THE VALIDITY OF COUNTERFACTUAL REASONING FOR FAIR PREDICTION WITH IMPERFECT MODELS

A PREPRINT

Jordan Menter
Department of Computer Science
University of Massachusetts
Amherst, MA 01002

January 30, 2023

ABSTRACT

Existing work on machine learning for fair prediction has leveraged causal reasoning, specifically counterfactuals, to account for the counterfactual nature of societal discrimination. Methods that leverage this reasoning require information about the true data-generating process (DGP). However, the true DGP is often subject to mechanisms of model misspecification, including latent confounding, measurement error, and selection bias, all potentially resulting from societal discrimination. When is the validity of counterfactual reasoning for fairness impacted by these mechanisms? To articulate an answer, we describe distributional parities that must hold in training data in order for $\hat{Y}$ to be counterfactually fair, regardless of the algorithm or estimator used to train $\hat{Y}$. Using the syntax of graphical models and d-separation, we describe the structures of model misspecification that (do not) threaten these requirements. We note that well-documented societal biases in observed data can result in model misspecifications that cause $\hat{Y}$ to violate counterfactual fairness. Empirically, we use a real dataset to investigate the impact of these forms of model misspecification on the task of fair prediction.

1 Introduction

Existing work on machine learning for the task of fair prediction has leveraged counterfactual reasoning to account for the counterfactual nature of societal discrimination. However, counterfactual reasoning requires assumptions about the true data-generating process (DGP) that are often untestable. The true DGP is also often subject to mechanisms that introduce misspecification of the true DGP, including latent confounding, selection, and measurement error. Given these phenomena, we can define a general research question: how is the validity of counterfactual reasoning for fairness impacted by these mechanisms of model misspecification? We hypothesize that the presence of the forms of model misspecification considered in this paper prevent the predictor $\hat{Y}$ from satisfying counterfactual fairness.

Our contributions are as follows:

- We articulate the “mechanisms" of counterfactual fairness by describing distributional parities that must hold in the observed data in order for $\hat{Y}$ to be counterfactually fair, regardless of the algorithm or estimator used to train $\hat{Y}$.
- We consider three canonical forms of causal model misspecification - latent confounding, endogenous selection bias, and (random and differential) measurement error - and use the syntax of graphical models to show where these misspecifications (do not) threaten the validity of these “mechanisms" of counterfactual fairness in general.
- For endogenous selection bias and measurement error, we show empirically that these forms of causal model misspecification result in statistically significant differences between predictions made by $\hat{Y}$ for original and

counterfactual values of sensitive variables - in short, $\hat{Y}$ is not counterfactually fair. This is the case even if $\hat{Y}$ is trained on non-descendants of sensitive variables.

2 Related Work

The fairness tasks to which causal reasoning has been applied can be broadly partitioned into three areas: prediction (postulating an input causal model which informs the construction of predictor $\hat{Y}$), explanation (characterizing mechanisms of societal discrimination using causal models) [35], and learning fair interventions [2, 12]. There are further finer-grained partitions within the literature, e.g. Salimi et al. use database repair algorithms to provide provable fairness guarantees on classifiers [27]. Additionally, an emerging line of work leverages causal models to account for downstream impacts of so-called fair predictions; [15] attempts to mitigate the discriminatory impact of real-world decisions on individuals, while [8] models the long-term dynamics of fair decisions in the context of dynamical systems (i.e. “feedback loops").

These partitions are somewhat in flux — Madras et al. argue for reframing fair classification as an intervention problem, for example [20] — but a non-trivial body of algorithmic fairness literature concerns constructing a fair predictor. This is where we focus our attention.

Causal methods for algorithmic fairness allow practitioners to accommodate causal influences at decision time. As causal reasoning deals in technical conceptions of causes, interventions, and counterfactuals, it captures core concepts of legal and ethical characterizations of justice, equality, and equity. Despite these strengths, the use of causal reasoning faces significant technical difficulties. One significant hurdle is the correct specification of the input causal model of the data-generating process, from which the fair predictor is constructed. Instances of causal model misspecification such as latent confounding, selection, or measurement error (that may be, in turn, the result of societal biases) can hinder this process.

Within the algorithmic fairness literature, calls to address some forms of the model misspecification we consider have already been made [7, 21]. Researchers who propose fairness definitions requiring knowledge of (some portion of) the underlying causal graph of the DGP are clearly aware of this specification problem. Recent literature has proposed methods for learning an approximately fair predictor without knowing the true causal model [26], integrating different causal models to provide counterfactually-fair decisions [26, 34], and using explicit latent variables to resolve data-generation and process mismatch [6]. Outside of the problem of fair prediction, recent work has also considered the problem of optimal fair policy learning in the presence of certain sources of statistical bias [22]. Our work deviates from these previous efforts in that we do not propose an algorithm that ostensibly circumvents the issue of model misspecification when learning a fair predictor. Rather, we theoretically and empirically interrogate the validity of a popular causal conception of fair prediction.

As a proposed solution to the problem of model misspecification, [13] develop a set of techniques to assess the impact of latent confounding on the counterfactual fairness of additive noise models. Our work deviates from this in that (i) we consider a wider range of common forms of model misspecification beyond latent confounding, and (ii) provide general theoretical results, agnostic to the estimator used to build $\hat{Y}$, that describe when these forms of model misspecification do (not) impact the validity of $\hat{Y}$.

Examples of this critical approach towards algorithmic fairness have appeared elsewhere in the literature, including [11], which empirically evaluates the behavior of various non-causal fairness metrics under scenarios of selection bias. [33], who evaluate three mainstream fairness definitions under different and competing sets of real-word assumptions, is also an example of the purpose of this work.

3 Causal Reasoning for Fair Prediction Using Counterfactuals

This section provides a sufficient background on causal graphical models and recent work on the application of counterfactuals to the problem of fair prediction.

3.1 Causal Reasoning and Counterfactuals

As in Pearl [23], we define a causal model as the triple (U, V, F) such that

1. U is a set of latent background variables that are not caused by any variables in the set V of observable variables

2. F is a set of structural equations [4] $\{f_1, ..., f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \ \{V_i\}$ and $U_{pa_i} \subseteq U$.

