|do(G = 1)) = P(O = 1|do(G = 0)) = \frac{1}{2}$. Therefore, it is interventionally fair (since $\mathbf{V} - \{O, G\} = \emptyset$). However, it is clearly unfair at individual level. If the variable U_o were endogenous (i.e. known to the algorithm), then the admission process is no longer interventionally fair, because it is not $\{U_o\}$-fair: $P(O = 1|do(G = 1), do(U_o = 1)) = P(O = 1|G = 1, U_o = 1) = 0$, while $P(O = 1|do(G = 1), do(U_o = 1)) = P(O = 1|G = 0, U_o = 1) = 1$.*

In practice, interventional fairness is too restrictive, as we show below. To make it practical, we allow the user to classify variables into *admissible* and *inadmissible*. The former variables through which it is permissible for the protected attribute to influence the outcome. In Example 1.1, the user would label department as admissible since it is considered a fair use in admissions decisions, and would (implicitly) label all other variables such as hobby as inadmissible. Only users can identify this classification, and therefore admissible variables are part of the problem definition:

Definition 3.3 (Fairness application). *A fairness application over a domain $\mathbf{V}$ is a tuple $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$, where $\mathcal{A} : Dom(\mathbf{X}) \to Dom(O)$ is an algorithm mapping input variables $\mathbf{X} \subseteq \mathbf{V}$ to an outcome $O \in \mathbf{V}$, $S \in \mathbf{V}$ is the protected attribute, and $\mathbf{A} \cup \mathbf{I} = \mathbf{V} - \{S, O\}$ is a partition of the variables into admissible and inadmissible.*

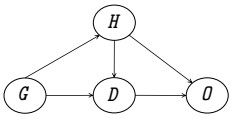
We can now introduce our definition of fairness:

Definition 3.4 (Justifiable fairness). *A fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ is* justifiably fair *if it is $\mathbf{K}$-fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$.*

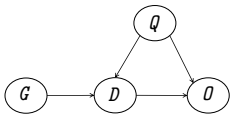
Notice that interventional fairness corresponds to the case where no variable is admissible, i.e., $\mathbf{A} = \emptyset$.

We give next a characterization of justifiable fairness in terms of the structure of the causal DAG:

Here D is not a proxy to G, because $D \perp\!\!\!\perp G$ by assumption.



a) College I



b) College II

College I	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	16	20	16	80	32	100
Female	16	80	16	20	32	100

College II	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	10	10	40	90	50	100
Female	40	50	10	50	50	100

Figure 3: Admission process representation in two colleges where the associational notions of fairness fail (see Ex.3.6).

THEOREM 3.5. *If all directed paths from S to O go through an admissible attribute in $\mathbf{A}$, then the algorithm is justifiably fair. If the probability distribution is faithful to the causal DAG, then the converse also holds.*

To ensure interventional fairness, a sufficient condition is that there exists no path from S to O in the causal graph (because $\mathbf{A} = \emptyset$). Hence, under faithfulness, interventional fairness implies fairness at individual-level, i.e., intervening on the sensitive attribute does not change the counterfactual outcome of individuals. Since this is too strong in most scenarios, we adopt justifiable fairness instead. We illustrate with an example.

EXAMPLE 3.6. *Fig 3 shows how fair or unfair situations may be hidden by coincidences but exposed through causal analysis. In both examples, the protected attribute is gender G, and the admissible attribute is department D. Suppose both departments in College I are admitting only on the basis of their applicants' hobbies. Clearly, the admission process is discriminatory in this college because department A admits 80% of its male applicants and 20% of the female applicants, while department B admits 20% of male and 80% of female applicants. On the other hand, the admission rate for the entire college is the same 32% for both male and female applicants, falsely suggesting that the college is fair. Suppose H is a proxy to G such that $H = G$ (G and H are the same), then proxy fairness classifies this example as fair: indeed, since Gender has no parents in the causal graph, intervention is the same as conditioning, hence $\Pr(O = 1|do(G = i)) = \Pr(O = 1|G = i)$ for $i = 0, 1$. Of the previous methods, only conditional statistical parity correctly indicates discrimination. We illustrate how our definition correctly classifies this examples as unfair. Assuming the user labels the department D as admissible, $\{D\}$-fairness fails because, by Eq.(2), $\Pr(O = 1|do(G = 1), do(D = \text{'A'})) = \sum_h \Pr(O = 1|G = 1, D = \text{'A'}, h)\Pr(h|G = 1) = \Pr(O = 1|G = 1, D = \text{'A'}) = 0.8$, and, similarly $\Pr(O = 1|do(G = 0), do(D = \text{'A'})) = 0.2$. Therefore, the admission process is not justifiably fair.*

