
PROMISES AND CHALLENGES OF CAUSALITY FOR ETHICAL MACHINE LEARNING

A PREPRINT

Aida Rahmattalabi*
University of Southern California
rahmatta@usc.edu

Alice Xiang
Sony AI
alice.xiang@sony.com

January 27, 2022

ABSTRACT

In recent years, there has been increasing interest in using causal reasoning for designing fair decision-making systems due to its compatibility with legal frameworks, interpretability for human stakeholders, and robustness to spurious correlations inherent in observational data, among other factors. This recent attention to causal fairness, however, has been accompanied with great skepticism due to the practical and epistemological challenges with applying current causal fairness approaches proposed in the literature. Motivated by the long-standing empirical work on causality in econometrics, social sciences, and biomedical sciences, in this paper we lay out the conditions for appropriate application of causal fairness under the “potential outcomes framework.” We highlight key aspects of causal inference that are often ignored in the causal fairness literature. In particular, we discuss the importance of specifying the nature and timing of proposed hypothetical interventions on social categories such as race or gender. Precisely, instead of postulating an intervention on *immutable attributes*, we propose a shift in focus to their *perceptions* and discuss the implications for fairness evaluation. We argue that such conceptualization of the hypothetical intervention is key in evaluating the validity of the causal assumptions and conducting sound causal analysis including avoiding post-treatment bias. Subsequently, we illustrate how causality can address the limitations of existing fairness metrics, including those that depend upon statistical correlations. Specifically, we introduce causal variants of common statistical notions of fairness, and we make a novel observation that under the causal framework there is no fundamental disagreement between different notions of fairness. Finally, we conduct extensive experiments where we demonstrate our approach for evaluating and mitigating unfairness, specially when post-treatment variables are present.

Keywords Fairness in Machine Learning · Causal Inference · Algorithmic Bias · Human-centered AI

1 Introduction

Recently, there has been a growing interest in applying causality for unfairness evaluation and mitigation [27, 24, 34, 7]. Causality provides a conceptual and technical framework for addressing questions about the effect of (hypothetical) interventions on, in this context, sensitive attributes such as race, gender, etc. This is in contrast with fairness criteria that merely rely on passive observations [5, 20, 13, 49, 25]. Observational criteria achieve fairness by constraining the relationships between variables, often in conflicting ways. Consequently, it has been shown that it is impossible to satisfy these criteria simultaneously on a dataset [25, 8, 44]. Causality helps unify these different perspectives by shifting the focus from association to causation in order to identify and mitigate the *sources of disparity*. This perspective is also more compatible with legal requirements of evaluating algorithmic bias discussed in earlier work [47].

Nevertheless, causal fairness too has been subject to criticism. One objection is around the validity of the assumptions in causal modeling. The majority of recent research on causal fairness has focused on structural causal models, which encode the relationships between variables via a Directed Acyclic Graph (DAG) [27, 34, 7]. In realistic settings, however,

*This work was conducted while the first author was an intern at Sony AI.

constructing the DAG model is a challenging task. In particular, it is generally difficult to come up with arguments for the absence of links without conducting controlled experiments [17]. Causal discovery from observational data also relies on strong untestable assumptions or do not generally pin down all possible causal details in a DAG [45, 29].

There are also concerns about considering categories such as race or gender as a cause [30, 26, 15, 22]. From one perspective, most of these attributes are determined at the time of an individual’s conception and are modeled as source nodes in a causal graph which can directly or indirectly influence the descendent variables. This view raises several major issues. Through such conceptualization, in order to evaluate and mitigate unfairness, one is inevitably required to identify all possible pathways through which sensitive attributes influence an outcome. In addition to the modelling challenge this view poses, in practice a single entity may not be held liable for the discrimination across an entire causal pathway. In this regard, many anti-discrimination mechanisms investigate whether a *particular person or institutional actor* has behaved in a discriminatory manner. For example, in the employment setting, a racial discrimination lawsuit aims to determine whether a firm has withheld some benefits such as hiring with regard to racial identity of the applicant. However, disparities in hiring rates for different groups might be a reflection of either discrimination or differences in the applicant pool’s qualifications. For example, if past discrimination in the educational system has led to some applicants having lower educational achievements, by hiring based on educational achievements, the employer will perpetuate the effects of this discrimination. Under anti-discrimination law, however, as long as the employer makes the hiring decision based on educational achievements that are legitimately connected with the job and business needs—with no regard to race—no liability is attached. In fact, if the employer seeks to proactively address past societal discrimination, this could lead to reverse discrimination lawsuits [31]. Another issue is post-treatment bias, which arises when one controls for post-treatment variables, resulting in biased estimates of the treatment effect [38]. Since some attributes such as race, gender, etc. are fixed at the time of one’s conception, almost all measurable variables become post-treatment. Hence, conditioning on those variables may lead to misleading estimates of discrimination. Removing those variables, e.g., as proposed in [27, 34], leaves little to no information for valid causal analysis.

Alternatively, many view attributes such as race or gender as social constructs that evolve over the course an individual’s life. Recently, [15, 22] studied epistemological and ontological aspects of counterfactuals in the context of fairness evaluation. In [15], the authors argue that social categories such as race may not admit counterfactual manipulation. In [22], the authors aim to address this problem by proposing a set of tenets which require a decision-maker to state implicit and unspecified assumptions about social ontology as explicitly as possible. Despite recent efforts, there has been limited empirical investigation on how the nature of the intervention impacts the scope and validity of causal analysis of sensitive attributes and conclusions one draws.

In this work, we investigate the practical and epistemological challenges of applying causality for fairness evaluation. In particular, we highlight two key aspects that are often ignored in the current causal fairness literature: *nature* and *timing* of the interventions on social categories such as race, gender, etc. Further, we discuss the impact of this specification on the plausibility of causal assumptions. To facilitate this discussion, we draw a distinction between intervening on immutable attributes and their perception, and demonstrate how such conceptualization allows us to disentangle the potential unfairness along causal pathways and attribute it to the respective actors. The idea that perceptions matter and can be manipulated is not new. For example, researchers have examined the effect of manipulated names associated with political speeches [42] and resumes [3]. Nevertheless, in the machine learning literature, little attention has been paid to the consequences for valid causal inference for unfairness evaluation and mitigation. We make the following contributions:

- We propose a causal framework to investigate and mitigate unfairness of a particular actor’s behavior, along a causal pathway. To the best of our knowledge, no prior work has aimed to isolate such effects for fair prediction. To tackle this problem, we highlight the importance of identifying the timing and nature of the intervention on social categories and its impact on conducting valid causal analysis including avoiding post-treatment bias.
- We illustrate how causality can address the limitations of existing fairness criteria, including those that depend upon statistical correlations. In particular, we introduce the causal variants of the popular statistical criteria for fairness and we make a novel observation that under the causal framework there is indeed no fundamental disagreement between different fairness definitions.
- We conduct extensive experiments where we demonstrate the effectiveness of our methodology for unfairness evaluation and mitigation compared to common baselines. Our results indicate that the causal framework is able to effectively identify and remove disparities at various stages of decision-making.