Structural equations allow for the calculation of counterfactual quantities. To define a counterfactual, assume we have two observed variables A and Y, and consider the statement “the value of Y if A had taken value a". We assume that the state of an observed variable is fully determined by its background variables U and structural equations F. Thus the counterfactual is modeled as the solution for Y for a given $U = u$, where the structural equations for A are replaced with $A = a$. This is denoted as $Y_{A \leftarrow a}(u)$ or Y_a [23].

We may be interested in the probability of $Y_{A \leftarrow a}(u)$ given some evidence $W = w$ and a causal model (U, V, F), where W, A, and Y are subsets of V. This task, counterfactual inference, requires three steps:

1. Abduction: for a given prior on U, compute the posterior $P(U|W = w)$
2. Action: Substitute structural equations for A with interventional values a; this results in a modified set of equations F_a
3. Prediction: Compute the implied distribution on the remaining elements of V using F_a and $P(U|W = w)$

3.2 Counterfactual Fairness

Let A be a (set of) sensitive variable(s), i.e. race, sex, etc. [16] define a fairness definition to ensure that the predictor $\hat{Y}$ is invariant to any possible counterfactual world of A.

Result 1 (Definition 5 in [16]). *For any $X = x$ and $A = a$, predictor $\hat{Y}$ is counterfactually fair if*

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

for all y and any a’ attainable by A.

In words, this requires that for an individual with a protected attribute $A = a$, $\hat{Y}$ is the same even if we observe the counterfactual $A = a'$. Thus this method requires reasoning about individual-level counterfactuals, a difficult task [25].

To guarantee (1), [16] propose a notion of “unawareness", i.e. constructing $\hat{Y}$ using variables that are not descendants of A. They make two propositions for doing so:

1. Construct $\hat{Y}$ using all variables V that are not descendants of A, or
2. If there are no V that are not descendants of A, construct $\hat{Y}$ using fair latent variables U_V extracted by marginalizing U_V via $P(U_V|V = v, A = a)$ (e.g. in general case of A causing V and V causing Y).

4 Sources of Causal Model Misspecification

In this section, we further define the three examples of model misspecification considered in this work: *measurement error*, *latent confounding*, and *endogenous selection bias*. Note that we consider structures that impact the validity of causal inference, as opposed to biases that impact the learning process of the predictor.

These structures could occur in a data-generating process independent of existing historical or societal biases; usually, the two concepts are considered separate, see Figure 2 in [21]. In this paper, we consider examples of latent confounding and endogenous selection bias that are both societal and statistical biases. However, we strongly caution against collapsing the two, a warning that has been repeated by others in the field [17, 21]. The term “fair model" or “fair algorithm" describes an equalization of parities, but this is not equivalent to satisfying societal or ethical notions of fairness. A causal method for fairness may be robust to certain forms of model misspecification in the true DGP, but we do not claim that it is unequivocally fair in all realities as a result.

We motivate the study of these particular sources of misspecification by returning to the social science literature — these sources of error are either frequently-occurring in these fields where fairness definitions are likely to be applied, or formalize realistic phenomena postulated and studied by social scientists.

4.1 Latent Confounding

A variable Z is a confounder if it is a common cause of two observed variables X_i, X_j in the DGP; we say it *confounds* X_i and X_j [23]. When Z is latent, a spurious association is induced between X_i and X_j. Much attention in the causal inference literature has been given to the effect of latent confounding on the effect of a treatment variable A on an outcome variable Y [10]; however, we focus on the case where there is latent confounding among observed variables that are inputs to a predictive model.

Example: Race and Socioeconomic Status

Although back-door paths do not constitute causal effects, they can contribute to problematic associations between sensitive and non-sensitive attributes. For example, a complex and unobserved historical process creates an individual's race and socioeconomic status at birth, thus confounding the two [31].

4.2 Measurement Error

The impacts of measurement error are still considered an open problem in the domain of algorithmic fairness [7, 21]. In this space, considerations of measurement error have largely been partitioned into errors on labels y (*label bias*) and errors on features x (*feature bias*) [7]. Recent literature claims that feature bias is less serious than label bias; in the absence of label bias, feature bias can be mitigated via the inclusion of group membership indicators in the observed training data, by which a statistical model may accommodate for group membership [7, 21].

We extend this discussion by focusing on both random and differential measurement error. Perfect measurement of ideal features or labels is highly unrealistic, and observed data are often *proxies* for these unmeasured variables. If these proxies are generated differently across groups, differential measurement bias may arise. More formally consider treatment A and proxy P; differential measurement error U_P, which has a direct path into P, is not independent of treatment, i.e. $f(U_P|A) \neq f(U_P)$ [10].

A large body of the literature is focused on differential measurement error of a treatment variable on an outcome [10]; however, more complicated realities may arise, such as settings where treatment (i.e. a sensitive attribute, in this case) causes differential measurement error on a (set of) feature(s) that are used as inputs to a model. We consider this specific setting here.

Example: Predictive Policing & Risk Assessment Predictive policing and pretrial risk assessment applications often utilize features that appear facially neutral, but are in fact proxies for protected attributes [21]. Minority communities are more highly policed, and minorities themselves are arrested more often for crimes they did not commit [1, 18]. This results in facially-neutral variables, such as the number of prior arrests, carrying implicit information about an individual's race. The recidivism risk prediction tool COMPAS made use of the number of prior arrests [24], resulting in higher false positive rates for black defendants compared to white defendants.

4.3 Endogenous Selection Bias

Endogenous selection bias [1] is characterized by the preferential selection of units in the data-generating process. Structurally, the nature of endogenous selection bias is characterized by conditioning on a collider variable, or a descendant of a collider, at any point in the causal process [9, 10].

The collider representing the selection mechanism, s, is a binary indicator that represents entry into the data pool ($s = 1$ when the unit is in the sample, otherwise $s = 0$). Preferential selection occurs because only $P(V|s = 1)$ is available for use. In this paper, we will consider the case where s is the child of two observed variables in the DGP; see [3] for a finer-grained structural classification of selection mechanisms.

As with latent confounding, much attention in the literature has been given to the effect of endogenous selection bias on estimation of the effect of a treatment variable A on outcome Y [3, 10]. However, we again note that reality may be more complex, and focus on the case where there is endogenous selection bias caused by two observed variables that may be inputs to a predictive model.