Now, consider the second table for College II, where both departments A and B admit only on the basis of student qualifications Q. A superficial examination of the data suggests that the admission is unfair: department A admits 80% of all females, and 100% of all male applicants; department B admits 20% and 44.4% respectively. Upon deeper examination of the causal DAG, we can see that the admission process is justifiably fair because the only path from Gender to the Outcome goes through department, which is an admissible attribute. To understand how the data could have resulted from this causal graph, suppose 50% of each gender have high qualifications and are admitted, while others are rejected, and that 50% of females apply to each department but more qualified females apply to department A than to B (80% v.s. 20%). Further, suppose fewer males apply to department A, but all of them are qualified. The algorithm satisfies demographic parity and proxy fairness but fails to satisfy conditional statistical parity since $\Pr(A = 1|G = 1, D = \text{'A'}) = 0.8$ but $\Pr(A = 1|G = 0, D = \text{'A'}) = 0.2$). Thus, conditioning on D falsely indicates discrimination in College II. One can check that the algorithm is justifiably fair, and thus our definition also correctly classifies this example; for example, $\{D\}$-fairness follows from Eq.(2): $\Pr(O = 1|do(G = i), do(D = d)) = \sum_q \Pr(O = 1|G = i, d, q))\Pr(q|G = i) = \frac{1}{2}$. To summarize, unlike previous definitions of fairness, justifiable fairness correctly identifies College I as discriminatory and College II as fair.

3.2 Testing Fairness on the Training Data

In this section we introduce a sufficient condition for testing justifiable fairness, which uses only the training data $D, \Pr$ (Sec. 2) and does not require access to the causal graph G. We assume only that G and Pr are Markov compatible (Sec. 2.2). The training data has an additional response variable Y. As before, we assume a fairness application ($\mathcal{A}$, $S, \mathbf{A}, \mathbf{I}$) is given and that the algorithm is a good prediction of the response variable, *i.e.* $\Pr(Y = 1|\mathbf{X} = \mathbf{x}) \approx \Pr(O = 1|\mathbf{X} = \mathbf{x})$; we call the algorithm a *reasonable* classifier to indicate that it satisfies this condition. Note that this is a typical assumption in pre-processing approaches such as [4] and is needed to decouple the the issues of model accuracy and fairness. If the distributions of $\Pr(Y = 1|\mathbf{X} = \mathbf{x})$ and $\Pr(O = 1|\mathbf{X} = \mathbf{x})$ could be arbitrarily far apart, no fairness claims can be made about a classifier that, for example, imposes a pre-determined distribution on the outcome predictions rather than learning an approximation of $\Pr(Y = 1|\mathbf{X} = \mathbf{x})$ from the training data.

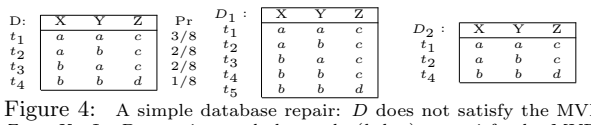
We first establish a technical condition for fairness based on the Markov boundary, and then simplify it. Recall that given a probability distribution Pr, the *Markov boundary* of a variable $Y \in \mathbf{V}$, denoted $\mathbf{MB}(Y)$, is a minimal subset of $\mathbf{V} - \{Y\}$ that satisfies the saturated conditional independence $(Y \perp\!\!\!\perp_{\Pr} \mathbf{V} - (\mathbf{MB}(Y) \cup \{Y\})|\mathbf{MB}(Y))$. Intuitively, $\mathbf{MB}(Y)$ shields Y from the influence of other variables. We prove:

THEOREM 3.7. *A sufficient condition for a fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ to be justifiably fair is $\mathbf{MB}(O) \subseteq \mathbf{A}$.*

The condition in Theorem 3.7 can be checked without knowing the causal DAG, but requires the computation of the Markov boundary; moreover, it is expressed in terms of the outcome O of the algorithm. We derive from here a sufficient condition that refers only to the response variable Y present in the training data.