2 Related Work

There are two main frameworks for causal inference: structural causal models [14], also referred to as DAGs, and the potential outcomes framework (POF) [40]. DAGs can be viewed as a sequence of steps for generating a distribution

from independent noise variables. Causal queries are performed by changing the value of a variable and propagating its effect through the DAG [14]. POF, on the other hand, starts by defining the counterfactuals with reference to an *intervention* and postulates potential outcomes under different interventions, albeit some unobserved. In general, DAGs encode more assumptions about the relationships of the variables; i.e., one can derive potential outcomes from a DAG, but potential outcomes alone are not sufficient to construct the DAG. Consequently, POF has been more widely adopted in empirical research, including bias evaluation outside of ML [3, 46]. More detailed discussion on the differences between the two frameworks in relation to empirical research can be found in [17]. Causal inference on immutable attributes has appeared in several works including [46, 24] via proxy variables and [12] through the perception of an immutable attribute. In this work, we follow the footsteps of [12] and provide a rigorous framework to reason about the causal effect of immutable attributes which helps avoid some of the common issues in causal inference including post-treatment bias.

Recently, there has been much interest in causality in the machine learning community, where the majority of works have adopted the DAG framework [27, 24, 51, 50, 28, 7] with a few exceptions that rely on POF [34, 23]. Specifically, [27] provides an individual-based causal fairness definition that renders a decision fair towards an individual if it is the same in the actual world and a counterfactual world where the individual possessed a different sensitive attribute. In [24], the authors propose *proxy discrimination* as (indirect) discrimination via proxy variables such as name, visual features, and language which are more amenable to manipulation. Additionally, [34, 7] study path-specific discrimination, where the former proposes to remove the descendants of the protected attribute under the unfair pathway and the latter aims to correct the those variables. In [23], the authors propose two causal definitions of group fairness: fair on average causal effect (FACE), and fair on average causal effect on the treated (FACT) and show how these quantities can be estimated for specific attributes such as race or gender as the treatment. The authors restrict their attention to the fairness evaluation task and do not discuss the distinction between pre- and post-treatment variables. Further, [51, 50] discusses counterfactual direct, indirect, and spurious effects and provides formulas to identify these quantities from observational data. These works rely on a causal model, or DAG, and develop different methodologies to identify and mitigate unfairness. However, a clear discussion of the causal assumptions is typically missing, which consequently hinders the adoption of these methods in practice. In addition, the validity of the causal assumptions are influenced by the nature of the postulated intervention and its timing, which is not clearly articulated in the current literature. In many applications, discrimination by specific individuals or institutional actors is the subject of a study not an entire causal pathway. Our work makes this distinction and discusses the importance of specifying the timing and nature of a hypothetical intervention to conduct such analyses.

Finally, we briefly review the observational notions of fairness. Demographic parity and its variants have been studied in numerous papers [10, 11, 9]. Also referred to as statistical parity, this fairness criteria requires the average outcome to be the same across different sensitive groups. Conditional statistical parity [11, 9] imposes a similar requirement after conditioning on a set of legitimate factors. In the classification setting, equalized odds and a relaxed variant, equality of opportunity, have been proposed [13] to measure the disparities in the error rate across different sensitive groups. The aforementioned criteria can be expressed using probability statements involving the observed random variables at hand, hence the name observational. These criteria are often easy to state and interpret. However, they suffer from a major limitation: it is impossible to simultaneously achieve these criteria on any particular dataset [25, 8, 44]. In this work, we revisit these notions and introduce their causal variants, where we show that under the causal framework, there is no fundamental disagreement between different criteria.

3 Causal Fairness: A Potential Outcomes Perspective

We consider a decision-making scenario where $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ is the available set of attributes for an individual which we aim to use in order to make a (discrete) decision $Y \in \{0, 1\}$. An individual is further characterized by a sensitive attribute $A \in \{0, 1\}$ for which fair treatment is important. We assume A is a single binary variable, however, our discussion can naturally be extended to cases where A has more than two levels. It also applies when there is more than one sensitive attribute, such as the intersection of race and gender, by considering their joint values. Causal fairness views the unfairness evaluation and mitigation problem as a counterfactual inference problem. For example, we aim to answer questions of type: *What would have been the hiring decision, if the person had been perceived to be of a different gender?* or *Would the person have been arrested if they had been perceived to be a different race?* Such causal criteria are centered around the notion of an *intervention* or *treatment* on social categories such as gender and race. Formally, we build on POF [40] and define $Y(A)$, $A \in \{0, 1\}$ as random variables describing the potential outcomes under different values of A , i.e., the outcome after we manipulate one’s sensitive attribute A (its perception). It is important to note that for any individual, only one of the values of $Y(A)$ is observed which is the outcome corresponding to the possessed value of A . Other outcomes are considered as counterfactual quantities and are treated as unobservable variables.

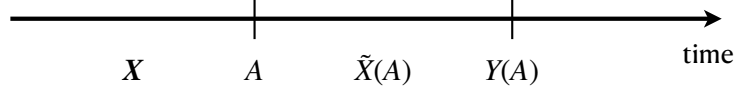


Figure 1: Decision-making timeline: the time when one’s sensitive attribute A is perceived determines pre- and post-treatment variables. Here, X is the vector of pre-treatment variables, \tilde{X} is a post-treatment variable and Y is the outcome or decision.

In this work, we take a decision-maker’s perspective considering how their perception of one’s sensitive attribute may lead to different decisions. Through this conceptualization, it is possible for discrimination to operate not just at one point in time and within one particular domain but at various points within and across multiple domains throughout the course of an individual’s life. For example, in the context of racial discrimination, earlier work has recognized potential points of discrimination across different domains including labor market, education, etc. [36]. Consequently, we need to specify the point in time at which we wish to measure and mitigate unfairness. In causal terms, this is closely related to the notion of timing of the intervention, i.e., the time at which one’s sensitive attribute is perceived by an actor. To illustrate, consider a hiring scenario and suppose we are interested in evaluating whether the hiring decision is fair with respect to gender or not. We can investigate unfairness at different stages, e.g., from the first time an individual comes into contact with the company (e.g., resume review), progresses in the system (during interviews), or when the final decision is being made. We may even take a much broader perspective and investigate the effect of gender from the point an individual attends college and study how gender affects education and subsequently the opportunities in the job market. Indeed, as we expand our view the causal inference problem we are faced with becomes increasingly more challenging but the conceptual framework remains valid.