Lastly, we note that the collider variable being conditioned on may not necessarily be an arbitrary selection mechanism s; we may also condition on another observed variable in the data-generating process, or its descendant. In this paper, however, we focus specifically on the case where we condition on s. Importantly, note that s is not part of the set of observable variables V in the DGP.

Example: Feedback in Predictive Policing

Preferential selection is a problem for the task of predictive policing. Due to historical biases, individuals are preferentially selected into the training data used to (re)build the predictive policing tool; this results in increased police

[1] As in [9], we avoid the term "selection bias" because has multiple meanings across domains.

presence in minority communities. As a result, the sample is more likely to contain minorities with higher arrest counts, or who are arrested for worse offenses. This sampling can cause a runaway feedback loop, as shown in [19].

5 Model Misspecification and the Validity of Counterfactual Fairness

In this section, we (i) define a “mechanism" of counterfactual fairness that describes distributional parities that *must* hold in D in order for $\hat{Y}$ to be counterfactually fair, and (ii) use the syntax of graphical models to prove when the aforementioned sources of causal model misspecification violate counterfactual fairness by violating this “mechanism".

5.1 The “Mechanism" of Using Counterfactuals for Fair Prediction

Because counterfactual fairness simply enforces a distributional parity, shown in Equation (1), it is difficult to define a specific core procedure that describes counterfactual fairness mechanistically. However, depending on the method use to make $\hat{Y}$ invariant to any counterfactual value of A, we can define additional distributional equalities that must hold in the training data D in order for $\hat{Y}$ to be counterfactually-fair. Once these are defined, we can articulate when aforementioned forms of model misspecification violate these distributional equalities, thus leading to a violation of (1).

Let X be the set of non-descendants of A.

5.1.1 Unawareness via Facially-Neutral Variables

This requires that the non-descendants of A be invariant to whatever counterfactual value of A we are in, i.e. $X_{A\leftarrow a}(U)$ and $X_{A\leftarrow a'}(U)$ have same distribution via the three inferential steps of counterfactual inference discussed in Section 3.1.

Thus, we are most interested in articulating where $X_{A\leftarrow a}(U)$ is (not) equivalent to $X_{A\leftarrow a'}(U)$, in the presence of different forms of model misspecification. If these two distributions are equivalent, this guarantees that any function $\hat{Y}$ of X is invariant with respect to the counterfactual values of A.

5.2 Unawareness via Extracting Latent Fair Variables

When all variables in the observed DGP are descendants of sensitive attributes, [16] suggests postulating background latent variables that act as non-deterministic causes of observable variables in the DGP; these postulated variables are then passed to $\hat{Y}$. This strategy is also suggested in [6] to mitigate some forms of causal model misspecification.

This requires that the distribution of the postulated background variables U be invariant to the counterfactual values of A, i.e. $P(U|W = w, A = a)$ is equivalent to $P(U|W = w, A = a')$.

Thus, we are most interested in articulating where the distributions of these extracted latent variables $P(U|W = w, A = a)$ and $P(U|W = w, A = a')$ are (not) equivalent. If these two distributions are equivalent, this guarantees that any function $\hat{Y}$ of U is invariant with respect to the counterfactual values of A.

5.3 Graphical Criteria for the Validity of Counterfactual Fairness

Consider the forms of model misspecification discussed in above sections. When do these forms of model misspecification (not) violate the distributional parities required for $\hat{Y}$ to be counterfactually fair? We can immediately define graphical criteria that articular where counterfactual fairness can (not) be achieved. We note that this does not require knowledge of the parametric form of these confounding, measurement, or selection mechanisms. For all following propositions, let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Also assume that $\mathcal{G}$ and the joint probability distribution Pr of V are faithful to each other [29].

Let $\hat{Y}_{fair}$ be the predictor that is the result of the application of one of the fairness methods in Section 3 to Pr. Before proceeding, we must articulate the following assumptions:

A1 Y is never an ancestor of sensitive variable(s) A.

A2 The ancestors or parents of demographic attributes are also demographic attributes (observed or otherwise).

A3 Demographic attributes referring to different individual characteristics are not ancestors or parents of each other.

A4 $\hat{Y}_{fair}$ is a $reasonable$ predictor of the response variable Y

These assumptions are meant specifically for the task of prediction. A2 and A3 specify a set of “blacklisted" edges that would not realistically occur in a real data-generating process. For example, by A2, *income* cannot cause *race*. By A3, there cannot be a directed edge between *age* and *race*. Assumption (3) also specifies that ancestors of observed demographic attributes must refer to the same individual characteristics, e.g. the parent of an individual’s *race* is *parent’s race*. This reasoning, that certain attributes are ancestrally closed, is discussed elsewhere in the algorithmic fairness literature [16].

A4 is necessary to decouple the problems of fairness and accuracy [5, 27]. For the purposes of this work, we would like to guarantee that any distance between $P(\hat{Y}_{fair}|X, A)$ and $P(Y|X, A)$ is due to changes in the “fairness" of $\hat{Y}_{fair}$ compared to the “fairness" of $P(Y|X, A)$.

We now proceed with some simple theoretical results regarding when the structures of latent confounding, endogenous selection bias, and measurement error are (not) problematic for training a predictor that satisfies counterfactual fairness.

5.3.1 Using Facially-Neutral Variables

We can define general criteria for the use of facially-neutral variables to build $\hat{Y}$, using concepts of d-separation and d-connection [23]:

Proposition 1. *If A and X are d-separated by some (possibly empty) set Z, $\{A, X\} \notin Z$, then constructing $\hat{Y}$ using X satisfies counterfactual fairness, even if A and Y are d-connected.*

Proof. By Theorem 1.2.4 in [23], if X and A are d-separated by a (possibly empty) set Z in $\mathcal{G}$, then X is independent of A conditional on Z. Therefore, $X_{A\leftarrow a}(U)$ and $X_{A\leftarrow a'}(U)$ have the same distribution by the inferential steps in Section 3.1. Thus, $\hat{Y} = f(X)$ is invariant with respect to the counterfactual values of A. □

Of course, we can then extend this result to account for cases where A and X are d-connected:

Proposition 2. *If A and X are d-connected by some set Z, $\{A, X\} \notin Z$, then constructing $\hat{Y}$ using X does not satisfy counterfactual fairness.*

Proof. By Theorem 1.2.4 in [23], if X and A are d-connected by a set Z in $\mathcal{G}$, then X is dependent on A conditional on Z. Therefore, $X_{A\leftarrow a}(U)$ and $X_{A\leftarrow a'}(U)$ cannot have the same distribution by the inferential steps in Section 3.1. Thus, $\hat{Y} = f(X)$ is not invariant with respect to the counterfactual values of A. □

Thus, we have articulated a general criteria for the use of facially-neutral variables as inputs to a $\hat{Y}$ that satisfies counterfactual fairness, agnostic to the estimator or algorithm being used to build $\hat{Y}$. Moreover, all three forms of model misspecification above can serve to d-connect X and A, even if A is not a direct cause of X. Theoretical results are provided below.