COLOLLARY 3.8. *Fix a training data $D, \Pr$, where $Y \in \mathbf{V}$ is the training label, and $\mathbf{A}, \mathbf{I}$ are admissible and inadmissible attributes. Then any reasonable classifier trained on a set of variables $\mathbf{X} \subseteq \mathbf{V}$ is justifiably fair w.r.t. a protected attribute S, if either: (a) $\Pr$ satisfies the CI $(Y \perp\!\!\!\perp \mathbf{X} \cap \mathbf{I}|\mathbf{X} \cap \mathbf{A})$, or (b) $\mathbf{X} \supseteq \mathbf{A}$ and $\Pr$ satisfies the saturated CI $(Y \perp\!\!\!\perp I|\mathbf{A})$.*

While condition (a) is the weaker assumption, condition (b) has the advantage that the CI is saturated. Our method



D:	X	Y	Z	Pr
t_1	a	a	c	3/8
t_2	a	b	c	2/8
t_3	b	a	c	2/8
t_4	b	b	d	1/8

D_1 :

	X	Y	Z
t_1	a	a	c
t_2	a	b	c
t_3	b	a	c
t_4	b	b	c
t_5	b	b	d

D_2 :

	X	Y	Z
t_1	a	a	c
t_2	a	b	c
t_4	b	b	d

Figure 4: A simple database repair: D does not satisfy the MVD $Z \twoheadrightarrow X$. In D_1, we inserted the tuple (b, b, c) to satisfy the MVD, and in D_2 we deleted the tuple (b, a, c) to satisfy the MVD.

for building a fair classifier is to repair the training data in order to enforce (b).

3.3 Building Fair Classifiers

A naive way to satisfy Corollary 3.8(a) is to set $\mathbf{X} = \mathbf{A}$, in other words to train the classifier only on admissible attributes This method guarantees fairness, but it is impractical and can negatively affect the accuracy of the classifier [32]. Instead, our approach is to repair the training data to enforce the condition in Corollary 3.8(b). We consider the saturated CI $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$ as an *integrity constraint* that should always hold in training data $D, \Pr$. CAPUCHIN performs a sequence of database updates (insertions and deletions of tuples) to obtain another training database D' to satisfy $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. We describe this repair problem in Sec. 4. In terms of the causal DAG, this approach can be seen as modifying the underlying causal model to enforce the fairness constraint. However, instead of intervening on the causal DAG, over which we have no control, we intervene on the data to ensure fairness. *Note that minimal repairs are crucial for preserving the utility of data.*

4. DATA REPAIR TO ENSURE FAIRNESS

We have shown in Corollary 3.8 that, if the training data D satisfies a certain saturated conditional independence (CI), then a classification algorithm trained on $D, \Pr$ is justifiably fair. We show here how to modify (repair) the training data to enforce the CI and thus ensure that any reasonable classifier trained on it will be justifiably fair.

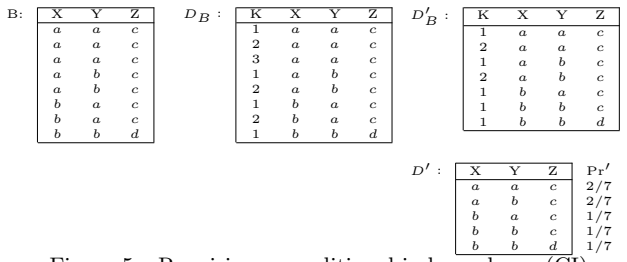
4.1 Minimal Repair for MVD and CI

We first consider repairing an MVD. Fix an MVD $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ and a database D that does not satisfy it. The minimal database repair problem is this: find another database D' that satisfies the MVD such that the distance between D and D' is minimized. In this section, we restrict the distance function to the symmetric difference, i.e, $|\Delta(D, D')|$.

EXAMPLE 4.1. *Consider the database D in Fig. 4 (ignoring the probabilities for the moment), and the MVD $Z \twoheadrightarrow X$. D does not satisfy the MVD. The figure shows two minimal repairs, D_1, D_2, one obtained by inserting a tuple, and the other by deleting a tuple.*

However, our problem is to repair for a saturated CI, not an MVD, since that is what is required in Corollary 3.8. The repair problem for a database constraint is well-studied in the literature, but here we need to repair to satisfy a CI, which is not a database constraint. We first formally define the repair problem for a CI and then show how to reduce it to the repair for an MVD. More precisely, our input is a database D and a probability distribution Pr, and the goal is to define a "repair" $D', \Pr'$ that satisfies the given CI.