Both timing and nature of the intervention impact the conclusions we draw. For example, under an unfair educational system, a hiring decision that is based on educational achievements will perpetuate those biases, even if it treats individuals fairly given their educational background. Similarly, a discriminatory interview process will result in an unfair hiring decision. However, the difference is that in the latter, the company is now liable for the discriminatory behavior as it stems from the a point in its decision-making process. The timing of the intervention is thus important in conducting causal analysis. In particular, consider an interview process which is discriminatory, resulting in unfair interview scores for a particular group. In our fairness evaluation, if we control for the interview score, we will find no relationship between gender and hiring decision contrary to our intuition that the decision-maker discriminates between female and male candidates through the interview score. This observation is due to post-treatment bias cautioned in the causal inference literature which happens when variables that are fixed after the intervention are used in evaluating the treatment effect [33]. Figure 1 demonstrates this over a decision-making timeline. After we fix the point of (hypothetical) intervention on A , variables $\tilde{X} \in \tilde{\mathcal{X}} \subseteq \mathbb{R}^m$ determined afterwards are considered as post-treatment variables and in principle are affected by A . Hence, we introduce the counterfactual values of $\tilde{X}(0)$, $\tilde{X}(1)$ to differentiate between pre-treatment and post-treatment variables. Consequently, the observed values of post-treatment variables are determined as $\tilde{X} = \tilde{X}(0)(1 - A) + \tilde{X}(1)A$.

Furthermore, the nature of the intervention influences the causal effect that we are able to uncover. For instance, in the study conducted in [3], the authors manipulated the names on the resumes to measure racial discrimination which only allowed them to capture the level of discrimination exhibited through the relationship between one’s name and perception of race. Under a different manipulation, e.g., zip code of the applicant, the outcome of the study would have been different. In observational studies, where the analyst has no control over how an individual’s sensitive attribute is perceived, a careful examination of mechanisms through which one’s attributes are perceived is necessary. Indeed, it is possible to identify several mechanisms affecting perceived attributes (e.g., name, clothing, language, etc.). In this case, it is possible to study the joint effect of the mechanisms by modeling the missing counterfactual values, under each mechanism, as random variables with a distribution. The distribution for each individual’s missing counterfactual value can then be represented by a stochastic mixture of distributions associated with each mechanism [12].

Building on the above discussion, we define fairness in terms of the treatment effect of a *specific intervention* on perceived sensitive attribute at a *particular point in time*. We refer to this notion as causal parity and under the POF, we can express it mathematically via the following definition.

Definition 1 (Causal Parity). *A decision-making process achieves causal parity if $\mathbb{E}[Y(1) - Y(0)] = 0$.*

In the above definition, $\tau = \mathbb{E}[Y(1) - Y(0)]$ is the treatment effect of A on Y . As stated earlier, both potential outcomes $Y(0)$, $Y(1)$ are not simultaneously observed for any individual. In order to conduct meaningful causal inference to

identify the treatment effects several assumptions are necessary. We review the assumptions and discuss how the precise specification of the intervention helps establish their plausibility.

Causal Assumptions for Identification

Assumption 1. *There is a set of established conditions under which causal inference becomes possible:*

- *Stable Unit Treatment Value Assumption (SUTVA): It states the treatment that one unit (individual) receives does not change the potential outcomes of other units.*
- *Consistency: Formally, $Y = Y(0)(1 - A) + Y(1)A$. In words, Y agrees with the potential outcome under the respective treatment. The implication of this assumption is that there are no two “flavors” or versions of treatment such that $A = 1$ under both versions but the potential outcome for Y would be different under the alternative versions.*
- *Positivity: At each level of pre-treatment variables \mathbf{X} , the probability of receiving any form of treatment is strictly positive. Mathematically,*

$$\mathbb{P}(\mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}) > 0) = 1 \quad \forall a \in \{0, 1\}, \mathbf{x} \in \mathcal{X}.$$

- *Conditional Exchangeability: it states that those individuals receiving the treatment should be considered exchangeable (with respect to potential outcomes \mathbf{Y} and the post-treatment variable $\tilde{\mathbf{X}}$) with those not receiving the treatment and vice versa. Mathematically,*

$$\mathbf{Y}, \tilde{\mathbf{X}} \perp A \mid \mathbf{X} = \mathbf{x} \quad \forall \mathbf{x} \in \mathcal{X},$$

where $\mathbf{Y} = \{Y(0), Y(1)\}$, $\tilde{\mathbf{X}} = \{\tilde{X}(0), \tilde{X}(1)\}$ and \mathbf{X} is the vector of pre-treatment variables.

Earlier works have emphasized the criticality of these assumptions in determining the causal effects [41]. Here, we highlight their importance in the context of fairness evaluation. SUTVA can also be viewed as a non-interference assumption and depends very much on the problem under study and the choice of the decision-maker. For example, for a recruiter as the decider, one should think carefully whether the recruiter’s decision to proceed with an application is independent from case to case. If a recruiter screened three candidates in a row with exceptional resumes, they might raise their standards when judging the fourth resume. In this case, SUTVA is violated as historical data on other candidates influences the future candidates outcomes.

The consistency assumption means, for example, that for candidates perceived as either male or female, an employer would not base the hiring decision on the level of “manliness.” Similarly, the degree of “blackness” of an individual should not affect the decision made for an individual. This assumption, however, can be potentially relaxed with information beyond what is typically assumed. For example, if an accurate estimate of the level of “manliness” or skin color were recorded, then the treatment could be conceptualized as having multiple levels [12]. Consistency can also be viewed as treatment invariance, which we discussed in the previous section in the context of nature of intervention on social categories. When intervening on social categories such as race, it is possible that different factors contribute to the perception of one’s sensitive attribute. Under consistency, one needs to make sure that there is sufficient data in order to capture the different levels of “race.” Without such nuanced data, it is still possible to measure the causal effect, but the interpretation changes, as the estimated causal effect is an average of multiple potential treatments.

The positivity assumption is also essential in order to identify the treatment effect. It requires that there is not a complete overlap between the treatment assignment and pre-treatment variables. For example, if all of the women in a hiring pool have a PhD, and all of the men only have a Master’s degree, then it is not possible to separate the effect of gender discrimination from the effect of the educational attainment on the employment decision. Positivity is often easy to verify from the data once the pre-treatment variables \mathbf{X} are determined.

Conditional exchangeability is one of the cornerstone assumptions for causal inference, which is in principle impossible to verify in observational studies. Conditional exchangeability in experimental settings can be obtained through stratified randomization. In order to increase the plausibility of this assumption in observational contexts, analysts typically include as many pre-treatment variables as possible to ensure that as many confounders as possible between treatment and outcome are accounted for. Intuitively, the goal is to ensure that once all of the pre-treatment variables \mathbf{X} are controlled for, the allocation of individuals between treatment and control is as close to random as possible. In the fairness setting, this would mean that, after controlling for \mathbf{X} , the only systematic difference between the two groups is the perception of their protected attribute (i.e., whether they were discriminated against), allowing for an empirical estimate of the effect of discrimination. We note that in the exchangeability assumption, we have the conditional

independence of the counterfactuals of both \tilde{X} and Y . This is a key distinction with earlier work [23] that does not differentiate between pre- and post-treatment variables.