Proposition 3. *Let X be a non-descendant of A. If there is a latent variable Z that confounds A and X, then $\hat{Y}_{fair}$ will not be counterfactually fair if it is constructed using X.*

Proof. Because A and X are d-connected, this is true by Proposition 2. □

Proposition 4. *Let X be a non-descendant of A. If there is a collider that is conditioned on between A and X, then $\hat{Y}_{fair}$ will not be counterfactually fair if it is constructed using X.*

Proof. Because A and X are d-connected, this is true by Proposition 2. □

Proposition 5. *Let X be a non-descendant of A. If there is a collider that is conditioned on between A and Y, then $\hat{Y}_{fair}$ will be counterfactually fair if it is constructed using unawareness.*

Proposition 6. *Let X be a non-descendant of A. If there is a latent variable Z that confounds A and Y, then $\hat{Y}_{fair}$ will be counterfactually fair if it is constructed using X.*

Proof. Because A and X are d-separated, this is true by Proposition 1. □

We can make a similar claim for biases that result from conditioning on a collider (or a descendant of a collider):

Proposition 7. *Let X be a non-descendant of A. If there is a collider that is conditioned on between A and Y, then $\hat{Y}_{fair}$ will be counterfactually fair if it is constructed using unawareness.*

Proof. Because A and X are d-separated, this is true by Proposition 1. □

5.3.2 Extracting Fair Latent Variables

Consider the causal model with observable variables W, A, and Y, where W is a child of A and Y is a child of W. Since there are no facially neutral variables in this causal model, according to [16] the practitioner postulates variable U as a cause of W, and uses $P(U|w,a)$ to pass information about W to $\hat{Y}$. With this example in mind, we have the following results.

As in the previous section, we can use d-separation and d-connection to articulate general criteria for the use of extracted latent variables to build $\hat{Y}$:

Proposition 8. *If A and U are d-separated by some (possibly empty) set Z, $\{A,U\} \notin Z$, then constructing $\hat{Y}$ using $P(U|w,a)$ satisfies counterfactual fairness, even if A and Y are d-connected.*

Proof. This is true by the same reasoning as Proposition 1. □

Proposition 9. *If A and U are d-connected by some set Z, $\{A,U\} \notin Z$, then constructing $\hat{Y}$ using $P(U|w,a)$ does not satisfy counterfactual fairness.*

Proof. This is true by the same reasoning as Proposition 2. □

Proposition 10. *If there is a collider that is conditioned on between A and W, then constructing $\hat{Y}_{fair}$ using $P(U|w,a)$ will not be counterfactually fair.*

Proof. Because U is a parent of W, the conditioned-on collider d-connects U and A. Then by Proposition 9, the predictor $\hat{Y}$ constructed using $P(U|w,a)$ is not counterfactually-fair. □

Finally, and most interestingly, we can note where certain kinds of model misspecification do *not* lead to U and A being d-connected, thus preserving counterfactual fairness in $\hat{Y}$:

Proposition 11. *If there is a latent confounder Z that confounds A and W, then $\hat{Y}_{fair}$ will be counterfactually fair if it is constructed using $P(U|w,a)$.*

Proof. Even with the presence of Z, A and U are still d-separated, so by Proposition 8, building $\hat{Y}_{fair}$ using $P(U|w,a)$ is counterfactually-fair. □

This result holds as well in the scenario that A is d-connected with W because of differential measurement error between A and W. Thus we see that extracting latent variables to build $\hat{Y}_{fair}$ results in a counterfactually-fair $\hat{Y}$ regardless of certain forms of model misspecification (latent confounding and measurement error); this is in contrast to the usage of facially-neutral variables to build $\hat{Y}$, which is susceptible to all forms of model misspecification considered here.

6 Experiments

In this section, we will explore empirically how two of these forms of model misspecification, endogenous selection bias and measurement error, cause $\hat{Y}$ to violate (1).

6.1 Considerations on Simulation

Our experiments require knowledge of the true underlying causal model $\mathcal{M}$, which is usually difficult, if not impossible, to obtain from observational data. We thus generate synthetic data from an underlying causal model that is close to a real DGP in the following manner.

First, we apply a structure-learning algorithm to the data-generating process $\mathcal{D}$ and obtain a causal model $\mathcal{M}_{\mathcal{D}}$. For experimental convenience, we may add or delete edges from $\mathcal{M}_{\mathcal{D}}$. To bias our experiments towards realism, we apply the same set of assumptions discussed in Section 5.3 when learning $\mathcal{M}_{\mathcal{D}}$.

We then sample from $\mathcal{M}_{\mathcal{D}}$ and induce some form of model misspecification to obtain $\mathcal{D}^*$ - misspecifications and their parameterizations are shown in Table 1. Our experiments are applied on $\mathcal{D}^*$, and $\mathcal{M}_{\mathcal{D}}$ is used as a source of “ground truth".

6.2 Data

We utilize a survey conducted by the Law School Admission Council across 163 law schools in the United States [32]. This dataset contains 21,791 records on law students’ entrance exam (LSAT) scores, grade-point average (GPA) prior to

Model Misspecification Type	Parameterization
Random Measurement Error	Variance of error
Differential Measurement Error	Variance of error
Endogenous Selection Bias	Strength of selection

Table 1: Forms of causal model misspecification that may be present in the true DGP, and their parameterizations.