We assume that all probabilities are rational numbers. Let the *bag associated* to $D, \Pr$ be the smallest bag B such that Pr is the empirical distribution on B. In other words, B is



B:

X	Y	Z
a	a	c
a	a	c
a	a	c
a	b	c
a	b	c
b	a	c
b	a	c
b	b	d

D_B :

K	X	Y	Z
1	a	a	c
2	a	a	c
3	a	a	c
1	a	b	c
2	a	b	c
1	b	a	c
2	b	a	c
1	b	b	d

D'_B :

K	X	Y	Z
1	a	a	c
2	a	a	c
1	a	b	c
2	a	b	c
1	b	a	c
1	b	b	c
1	b	b	d

D' :

X	Y	Z	Pr′
a	a	c	2/7
a	b	c	2/7
b	a	c	1/7
b	b	c	1/7
b	b	d	1/7

Figure 5: Repairing a conditional independence (CI).

obtained by replicating each tuple $t \in D$ a number of times proportional to $\Pr(t)$. If Pr is uniform, then $B = D$.

DEFINITION 4.2. *The minimal repair of $D, \Pr$ for a saturated CI $(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$ is a pair $D', \Pr'$ such that $\Pr'$ satisfies the CI and $|\Delta(B, B')|$ is minimized, where B and B' are the bags associated with $D, \Pr$ and $D', \Pr'$, respectively.*

Recall that $\mathbf{V}$ denotes the set of attributes of D. Let Pr be any probability distribution on the variables $\{K\} \cup \mathbf{V}$, where K is a fresh variable not in $\mathbf{V}$.

LEMMA 4.3. *If* Pr *satisfies $(K\mathbf{X}; \mathbf{Y}|\mathbf{Z})$, then it also satisfies $(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$.*

We now describe our method for computing a minimal repair of $D, \Pr$ for some saturated CI. First, we compute the bag B associated to $D, \Pr$. Next, we add the new attribute K to the tuples in B and assign distinct values to $t.K$ to all duplicate tuples t, thus converting B into a set D_B with attributes $K \cup \mathbf{V}$. Importantly, we use as few distinct values for K as possible, i.e., we enumerate the instances of each unique tuple. More precisely, we define:

$$D_B = \big\{(i, t) \mid t \in B, i = 1, \ldots, |t_B|\big\} \tag{5}$$

were $|t_B|$ denotes the number of occurrences (or multiplicity) of a tuple t in the bag B. Then, we repair D_B w.r.t. to the MVD $\mathbf{Z} \twoheadrightarrow K\mathbf{X}$, obtaining a repaired database D'_B. Finally, we construct a new training set $D' = \Pi_{\mathbf{V}}(D'_B)$, with the probability distribution obtained by marginalizing the empirical distribution on D'_B to the variables $\mathbf{V}$.

EXAMPLE 4.4. *Fig 4 shows two repairs D_1 and D_2 of the database D, in Example 4.1, w.r.t the MVD $Z \twoheadrightarrow X$. Consider now the probability distribution,* Pr *shown in the figure. Suppose we want to repair it for the CI $(X; Y|Z)$. Clearly, both D_1 and D_2, when endowed with the empirical distribution* do *satisfy this CI, but they are very poor repairs because they completely ignore the probabilities in the original training data, which are important signals for learning. Our definition captures this by insisting that the repaired bag B' be close to the bag B associated to $D, \Pr$ (see B in Fig. 5), but the sets D_1 and D_2 are rather far from B. Instead, our method first converts B into a set D_B by adding a new attribute K (see Fig. 5) then, it repairs D_B for the MVD $Z \twoheadrightarrow KX$, obtaining D'_B. The final repair $D', \Pr'$ consists of the empirical distribution on D'_B, but with the attribute K and duplicates removed.*

The problem of computing minimal repairs for MVDs and CIs, as introduced in this section, is essentially an optimization problem. A suit of techniques for addressing these problems has been introduced in [33, 32] that exploit reduction to the MaxSAT and Matrix Factorization.