In more complicated settings, where an individual interacts with multiple parts of a system, we may have more than one choice of decision-maker to study. In such situations, an analyst may have to balance the need to make the exchangeability assumption plausible against the desire to study a decision-maker’s behavior early in the decision-making chain. Choosing the timing of the intervention towards the later interactions renders more measured variables pre-treatment which in turn can make the exchangeability assumption more plausible. However, by treating such variables as pre-treatment and thus conditioning on them in the analysis, the analyst forgoes the detection of any prior discrimination that may have affected the values of these variables. In cases where there is sufficient data to detect discrimination starting from earlier stages of decision-making, it may be still important to pin down the different sources of discrimination throughout the decision-making process. For example, in the hiring context, suppose from the onset (the first interaction of the applicant with the company), a rich set of data about the applicant’s background and qualifications is collected that allows an analyst to determine the hiring process is unfair towards a group. In such a case, it is important to understand whether discrimination is attributed to the recruitment process, the interview stage or the final hiring process. Additionally, there may be a long delay between the time of perceiving an individual’s sensitive attribute and outcome. In this case, it may be helpful to use post-treatment variables to improve the precision [1].

Fairness Evaluation

So far, we have examined the causal assumptions and their implications in the context of fairness evaluation. Once the plausibility of the assumptions are established, we can proceed to estimate the treatment effect of A on Y . While there are many approaches in estimating the causal effect, we mainly focus on direct regression method. We first consider a case where post-treatment variables are absent. Under causal assumptions, treatment effect of A can be formulated as

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}]] = \mathbb{E}[\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, A = 1] - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, A = 0]], \quad (1)$$

which can be estimated from observational data via two separate regression models. When post-treatment variables are present, it may be helpful to use them in order to improve the precision of treatment effect estimates. In this case, simply conditioning on those variables will introduce bias in the analysis. Instead, we should treat them as dependants on A . In order to emphasize the causal effect of post-treatment variables on the potential outcomes, we consider potential outcomes $Y(A, \tilde{X}(A))$ that are indexed by both the treatment and the post-treatment counterfactuals. We estimate the treatment effect of A on Y is given as $\tau = \mathbb{E}[Y(1, \tilde{X}(1)) - Y(0, \tilde{X}(0))]$. In the mediation literature, this quantity is known as *total effect* [16]. Estimating the total effect poses a considerable identification challenge as it depends on four ($\tilde{X}(0)$, $\tilde{X}(1)$, $Y(0, \tilde{X}(0))$, $Y(1, \tilde{X}(1))$) counterfactuals which are not simultaneously observed for any individual. To tackle this problem, we propose to use imputation [39] which is commonly used in causal inference literature to assign values to unobserved variables in the data. Precisely, in order to attain the causal effect of A on Y , we sequentially impute the missing variables were conditional on the previous step. Precisely, we first impute the counterfactual post-treatment variables \tilde{X} as a function of the pre-treatment variables \mathbf{X} and A . Next, we impute unobserved $Y(A, \tilde{X}(A))$ values as a function of pre-treatment variables \mathbf{X} , post-treatment counterfactuals \tilde{X} and A . Similar sequential imputation techniques have been used in causal inference literature in order to evaluate the long-term impact of policy shifts [48].

Unfairness Mitigation

In the previous section, we focused solely on fairness evaluation which we formulated as a causal inference problem on the effect of A on Y . Here, we discuss how we can mitigate unfairness if the treatment effect of A on Y is non-zero. Similar to the previous section, we distinguish between pre- and post-treatment variables as the post-treatment variables are affected by A . The core idea of our unfairness mitigation approach is to adjust the post-treatment and outcome variables to achieve $\tau = 0$. The idea of adjusting downstream variables, affected by sensitive attributes, has been recently investigated in the fair ML literature and in the context for mitigating path-specific effects under the DAG framework [7]. In this work, we are interested in mitigating unfairness that attributed to a specific actor’s decision-making process, rather than an entire causal path. Intuitively, our approach is based on the assumption that in a fair world, everyone is treated with no regard to their group membership. In other words, we deem a decision-making process fair if everyone is treated as if they belong to the same group, which we refer to as a baseline group. The baseline group can be viewed as either the majority group or a historically advantaged group.

We first consider a setting with no post-treatment variables and assume $\mathbb{E}[Y(1) - Y(0)] \neq 0$. Let $A = 0$ be the baseline group. If we had access to $Y(0)$ for every individual in the population, we could use that in order to learn a fair classifier. That is, if we observed the outcome of individuals had they belonged to the baseline group, we could

use this data to learn a predictive model. The reason is that when we use $Y(0)$ to learn a model, we are effectively eliminating decision-maker’s unfavorable attitude towards membership to group $A = 1$. Consequently, we can model the prediction problem as $\mathbb{P}(Y(0) = 1 \mid \mathbf{X} = \mathbf{x})$. In the presence of post-treatment variables, we employ a similar approach in order to eliminate the discriminatory effects of A . Therefore, we formulate the prediction problem as $\mathbb{P}(Y(0) = 1 \mid \mathbf{X} = \mathbf{x}, \tilde{X}(0) = \tilde{x})$, in which we use $\tilde{X}(0)$ which is the value of the post-treatment variable had the individual belonged to group $A = 0$. Remarkably, under this formalization causal parity is automatically achieved as $\mathbb{E}[Y(0) - Y(1)] = 0$. In words, we are practically assuming that an individual’s potential outcomes for an individual is the same and is equal to $A = 0$, regardless of the observed value of A . A key challenge with this approach is that $Y(0)$ values are not observed for every individual. We leverage imputation from causal inference literature to tackle this problem [39].

4 Trade-offs under the Lens of Causality

We now turn to another important aspect of our analysis. We introduce causal variants of common statistical criteria of fairness to study their behavior under the causal lens.

Causal Fairness Definitions

We center our discussion on the criteria with known impossibility results in the fair ML literature.

Definition 2 (Conditional Causal Parity). *A decision-making process achieves conditional causal parity if*

$$\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}] = 0 \forall \mathbf{x} \in \mathcal{X}.$$

The above definition is closely related to conditional statistical parity which aims to evaluate fairness after controlling for a limited set of “legitimate” factors [20]. The set of legitimate factors significantly impacts the conclusions we draw. However, it is typically assumed as given, e.g., by domain experts. In contrast, in our definition \mathbf{X} collects all the pre-treatment variables. Hence, once the nature of the intervention is explicitly defined, all remaining pre-treatment variables can be considered as legitimate since the main effect we aim to identify is the effect of the treatment.