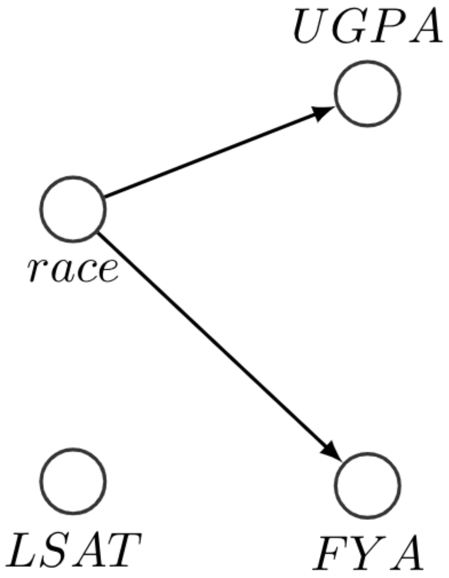


Figure 1: The structure of the true DGP used for experiments.

entering law school, and first-year average grade (FYA). The dataset also contains information on students' *race* and *sex*.

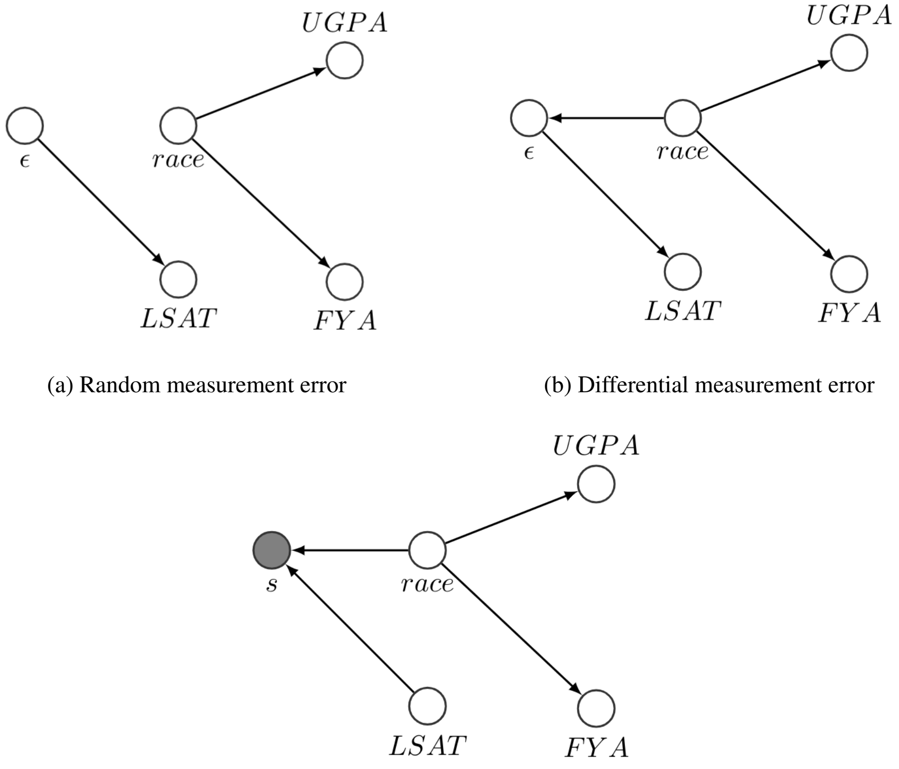
We pre-process the dataset by standardizing GPA and LSAT to have a mean of 0 and a standard deviation of 1 (FYA is already standardized in the original dataset). For simplicity, we filter records so that *race* has two levels, *White* and *Black* (retaining 19,567 out of 21,791 records). We use `bnlearn` [30] to learn the “true" DGP as a conditional linear Gausian Bayesian network.

For the sake of simplicity in our experimental pipeline, we remove the *sex* variable from our simulated DGP, so that *race* is the only sensitive variable being considered. Futhermore, we remove the edge from race to LSAT, so that we can test the method of constructing $\hat{Y}$ using only `LSAT`, a facially-neutral variable. The final true DGP used for our experiments is shown in Figure 1.

6.3 Distributions and Parameterizations for Simulation

For simulation purposes, the observed variables in the Law School dataset take on the following distributions that were learned using the R package `bnlearn` [30]. Continuous nodes with discrete parents take on distributions described by a set of linear regression models, one for each configuration of its discrete parents.

$Race \sim Bern(0.5)$
$FYA_{R_0} = 0.213 + \epsilon_{R_0} \sim \mathcal{N}(0, 0.884)$
$FYA_{R_1} = -0.828 + \epsilon_{R_1} \sim \mathcal{N}(0, 0.926)$
$UGPA_{R_0} = 0.245 + \epsilon_{R_0} \sim \mathcal{N}(0, 0.898)$
$UGPA_{R_1} = -0.719 + \epsilon_{R_1} \sim \mathcal{N}(0, 1.020)$
$LSAT = 0.047 + \epsilon \sim \mathcal{N}(0, 0.894)$



(a) Random measurement error

(b) Differential measurement error

(c) Endogenous selection bias (s is conditioned on)

Figure 2: The true DGP's under various forms of model misspecification.

	$race = 0$	$race = 1$
$LSAT \leqslant 0$	$s = 0$	$s = 1$
$LSAT > 0$	$s = 1$	$s = 0$

Table 2: CPT of selection mechanism s for experiments

We now discuss in greater detail the parameterizations used on our chosen forms of model misspecification.

Random Measurement Error In the case of random measurement error, the sources of error ϵ are independent of any observed variables in the DGP [10]. Thus for each of the measured variables *UGPA*, *LSAT*, and *FYA*, we inject varying degrees of measurement error, e.g. $UGPA_m = UGPA + \epsilon_{UGPA}$, where $\epsilon_{UGPA} \sim \mathcal{N}(0, \sigma^2)$ for values of $\sigma^2 \in \{.1, .2, .3, .4, .5, 1, 2, 3, 4, 5\}$.

Differential Measurement Error Under random measurement error, the source of error ϵ is independent of any observed variables in the DGP. Under the phenomena of differential measurement error, however, this ϵ is dependent on some or all observed variables in the DGP [10]. Specifically, we focus on instances where an individual's measured values for `LSAT` depends on their `race`. More specifically, for individuals in the data that are white, we fix the σ^2 of the associated ϵ terms at 1; for individuals in the data that are black for female, we vary $\sigma^2 \in \{.1, .2, .3, .4, .5, 1, 2, 3, 4, 5\}$. The literature on the real-world phenomena of the effect of race etc. on measured variables is mostly centered on mechanisms self-disclosure, e.g. [28] - therefore this experimental scenario does not capture exactly how societal discrimination works in practice, but is a useful inroad.