5. EXPERIMENTAL RESULTS

This section presents experiments that evaluate the feasibility and efficacy of CAPUCHIN. We aim to evaluate the end-to-end performance of CAPUCHIN in terms of utility and fairness, with respect to our repair method. We refer the reader to [32] for more experiments.

5.1 Setup

We report the empirical utility of each classifier using Accuracy (ACC) via 5-fold cross-validation. We evaluate using three classifiers: Linear Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF).

To assess the effectiveness of the proposed approaches, we used the ratio of observational discrimination (ROD) defined in [32] as follows: Given a fairness application ($\mathcal{A}$, $S, \mathbf{A}, \mathbf{I}$), let $\mathbf{A}_b = \mathbf{MB}(O) - \mathbf{I}$. We quantify the *ratio of observational discrimination (ROD)* of $\mathcal{A}$ against S in a context $\mathbf{A}_b = \mathbf{a}_b$ as $\delta(S; O|\mathbf{a}_b) \stackrel{\text{def}}{=} \frac{\Pr(O=1|S=0,\mathbf{a}_b)\Pr(O=0|S=1,\mathbf{a}_b)}{\Pr(O=0|S=0,\mathbf{a}_b)\Pr(O=1|S=1,\mathbf{a}_b)}$. Intuitively, ROD calculates the effect of membership in a protected group on the odds of the positive outcome of $\mathcal{A}$ for subjects that are similar on $\mathbf{A}_b = \mathbf{a}_b$ ($\mathbf{A}_b$ consists of admissible attributes in the Markov boundary of the outcome). ROD is sensitive to the choice of a context $\mathbf{A}_b = \mathbf{a}_b$ by design. The overall ROD denoted by $\delta(S, O|\mathbf{A}_b)$ can be computed by averaging $\delta(S, O|\mathbf{a}_b)$ for all $\mathbf{a}_b \in \mathbf{A}_b$.

5.2 End-To-End Results

In the following experiments, a fairness constraint was enforced on training data using CAPUCHIN repair algorithms (cf. Sec 4). Specifically, each dataset was split into five training and test datasets. All training data were repaired separately using Matrix Factorization (MF), Independent Coupling (IC) and two versions of the MaxSAT approach (see [32] for details of MF and IC methods): MS(Hard), which feeds all clauses of the lineage of a CI into MaxSAT, and MS(Soft), which only feeds small fraction of the clauses. We tuned MaxSAT to enforce CIs approximately. We then measured the utility and discrimination metrics for each repair method as explained in Sec 5.1. For all datasets, the chosen training variables included the Markov boundary of the outcome variables, which were learned from data using the Grow-Shrink algorithm [23] and permutation test [30].

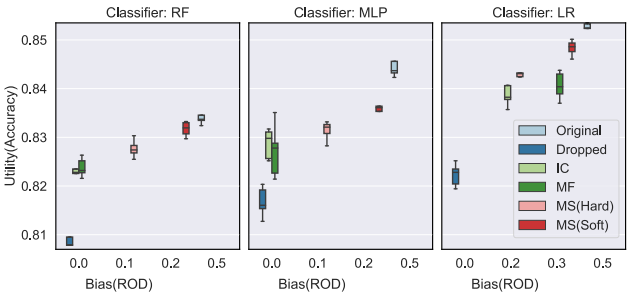


Figure 6: Performance of CAPUCHIN on Adult data.

Adult data. This data reflects historical income inequality that can be reinforced by ML algorithms. We used CAPUCHIN to remove the mentioned sources of discrimination from Adult data. Specifically, we categorized the attributes in the Adult dataset as follows: (**S**) sensitive attributes: gender (male, female); (**A**) admissible attributes: hours per week, occupation, age, education, etc.; (**N**) inadmissible attributes: marital status; (Y) binary outcome: high income. As is common in the literature, we assumed that the potential influence of gender on income through some or all of the admissible variables was fair; However, the direct influence of gender on income, as well as its indirect influence on income through marital status, were assumed to be discriminatory. To remove the bias, we enforced the CI ($Y \perp\!\!\!\perp \mathbf{S}, \mathbf{N}|\mathbf{D}$) on training datasets using the CAPUCHIN repair algorithms. Then, we trained the classifiers on both original and repaired training datasets using the set of variables $\mathbf{A} \cup \mathbf{N} \cup \mathbf{S}$. We also trained the classifiers on original data using only $\mathbf{A}$, i.e., we dropped the sensitive and inadmissible variables.