Definition 3 (Causal Equalized Odds). *A predictor \hat{Y} satisfies causal equalized odds if:*

$$\begin{aligned} \mathbb{P}(\hat{Y} = 1 \mid Y(0) = 1) &= \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 1) \\ \mathbb{P}(\hat{Y} = 1 \mid Y(0) = 0) &= \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 0) \end{aligned}$$

The above definition is the causal counterpart of equalized odds proposed in [13]. It states that the probability of receiving a positive prediction $\hat{Y} = 1$ in worlds where everyone is treated as $A = 0$ or $A = 1$ should be the same. Therefore, an individual does not have any preferences to be in either of these worlds since in either world the prediction is the same. Next, we define the causal variant of calibration [25]. Calibration is defined in the context of risk scores.

Definition 4 (Causal Calibration). *Let $S \in \mathcal{S}$ denote a random variable encoding an individual’s risk score. The risk assignment is well-calibrated within groups if it satisfies the following condition:*

$$\mathbb{P}(Y(0) = 1 \mid S = s) = \mathbb{P}(Y(1) = 1 \mid S = s) \forall s \in \mathcal{S}.$$

Causal calibration states that a risk score S should have the same meaning in worlds where everyone is treated as $A = 0$ or $A = 1$, i.e., the proportion of positive outcomes in either worlds should be the same for any fixed $S = s$. Subsequently, we can define causal positive predictive parity.

Definition 5 (Causal Positive Predictive Parity). *A predictor \hat{Y} satisfies causal positive predictive parity if:*

$$\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1) = \mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1).$$

Causal predictive parity is applicable in the binary decision-making scenarios and has a similar interpretation as causal calibration in that it requires the proportion of positive outcomes in worlds with $A = 0$ and $A = 1$ to be the same for any fixed $\hat{Y} = 1$. Therefore, an individual with positive prediction does not feel being discriminated against since in both worlds, the rate of positive outcome is the same.

Trade-offs among Causal Criteria of Fairness

We investigate two main impossibility results known for the statistical fairness criteria and show that there is no fundamental disagreement between their causal variants.

Causal Parity and Conditional Causal Parity. It is easy to see that statistical parity and conditional statistical parity may not be satisfied simultaneously on a dataset. The Berkeley college admission study is a notorious example [4]. In this study, it was shown that while female students were admitted at a lower rate compared to the male students, after controlling for department choice, the difference in admission rates became insignificant among the two groups. This observation can be expressed formally as below.

Observation 1. *There exists a joint distribution $p(\mathbf{X}, A, Y)$ such that conditional statistical parity does not imply statistical parity, i.e.,*

$$\mathbb{E}[Y \mid \mathbf{X} = x, A = 1] - \mathbb{E}[Y \mid \mathbf{X} = x, A = 0] = 0 \forall x \in \mathcal{X} \not\Rightarrow \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] = 0.$$

Contrary to the above result, it is straightforward to show that conditional causal parity implies causal parity.

Proposition 1. *Conditional causal parity implies causal parity. Mathematically,*

$$\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = x] = 0 \forall x \in \mathcal{X} \implies \mathbb{E}[Y(1) - Y(0)] = 0.$$

Proof. The proof follows simply from taking the expectation over \mathbf{X} . □

The intuition behind the above result is that $\mathbb{E}[Y \mid A]$ merely measures the statistical dependence between Y and A and does not differentiate between different sources of dependence, e.g., female students applying for more competitive departments than male students, or a discriminatory admission process. We note that conditional causal parity is a more stringent requirement than causal parity and the reverse implication does not generally hold true in Proposition 1.

Causal Positive Predictive Parity and Causal Equalized Odds. It is well-known that one can not achieve positive predictive parity or calibration together with equalized odds simultaneously unless either the base rates $\mathbb{P}(Y \mid A = a)$ are equal or the classifier is perfect [25, 8]. Here, we show no such restrictions are necessary for their causal variants. We first define $f : \mathcal{S} \rightarrow \{0, 1\}$ as a mapping from the risk score S to binary prediction \hat{Y} . For example, $f(S) = \mathbb{I}(S > \theta)$ classifying the data points based on a threshold. We now present our main results.

Theorem 1. *Causal calibration implies causal parity and causal equalized odds.*

Proof. First, by taking the expectation over $s \in \mathcal{S}$ it is straightforward to show that causal calibration implies causal parity. Next, we apply Bayes theorem on the causal calibration definition. It follows that:

$$\begin{aligned} \mathbb{P}(S = s \mid Y(0) = 1)\mathbb{P}(Y(0) = 1) &= \mathbb{P}(S = s \mid Y(1) = 1)\mathbb{P}(Y(1) = 1) & \forall s \in \mathcal{S} \\ \Rightarrow \mathbb{P}(S = s \mid Y(0) = 1) &= \mathbb{P}(S = s \mid Y(1) = 1) & \forall s \in \mathcal{S} \\ \Rightarrow \mathbb{P}(\hat{Y} = f(s) \mid Y(0) = 1) &= \mathbb{P}(\hat{Y} = f(s) \mid Y(1) = 1) & \forall s \in \mathcal{S}. \end{aligned}$$

Similarly, we can show:

$$\mathbb{P}(Y(0) = 0 \mid S = s) = \mathbb{P}(Y(1) = 0 \mid S = s) \forall s \in \mathcal{S} \Rightarrow \mathbb{P}(\hat{Y} = f(s) \mid Y(0) = 0) = \mathbb{P}(\hat{Y} = f(s) \mid Y(1) = 0).$$

□

Causal parity, i.e., $\mathbb{P}(Y(0) = 1) = \mathbb{P}(Y(1) = 1)$, is satisfied if decisions are made regardless of one's group membership and is different from the equal base rate assumption which does not necessarily hold in many applications. The above result shows that there is an inherent compatibility between different causal fairness criteria as achieving one automatically implies one or two other criteria.

Lemma 1. *If $A/B = C/D$ and $(1 - A)/(1 - B) = (1 - C)/(1 - D)$, then $A = C$ and $B = D$.*

Proof.

$$\left. \begin{aligned} \frac{A}{B} = \frac{C}{D} &\Rightarrow \frac{A-B}{B} = \frac{C-D}{D} \\ \frac{1-A}{1-B} = \frac{1-C}{1-D} &\Rightarrow \frac{A-B}{1-B} = \frac{C-D}{1-D} \end{aligned} \right\} \Rightarrow \frac{1-B}{B} = \frac{1-D}{D}.$$

It follows that $A = C$ and $B = D$. □

Theorem 2. *Causal equalized odds implies causal parity and causal positive predictive parity.*

Proof.

$$\begin{aligned}\mathbb{P}(\hat{Y} = 1 \mid Y(0) = 1) &= \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 1) \Rightarrow \frac{\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1)}{\mathbb{P}(Y(0) = 1)} = \frac{\mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1)}{\mathbb{P}(Y(1) = 1)} \\ \mathbb{P}(\hat{Y} = 1 \mid Y(0) = 0) &= \mathbb{P}(\hat{Y} = 1 \mid Y(1) = 0) \Rightarrow \frac{\mathbb{P}(Y(0) = 0 \mid \hat{Y} = 1)}{\mathbb{P}(Y(0) = 0)} = \frac{\mathbb{P}(Y(1) = 0 \mid \hat{Y} = 1)}{\mathbb{P}(Y(1) = 0)}.\end{aligned}$$

From Lemma 1, it follows that $\mathbb{P}(Y(0) = 1) = \mathbb{P}(Y(1) = 1)$ and $\mathbb{P}(Y(0) = 1 \mid \hat{Y} = 1) = \mathbb{P}(Y(1) = 1 \mid \hat{Y} = 1)$, where the first and second equations correspond to causal parity and causal positive predictive parity, respectively. \square

We conclude this section by providing a complementary result that relates conditional causal parity to causal calibration and causal positive predictive parity.