Endogenous Selection Bias In the Law School example, we can imagine a scenario wherein due to historical biases or feedback mechanism, black individuals with lower-than-average (z-standardized) LSAT scores are selected into the sample. Thus there is a collider s in the true DGP that is a child of $race$ and $LSAT$ that is not a member of the set of observed variables. We can define the conditional probability table of s in Table 2; the data is filtered to include all records where $s = 1$.

We can parameterize these experiments by varying the amount of noise present in this selection mechanism. If each cell in Table 2 is a coin flip, we can increase noise by increasing the probability that a cell takes on a random value in $\{0, 1\}$, rather than those shown.

6.4 Evaluating the Counterfactual Fairness of $\hat{Y}$

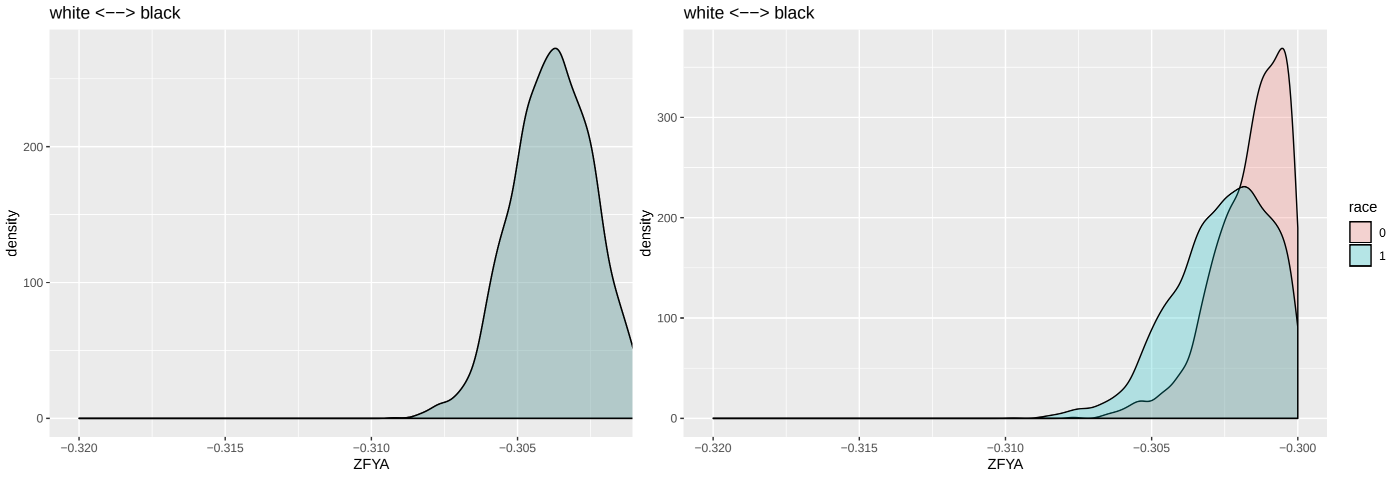
We would like to test if an algorithm is counterfactually fair under different scenarios of model misspecification. To do so, we do the following:

- Specify the true DGP from which synthetic data will be sampled
- Sample from this network and obtain a dataset D. Randomly shuffle and partition D into a training set D_{Tr} and a test set D_{Te}.
- Train $\hat{Y}$ on D_{Tr}, using only the facially-netural variable `LSAT` as an input.
- Obtain predictions Pr and Pr' of $\hat{Y}$ on D_{Te}, for both the original values of race and counterfactual values of race. If $\hat{Y}$ is counterfactually fair, the distributions of Pr and Pr' should show significant overlap. This can be measured further using the two-sample Kolmogorov Smirnov (KS) test.

7 Results

7.1 Measurement Error

From Figure 2, we would expect that under random measurement error, $\hat{Y}$ would be counterfactually fair, since `race` and `LSAT` are d-separated. Similarly, we would expect that under differential measurement error, $\hat{Y}$ would not be counterfactually fair, since in that case `race` and `LSAT` are d-connected. Figure 3 corroborates this reasoning. Under random measurement error, the original and counterfactual distributions of the predicted `FYA` appear exactly the same; under differential measurement error, we see much less overlap.



(a) Random measurement error

(b) Differential measurement error

Figure 3: Distributions of model predictions for `FYA` for original (red) and counterfactual (blue) values of `race`, for $\sigma^2 = 4$. The original value for `race` is `white`, and its counterfactual value is `black`. In (a), we see complete overlap of the original and counterfactual distributions; in (b), we see that the counterfactual distribution has much higher variance than the original distribution.

This is again corroborated in Figure 4, where we see the distances between the original and counterfactual distributions of predicted `FYA`, for differential and random measurement error. With the exception of when $\sigma^2 = 1$, the original and counterfactual distributions are significantly different under differential measurement error, and not so under random measurement error.