Fig. 6 compares the utility and bias of CAPUCHIN repair methods on Adult data. As shown, our repair methods delivered surprisingly good results: when partially repairing data using the MaxSAT approach, i.e, using MS(Soft), almost 50% of the bias was removed while accuracy decreased by only 1%.

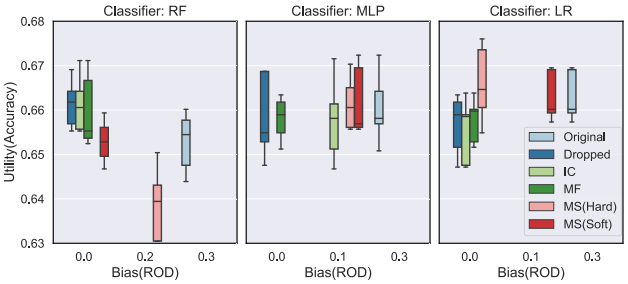


Figure 7: Performance of CAPUCHIN on COMPAS data.

COMPAS. For the second experiment, we used the ProPublica COMPAS dataset [20]. This dataset contains records for all offenders in Broward County, Florida in 2013 and 2014. We categorized the attributes in COMPAS data as follows: (**S**) protected attributes: race (African American, Caucasian); (**A**) admissible attributes: number of prior convictions, severity of charge degree, age; **(Y)** binary outcome: a binary indicator of whether the individual is a recidivist. As is common in the literature, we assumed that it was fair to use the admissible attributes to predict recidivism even though they can potentially be influenced by race, and our only goal in this experiment was to address the direct influence of race. We pursued the same steps as explained in the first experiment. Fig. 7 compares the bias and utility of CAPUCHIN repair methods to original data. As shown, all repair methods successfully reduced the ROD. However, we observed that MF and IC performed better than MS on COMPAS data (as opposed to Adult data).

6. CONCLUSIONS

We considered a causal approach for fair ML, reducing it to a database repair problem. We showed that conventional associational and causal fairness metrics can over- and under-report discrimination. We defined a new notion of fairness, called *justifiable fairness*, that addresses shortcomings of the previous definitions and argued that it is the strongest notion of fairness that is testable from data. We then proved sufficient properties for justifiable fairness and use these results to translate the properties into saturated conditional independence that we can be seen as multivalued

dependencies with which to repair the data. We then proposed multiple algorithms for implementing these repairs. Our experimental results show that our algorithms successfully mitigate discrimination due to biased training data, are robust to unseen test data.

7. REFERENCES

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] Leopoldo E. Bertossi. *Database Repairing and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

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

[4] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[5] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[7] Rachel Courtland. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558, 2018.

[8] Jeffrey Dastin. Rpt-insight-amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[11] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

[12] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[13] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, 2016. `www.bloomberg.com/graphics/2016-amazon-same-day/`.

[14] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[15] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[16] Batya Kenig, Pranay Mundra, Guna Prasaad, Babak Salimi, and Dan Suciu. Mining approximate acyclic schemes from relations. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, June 14-19, 2020*, pages 297–312. ACM, 2020.

[17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[18] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[19] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *CoRR*, abs/1703.06856, 2017.

[20] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[21] Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. Computing optimal repairs for functional dependencies. *ACM Transactions on Database Systems (TODS)*, 45(1):1–46, 2020.

[22] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[23] Dimitris Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[24] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[25] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[27] Donald B Rubin. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1970.

[28] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[29] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[30] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1021–1035. ACM, 2018.

[31] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. Causal relational learning. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, June 14-19, 2020*, pages 241–256. ACM, 2020.

[32] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

[33] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810. ACM, 2019.

[34] Andrew D Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[35] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[36] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 440:1–440:14, 2018.

[37] Lauren Weber and Elizabeth Dwoskin. Are workplace personality tests fair? *Wall Strreet Journal*, 2014.

[38] SK Michael Wong, Cory J. Butz, and Dan Wu. On the implication problem for probabilistic conditional independency. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(6):785–805, 2000.

[39] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1920–1953, 2017.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.

[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 962–970, 2017.

[42] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.