Proposition 2. *Given a risk score as a function of pre-treatment variables \mathbf{X} , i.e., $S = h(\mathbf{X})$, it holds that conditional causal parity implies causal calibration and causal positive predictive parity.*

$$\begin{aligned}\mathbb{P}(Y(0) = 1 \mid \mathbf{X} = \mathbf{x}) &= \mathbb{P}(Y(1) = 1 \mid \mathbf{X} = \mathbf{x}) && \forall \mathbf{x} \in \mathcal{X} \\ \Rightarrow \mathbb{P}(Y(0) = 1 \mid \mathbf{X} \in h^{-1}(s)) &= \mathbb{P}(Y(1) = 1 \mid \mathbf{X} \in h^{-1}(s)) && \forall s \in \mathcal{S} \\ \Rightarrow \mathbb{P}(Y(0) = 1 \mid h(\mathbf{X}) = s) &= \mathbb{P}(Y(1) = 1 \mid h(\mathbf{X}) = s) && \forall s \in \mathcal{S} \\ \Rightarrow \mathbb{P}(Y(0) = 1 \mid \hat{Y} = f(s)) &= \mathbb{P}(Y(1) = 1 \mid \hat{Y} = f(s)) && \forall s \in \mathcal{S}.\end{aligned}$$

Consequently, causal parity and causal equalized odds will be satisfied. The above result implies that for a given set of pre-treatment variables \mathbf{X} , if conditional causal parity is satisfied, all other causal fairness criteria discussed in the present work will be satisfied provided that the risk score function h and the classifier f are functions of the pre-treatment variables. Conditional causal parity can be achieved using the imputation technique described in the previous section. Finally, it is important to note that the above results are based on the assumption that the joint distribution of variables is known. In practice, factors such as inadequate sample sizes, modelling choices, hyper-parameter selection, etc. can influence the performance of models across different groups. In the standard ML setting, previous work has aimed to address some of these limitations through careful model selection or additional training data collection, etc. [6].

5 Empirical Results

We consider a stylized hiring scenario to illustrate our causal unfairness evaluation and mitigation approach. Specifically, we consider a decision-making process that involves two interactions: interview and final hiring decision. We study how the timing of the intervention impacts our conclusions. We use A to represent gender, which we draw from a Bernoulli distribution $Bern(0.75)$ with the majority class being male $A = 1$. An individual's qualification is described by a random variable X drawn from a normal distribution $\mathcal{N}(2\alpha(A - 0.5), 1)$, where α controls the difference in the average qualifications between genders. Each candidate has a score S reflecting their performance during the interview. We model the score as a binary variable whose mean depends on the qualifications and possibly gender. We have $\mathbb{P}(S = 1) = \sigma(X + 2\beta(A - 0.5))$, where $\sigma(z) = 1/(1 + e^{-z})$ is the logistic function and $\beta \geq 0$ determines the level of discrimination in S , e.g., when $\beta > 0$ being a male $A = 1$ increases one's probability of receiving a higher score. Subsequently, a decision Y is made indicating whether the candidate receives an offer or not. We use the probabilistic model $\mathbb{P}(Y = 1) = \sigma(X + S + 2\gamma(A - 0.5))$, with $\gamma \geq 0$ controlling the level of discrimination in Y for a fixed X, S . The vector of potential outcomes, for both S and Y , can be obtained by substituting the respective value of A in the model. We present results across a wide range of α, β and γ values.

According to our causal framework, we need to specify the point in time from which the effect of gender needs to be assessed. There are two possibilities: after interview is conducted or before the interview (as one may be concerned about an unfair interview process). Naturally, the above choice will impact our conclusions about whether the system is fair or not. We generate 100,000 data points $(X, A, S(0), S(1), Y(0, S(0)), Y(1, S(1)))$ according to the process explained above. For post-interview fairness evaluation we can use the observed S values as the score is a pre-treatment variable. However, when evaluating fairness before the interview, the score becomes post-treatment. In order to impute the missing counterfactual score values $S(A)$, we use logistic regression to model $\mathbb{P}(S = 1 \mid \mathbf{X} = \mathbf{x}, A = a)$, $a \in \{0, 1\}$, from which we sample (10 samples). We use a second logistic regression $\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}, S(a) = s, A = a)$, $\forall a, s \in \{0, 1\}$ to impute $Y(A, S(A))$ values. This approach is based on