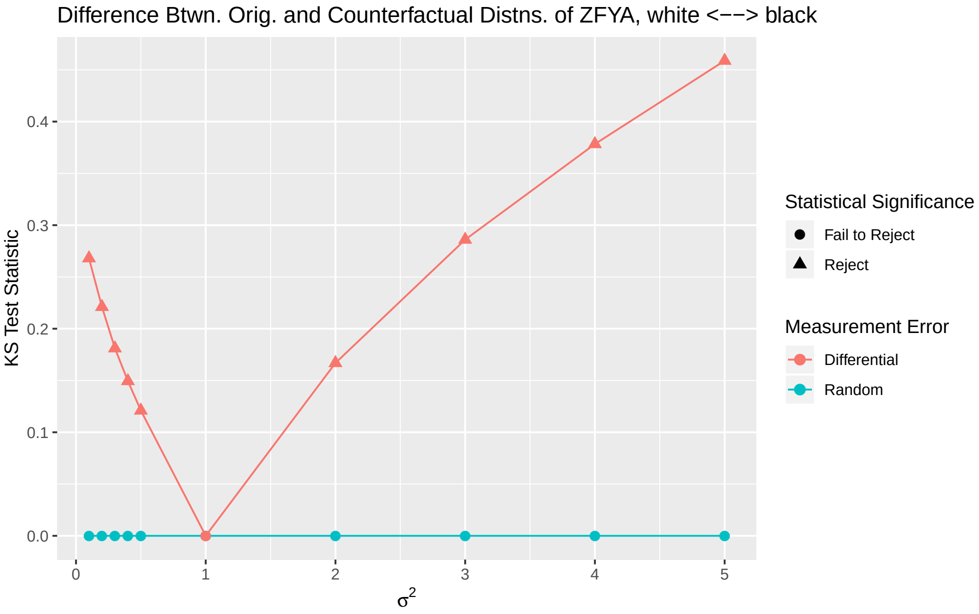
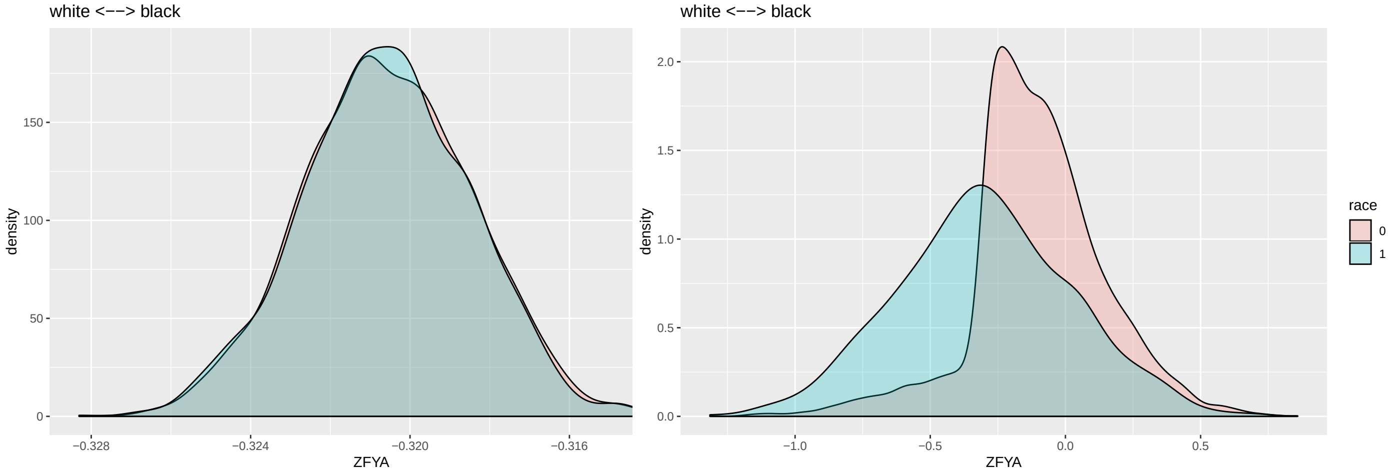


Figure 4: For all values of σ^2, the associated KS test statistic and statistical significance of said test, for the distance between the original and counterfactual distributions of predicted FYA. Note that when $\sigma^2 = 1$, the standard deviation of ϵ for both original and counterfactual values of race is the same, hence the equivalent test statistic. Otherwise, for all values of σ^2, differential measurement error causes statistically-significant differences between the original and counterfactual distributions; random measurement error does not.



(a) Selection probability of 0.5

(b) Selection probability of 0.9

Figure 5: Distributions of model predictions for FYA for original (red) and counterfactual (blue) values of race, for selection probabilities of 0.5 and 0.9. The original value for race is white, and its counterfactual value is black. In (a), we see almost complete overlap of the original and counterfactual distributions; in (b), we see that the original distribution has increased density on more positive values of predicted FYA, compared to the counterfactual distribution.

7.2 Endogenous Selection Bias

From Figure 2, we see that the selection mechanism s d-connects race and LSAT; thus we would expect any selection mechanism with a probability other than 0.5 (random) to cause $\hat{Y}$ to violate counterfactual fairness. This reasoning is corroborated in Figures 5 and 6. We see that, for all probabilities other than 0.5, endogenous selection bias causes statistically-significant differences between the original and counterfactual distributions.

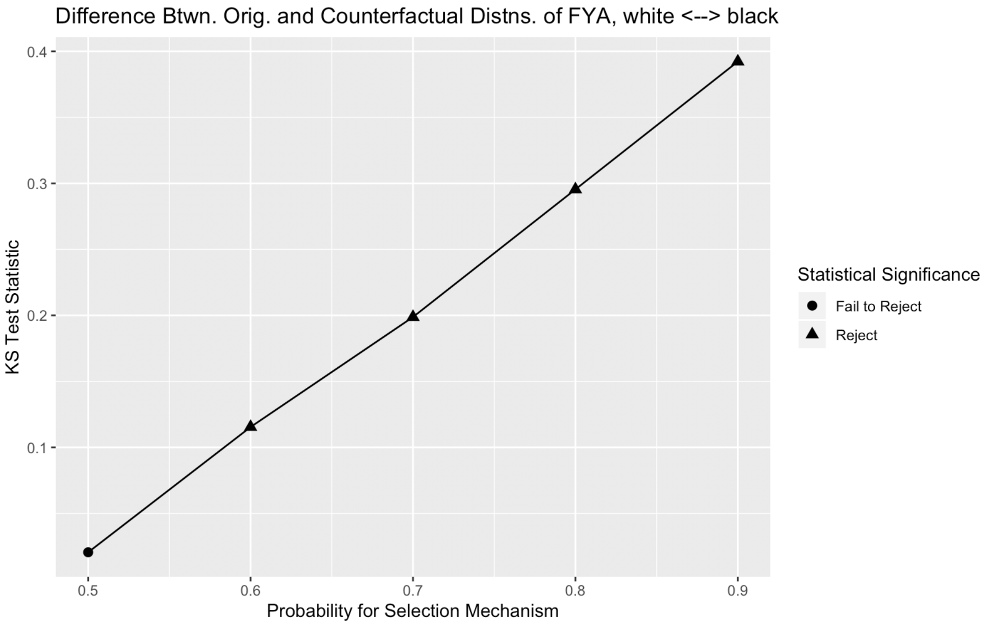


Figure 6: For all considered selection probabilities, the associated KS test statistic and statistical significance of said test, for the distance between the original and counterfactual distributions of predicted FYA. For all probabilities other than 0.5 (random), endogenous selection bias causes statistically-significant differences between the original and counterfactual distributions.