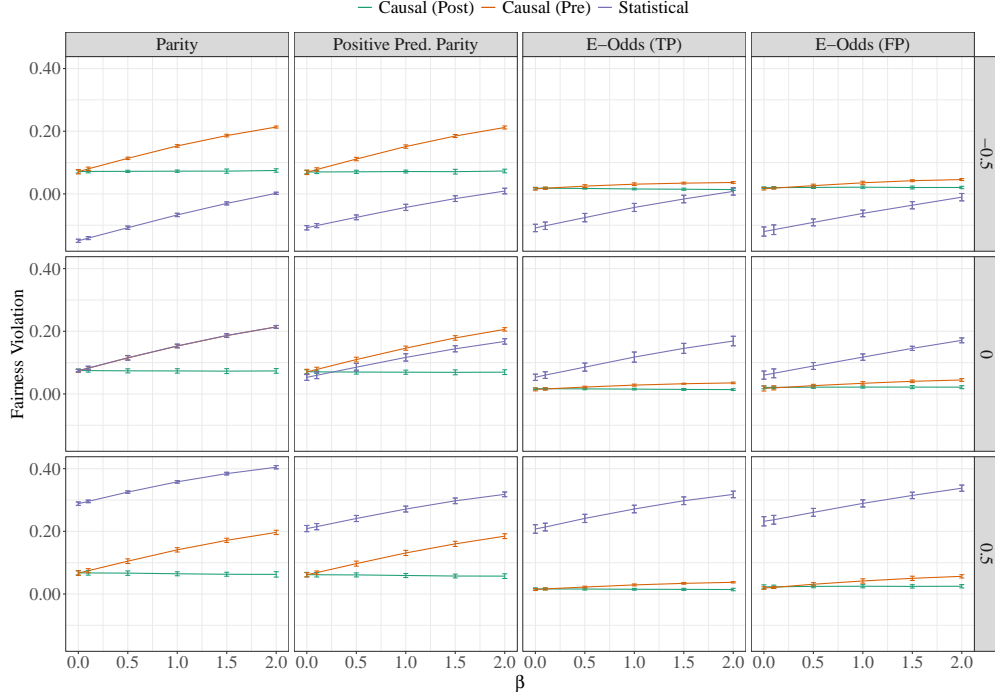


Figure 2: Synthetic results in the hiring scenario. Colors denote the evaluation method: causal pre-interview, causal post-interview and statistical. From top to bottom, each row corresponds to a different value of $\alpha \in \{-0.5, 0, 0.5\}$. Column are different fairness evaluation criteria. On the x -axis, we vary the value of β , which reflects the dependence of the interview score on one’s gender. The y -axis shows fairness violation across four different definitions. We note that for causal approaches we use the causal variants of the fairness definitions. The value of γ is set to 0.2. The error bars show 95% confidence interval. Depending on the joint setting of the parameters, statistical criteria may erroneously result in an over- or under-estimation of fairness violation. Further, post-interview fairness evaluation does not capture discrimination at earlier points in time.

multiple imputation in the causal inference literature [39]. We then use these counterfactual values in the expression that evaluates the treatment effect of A .

We compare our causal criteria against statistical fairness definitions, where we measure the fairness violation of a logistic regression model trained to predict Y using observed values of X , S and A . Figure 2 depicts a summary of results.

We can make several key observations. First, the post-interview causal plot remains flat across different values of α , β exhibiting a constant fairness violation at 0.07 due to constant $\gamma = 0.2$ which is independent of prior discrimination in the interview step. This suggests that early discrimination can not be captured when one chooses a later time as the point of fairness investigation. In other words, any unfairness in the pre-treatment variables used in the analysis will remain undetected. Pre- and post-interview lines only intersect at $\beta = 0$ and pre-interview fairness violation increases monotonically with β across all causal fairness definitions. Statistical fairness definitions exhibit significantly different results. For example, when $\alpha = -0.5$, all statistical lines lie below the causal ones which suggests that they underestimate the true level of discrimination. This is due to the fact that when $\alpha < 0$, males qualification is lower than females on average. However, since $\beta, \gamma > 0$ the interview score and the final decision are in favor of male candidates. Since the statistical criteria fail to disentangle different sources of disparities, these opposing effects are cancelled, resulting in lower estimates of unfairness. On the other hand, when $\alpha > 0$, these effects reinforce each other resulting in an over-estimation of unfairness. Only when $\alpha = 0$, do statistical parity and causal parity, in Causal (Pre), match which indicates the sensitive of statistical criteria to baseline differences between groups (average qualifications). For $\beta, \gamma = 0$ (no discrimination in interview or the hiring process), our results indicate near-zero estimates for all causal definitions of fairness across different values of α . This confirms that it is indeed possible to satisfy different causal fairness definitions simultaneously, even when there are baseline differences between the qualifications of different groups. Conversely, statistical criteria yield non-zero estimates except for the case where $\alpha, \beta, \gamma = 0$ which points to the equal base rate condition highlighted in previous work [25].

	Fairness Violation (Statistical Criteria)				
	Parity	Positive Pred. Parity	E-Odds (TP)	E-Odds (FP)	Accuracy (%)
No Fairness	0.31	0.02	0.26	0.21	72.7
Re-weighting	0.10	0.11	0.04	0.03	71.8
PrejudiceRemover	0.16	0.04	0.02	0.24	74.0
RejectOption	0.05	0.19	0.09	0.16	72.0
Causal (Pre)	0.03	0.14	-0.02	-0.04	70.0
Causal (Post)	0.17	0.07	-0.11	-0.09	72.0

Table 1: Fairness violation of statistical criteria and the classification accuracy.

Next, we study the power of our approach to mitigate unfairness. We focus on the setting where $\alpha = 0$ and $\beta, \gamma \neq 0$. This is because we aim to remove unfairness in the decision-making process, which is associated with β and γ . Since statistical approaches are not able to disentangle different sources of unfairness, by setting α we are able to compare our results against those criteria. Specifically, we train a model using our approach by imputing the missing potential outcomes. We compare the accuracy and fairness violation with an unconstrained model (No Fairness), as well as the same model after applying one of three common unfairness mitigation algorithms in the literature: pre-processing method (Re-weighting) in [18] which generates weights for the training examples in each combination of A and Y differently to ensure statistical parity, in-processing method (PrejudiceRemover) of [21] that adds a regularization term to the learning objective, and post-processing approach (RejectOption) in [19] which gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary. We rely on the implementations in AI Fairness 360 package [2]. The training model used in all of the methods is a logistic regression model. For fairness violation, we consider both statistical criteria and their causal variants.

Table 1 summarizes the statistical fairness violation results for $\beta = 1.0$ and $\gamma = 0.2$. Among all fair baselines, RejectOption and Causal (Pre) perform significantly better in terms of statistical parity. Despite the fact that other fair baselines are also designed to remove average disparities between groups, they still exhibit significant disparities. On the other hand, RejectOption performs worse than Causal (Pre) with respect to all other criteria. We also note the difference in pre- and post-interview results. The increase in parity violation in Causal (Post) can be explained by the fact that it only adjusts for the outcome variable and assumes disparities in the interview score are acceptable. Disparities in score will in turn result in different outcomes across groups but in post-interview analysis this effect remains undetected. Finally, in terms of accuracy, Causal (Post) achieves comparable results to the other methods. The difference in the accuracy of Causal (Pre) and (Post) is in part due to the fact that the accuracy is measured with respect to the observed unfair outcomes. As a result, Causal (Pre) which significantly reduces the gap between female and male candidates may not conform to the historical decisions. Finally, these results highlight the importance of determining the timing of the intervention. Specifically, they suggest that through the causal framework, we are able to identify and remove sources of disparities by actively adjusting the affected variables. Finally, we evaluated our approach based on causal criteria of fairness, choosing pre-interview as the starting time of fairness assessment. In Table 2, we observe no violation of fairness in Causal (Pre) as expected. However, Causal (post) exhibits small violations which is due to the fact that it only mitigates unfairness due to γ and in the hiring decision.

	Fairness Violation (Causal Criteria)			
	Parity	Positive Pred. Parity	E-Odds (TP)	E-Odds (FP)
Causal (Pre)	0.00	0.00	0.00	0.00
Causal (Post)	0.06	0.02	0.05	0.01

Table 2: Fairness violation of causal criteria.