8 Conclusions and Future Work

Causality provides an attractive framework with which we can reason about fairness and discrimination with respect to the task of fair prediction. However, the model specification assumptions implicit in these methods have not yet been rigorously studied in the literature. For one causal fairness method, counterfactual fairness, we have articulated distributional parities required in the training data and implied by its definition, and have shown theoretically when different forms of model misspecification will violate these parities, and thus violate counterfactual fairness. We have corroborated this theory with empirical results on a real dataset.

Future work may continue to interrogate the usage of causal reasoning for the task of fair prediction. There are numerous proposed definitions of algorithmic fairness that leverage concepts from causality [6] [14] - we have only discussed one usage of counterfactuals. Moreover, the task of fair prediction using causal methods can be decoupled into two tasks: (1) causal discovery to estimate the true causal model $\mathcal{M}_{DGP}$ with $\mathcal{M}_{learned}$, and (2) fair prediction using $\mathcal{M}_{learned}$. This work focuses solely on (2), but (1) is a necessary avenue of research as well.

References

[1] ALEXANDER, M. *The New Jim Crow*. New Press, 2012.

[2] BARABAS, C., VIRZA, M., DINAKAR, K., ITO, J., AND ZITTRAIN, J. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *FAT* (2018), vol. 81 of *Proceedings of Machine Learning Research*, PMLR, pp. 62–76.

[3] BAREINBOIM, E., TIAN, J., AND PEARL, J. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), AAAI'14, AAAI Press, pp. 2410–2416.

[4] BOLLEN, K. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

[5] CALMON, F., WEI, D., VINZAMURI, B., NATESAN RAMAMURTHY, K., AND VARSHNEY, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3992–4001.

[6] CHIAPPA, S., AND P. S. GILLAM, T. Path-specific counterfactual fairness. In *arXiv:1802.08139*. 2018.

[7] CORBETT-DAVIES, S., AND GOEL, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR abs/1808.00023* (2018).

[8] CREAGER, E., MADRAS, D., PITASSI, T., AND ZEMEL, R. Causal modeling for fairness in dynamical systems. In *arXiv preprint arXiv:1909.09141v1*. 2019.

[9] ELWERT, F., AND WINSHIP, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology 40*, 1 (2014), 31–53. PMID: 30111904.

[10] HERNAN, M. A., AND ROBINS, J. M. *Causal Inference*. Chapman Hall/CRC, forthcoming, 2019.

[11] HINNEFELD, J. H., MAMMO, N., COOMAN, P., AND DEESE, R. Evaluating fairness metrics in the presence of dataset bias. In *arXiv:1802.08139*. 2018.

[12] JACKSON, J. W., AND VANDERWEELE, T. J. Decomposition analysis to identify intervention targets for reducing disparities. In *arXiv:1703.05899*. 2018.

[13] KILBERTUS, N., BALL, P. J., KUSNER, M. J., WELLER, A., AND SILVA, R. The sensitivity of counterfactual fairness to unmeasured confounding. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)* (July 2019), AUAI Press, p. 213.

[14] KILBERTUS, N., ROJAS-CARULLA, M., PARASCANDOLO, G., HARDT, M., JANZING, D., AND SCHÖLKOPF, B. Avoiding discrimination through causal reasoning. In *NIPS* (2017), pp. 656–666.

[15] KUSNER, M., RUSSELL, C., LOFTUS, J., AND SILVA, R. Making decisions that reduce discriminatory impacts. In *Proceedings of the 36th International Conference on Machine Learning* (Long Beach, California, USA, 09–15 Jun 2019), K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 3591–3600.

[16] KUSNER, M. J., LOFTUS, J., RUSSELL, C., AND SILVA, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4066–4076.

[17] LIPTON, Z., AND STEINHARDT, J. Troubling trends in machine learning scholarship. *Presented at the Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning.* (2018).

[18] LUM, K. Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour 1* (June 2017), 0141.

[19] LUM, K., AND ISAAC, W. To predict and serve? *Significance* (2016), 14 – 18.

[20] MADRAS, D., CREAGER, E., PITASSI, T., AND ZEMEL, R. Fairness through causal awareness: Learning latent-variable models for biased data. In *arXiv:1809.02519*. 2018.

[21] MITCHELL, S., POTASH, E., AND BAROCAS, S. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. In *arXiv:1811.07867*. 2018.

[22] NABI, R., MALINSKY, D., AND SHPITSER, I. Learning optimal fair policies. In *arXiv:1809.02244*. 2018.

[23] PEARL, J. *Causality*. Cambridge University Press, 2009.

[24] PROPUBLICA. Compas recidivism risk score data and analysis.

[25] RUBIN, D. B. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association 103*, 484 (2008), 1350–1353.

[26] RUSSELL, C., KUSNER, M. J., LOFTUS, J., AND SILVA, R. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6414–6423.

[27] SALIMI, B., RODRIGUEZ, L., HOWE, B., AND SUCIU, D. Capuchin: Causal database repair for algorithmic fairness. *CoRR abs/1902.08283* (2019).

[28] SAMPLES, T. C., WOODS, A., DAVIS, T. A., RHODES, M., SHAHANE, A., AND KASLOW, N. J. Race of interviewer effect on disclosures of suicidal low-income african american women. *Journal of Black Psychology 40*, 1 (2014), 27–46.

[29] SCHEINES, R. An introduction to causal inference. In *Causality in Crisis? University of Notre Dame* (1997), Press, pp. 185–200.

[30] SCUTARI, M. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software 35*, 3 (2010), 1–22.

[31] VANDERWEELE, T. J., AND ROBINSON, W. R. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. 473–484. Exported from https://app.dimensions.ai on 2019/05/13.

[32] WIGHTMAN, L. F. Lsac national longitudinal bar passage study. *LSAC Research Report Series* (1998).

[33] YEOM, S., AND TSCHANTZ, M. C. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. *CoRR abs/1808.08619* (2018).

[34] ZENNARO, F. M., AND IVANOVSKA, M. Pooling of causal models under counterfactual fairness via causal judgement aggregation. *CoRR abs/1805.09866* (2018).

[35] ZHANG, J., AND BAREINBOIM, E. Fairness in decision-making - the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018* (2018).