6 Conclusion

As empirical evidence on ethical implications of algorithmic decision-making is growing [37, 32, 35, 43], a variety of approaches have been proposed to evaluate and minimize the harms of these algorithms. In the statistical fairness literature, it is well-established that it is not possible to satisfy every fairness criterion simultaneously, which results in significant trade-offs in selecting a metric. On the other hand, in the causal fairness literature, there is substantial ambiguity around how the proposed methods should be applied to a particular problem. Also, these methods rely on assumptions that are often too strong to be applicable in practice. In this work, we addressed some of these limitations.

First, we illustrated the utility of applying concepts from the “potential outcomes framework” to algorithmic fairness problems. In particular, we emphasized the timing and nature of the intervention as two key aspects of causal fairness

analysis. That is, for any valid causal analysis, it is critical to precisely define the starting point of the fairness evaluation and the postulated intervention. We argue that fairness evaluation is not a static problem and unfairness can happen at various points and within and across multiple domains. This is contrast with methods that rely on fixed DAG models.

Next, we demonstrated how such a causal framework can address the limitations of existing approaches. Specifically, our theoretical investigation indicates that there is an inherent compatibility between the causal fairness definitions we propose. Finally, we showed the effectiveness of our approach in evaluating and mitigating unfairness associated with different stages of decision-making. We hope that our empirical observations spark additional work on collecting new datasets that lend themselves to temporal fairness evaluation.

References

- [1] Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. 2019. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. *SSRN* (2019). <https://doi.org/10.3386/w26463>
- [2] Rachel K.E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* (2018).
- [3] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004). <https://doi.org/10.1257/0002828042002561>
- [4] P. J. Bickel, E. A. Hammel, and J. W. O’Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975). <https://doi.org/10.1126/science.187.4175.398>
- [5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDM Workshops 2009 - IEEE International Conference on Data Mining*. <https://doi.org/10.1109/ICDMW.2009.83>
- [6] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems*, Vol. 2018-December.
- [7] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. <https://doi.org/10.1609/aaai.v33i01.33017801>
- [8] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. <https://doi.org/10.1089/big.2016.0047>
- [9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. Part F129685. <https://doi.org/10.1145/3097983.3098095>
- [10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS 2012 - Innovations in Theoretical Computer Science Conference*. <https://doi.org/10.1145/2090236.2090255>
- [11] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 2015-August. <https://doi.org/10.1145/2783258.2783311>
- [12] D. James Greiner and Donald B. Rubin. 2011. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics* 93, 3 (2011). https://doi.org/10.1162/REST{_}a{_}00110
- [13] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*.
- [14] Christopher Hitchcock and Judea Pearl. 2001. Causality: Models, Reasoning and Inference. *The Philosophical Review* 110, 4 (2001). <https://doi.org/10.2307/3182612>
- [15] Lily Hu and Issa Kohler-Hausmann. 2020. What’s sex got to do with machine learning?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 513–513.
- [16] Kosuke Imai, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* 25, 1 (2010). <https://doi.org/10.1214/10-STS321>

- [17] Guido W. Imbens. 2020. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. <https://doi.org/10.1257/JEL.20191597>
- [18] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012). <https://doi.org/10.1007/s10115-011-0463-8>
- [19] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. <https://doi.org/10.1109/ICDM.2012.45>
- [20] Faisal Kamiran, Indre Žliobaite, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35, 3 (2013). <https://doi.org/10.1007/s10115-012-0584-8>
- [21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7524 LNAI. https://doi.org/10.1007/978-3-642-33486-3_3
- [22] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *FACt 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3442188.3445886>
- [23] Aria Khademi, David Foley, Sanghack Lee, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. <https://doi.org/10.1145/3308558.3313559>
- [24] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, Vol. 2017-December.
- [25] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. (2017). <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
- [26] Issa Kohler-Hausmann. 2019. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review* 113, 5 (2019).
- [27] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, Vol. 2017-December.
- [28] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3287560.3287564>
- [29] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018). <https://doi.org/10.1111/phc3.12470>
- [30] Alexandre Marcellesi. 2013. Is race a cause? *Philosophy of Science* 80, 5 (2013). <https://doi.org/10.1086/673721>
- [31] Charles E. Mitchell. 2013. An analysis of the U.S. Supreme Court’s decision in Ricci v. DeStefano: The New Haven firefighter’s case. *Public Personnel Management* 42, 1 (2013). <https://doi.org/10.1177/0091026013484574>
- [32] John Monahan and Jennifer L. Skeem. 2016. Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology* 12 (2016). <https://doi.org/10.1146/annurev-clinpsy-021815-092945>
- [33] Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres. 2018. How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science* 62, 3 (2018). <https://doi.org/10.1111/ajps.12357>
- [34] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.
- [35] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. <https://doi.org/10.1145/3287560.3287593>
- [36] Lincoln Quillian. 2006. Measuring Racial Discrimination. *Contemporary Sociology: A Journal of Reviews* 35, 1 (2006). <https://doi.org/10.1177/009430610603500165>
- [37] Lisa Rice and Deidre Swesnik. 2012. Discriminatory effects of credit scoring on communities of color.

- [38] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (4 1983), 41–55. <https://doi.org/10.1093/biomet/70.1.41>
- [39] Donald B. Rubin. 1996. Multiple Imputation after 18+ Years. *J. Amer. Statist. Assoc.* 91, 434 (1996). <https://doi.org/10.1080/01621459.1996.10476908>
- [40] Donald B Rubin. 2005. Causal Inference Using Potential Outcomes. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331. <https://doi.org/10.1198/016214504000001880>
- [41] Donald B. Rubin. 2007. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* 6, 1 (2007). <https://doi.org/10.1214/aos/1176344064>
- [42] Virginia Sapiro. 1981. If U.S. Senator Baker Were A Woman: An Experimental Study of Candidate Images. *Political Psychology* 3, 1/2 (1981). <https://doi.org/10.2307/3791285>
- [43] Tom Simonite. 2020. Meet the secret algorithm that’s keeping students out of college.
- [44] Arvind Narayanan Solon Barocas, Moritz Hardt. 2020. Fairness in Machine Learning Limitations and Opportunities. *Book* (2020).
- [45] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics* 3, 1 (2016). <https://doi.org/10.1186/s40535-016-0018-x>
- [46] Tyler J. VanderWeele and Whitney R. Robinson. 2014. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 25, 4 (2014). <https://doi.org/10.1097/EDE.000000000000105>
- [47] A. Xiang. 2021. Reconciling Legal and Technical Approaches to Algorithmic Bias. *Tennessee Law Review* 88, 3 (2021).
- [48] Alice Xiang and Donald B. Rubin. 2015. Assessing the potential impact of a nationwide class-based affirmative action system. *Statist. Sci.* 30, 3 (2015). <https://doi.org/10.1214/15-STS514>
- [49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *26th International World Wide Web Conference, WWW 2017*. <https://doi.org/10.1145/3038912.3052660>
- [50] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, Vol. 2018-December.
- [51] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making the causal explanation formula. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.