Statistical Science
Advance publication—final volume, issue, and page numbers to be assigned.
https://doi.org/10.1214/22-STS872

Principal Fairness for Human and Algorithmic Decision-Making

Kosuke Imai and Zhichao Jiang

Abstract. Using the concept of principal stratification from the causal inference literature, we introduce a new notion of fairness, called principal fairness, for human and algorithmic decision-making. Principal fairness states that one should not discriminate among individuals who would be similarly affected by the decision. Unlike the existing statistical definitions of fairness, principal fairness explicitly accounts for the fact that individuals can be impacted by the decision. This causal fairness formulation also enables online or post-hoc fairness evaluation and policy learning. We also explain how principal fairness relates to the existing causality-based fairness criteria. In contrast to the counterfactual fairness criteria, for example, principal fairness considers the effects of decision in question rather than those of protected attributes of interest. Finally, we discuss how to conduct empirical evaluation and policy learning under the proposed principal fairness criterion.

Key words and phrases: Algorithmic fairness, causal inference, potential outcomes, principal stratification.

Although the notion of fairness has long been studied, the increasing reliance on algorithmic decision-making in today's society has led to the fast-growing literature on algorithmic fairness (see, e.g., [2, 5, 10, 11, 29] and references therein). In this paper, we introduce a new definition of fairness, called *principal fairness*, for human and algorithmic decision-making. Unlike the existing *statistical fairness* criteria [9, 18, 22, 36], principal fairness incorporates causality into fairness. This causal fairness formulation also enables online or post-hoc fairness evaluation and policy learning, going beyond evaluation based on historical data for which most of the existing statistical fairness criteria are designed.

Furthermore, we explain how principal fairness relates to the existing causality-based fairness criteria. In particular, different from the *counterfactual equalized odds* criteria, principal fairness considers joint potential outcomes, and thus takes into account how the decision affects the outcome [12]. Moreover, when compared to the *counterfactual fairness* criteria [8, 28, 30, 37], principal fairness focuses on the effects of decision in question rather than those of protected attributes of interest. We characterize the formal relations between principal fairness and these other fairness criteria.

Principal fairness states that one should not discriminate among individuals who would be similarly affected by the decision. Consider a judge who decides, at a first appearance hearing, whether to detain or release an arrestee pending disposition of any criminal charges (see [20] for a related empirical study, which motivates this example). Suppose that the outcome of interest is whether the arrestee commits a new crime before the case is resolved. According to principal fairness, the judge should not discriminate between arrestees if they would behave in the same way under each of two potential scenarios—detained or released. For example, if both of them would not commit a new crime regardless of the decision, then the judge should not treat them differently.

Therefore, principal fairness is related to individual fairness [14], which demands that similar individuals should be treated similarly. The critical difference, however, is that for principal fairness the similarity is measured based on the potential outcomes rather than observed variables such as the observed outcome, covariates or any function of them. Principal fairness can also be seen as a causal formulation of disparate impact rather than disparate treatment [3]. This means that a decision, which is fair for one outcome, may not be fair for another outcome.

Kosuke Imai is Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge, Massachusetts 02138, USA (e-mail: imai@harvard.edu). Zhichao Jiang (corresponding author) is Professor, School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275, China (e-mail: jiangzhch7@mail.sysu.edu.cn).

1. THREE EXISTING STATISTICAL FAIRNESS CRITERIA

We begin by briefly reviewing the three existing statistical fairness criteria. Let $D_i \in \{0, 1\}$ be the binary decision variable and $Y_i \in \{0, 1\}$ be the binary outcome variable of interest. For the simplicity of exposition, we assume that the outcome and treatment variables are both binary, but the framework can be extended to other variable types.

We consider the following popular statistical fairness criteria.

DEFINITION 1.1 (Statistical fairness). A decision-making mechanism is fair with respect to the outcome of interest Y_i and the protected attribute A_i if the resulting decision D_i satisfies a certain conditional independence relationship. Prominent examples of such relationships used in the literature are given below:

(a) OVERALL PARITY: $\Pr(D_i \mid A_i) = \Pr(D_i)$
(b) CALIBRATION: $\Pr(Y_i \mid D_i, A_i) = \Pr(Y_i \mid D_i)$
(c) ACCURACY: $\Pr(D_i \mid Y_i, A_i) = \Pr(D_i \mid Y_i)$

In our criminal justice example, let $D_i = 1$ represent the judge's decision to detain an arrestee, while we use $D_i = 0$ to denote the decision to release. In addition, let $Y_i = 1$ denote that the arrestee commits a new crime whereas $Y_i = 0$ represents no new crime being committed. Suppose that the protected attribute is race. Then the overall parity implies that a judge should detain the same proportion of arrestees across racial groups. In contrast, the calibration criterion requires a judge to make decisions such that the fraction of detained (or released) arrestees who commit a new crime is identical across racial groups. Finally, according to the accuracy criterion, a judge must make decisions such that among those who committed (or did not commit) a new crime, the same proportion of arrestees had been detained across racial groups.

Table 1 shows a numerical example of observed data, which does not satisfy any of the three statistical fairness criteria. For example, among those who committed a new crime, the detention rate is much higher for Group A than Group B, implying that the accuracy criterion is not satisfied. In addition, among those who are detained, the rate of new crime is much higher for Group A than Group B, failing to satisfy the calibration criterion.

As mentioned earlier, the major shortcoming of these popular statistical fairness criteria is that it does not incorporate the causal impact of decision on the outcome of interest. In the current example, these fairness criteria do not take into account how the judge's decision affects the arrestee's behavior. This also means that the existing statistical fairness criteria are not applicable for an online or post-hoc fairness evaluation although they may be used for fairness evaluation based on historical data.

Next, we introduce the principal fairness criterion that addresses this problem of the existing statistical criteria.

2. PRINCIPAL FAIRNESS

To formally define principal fairness, we follow the standard causal inference literature and use $Y_i(d)$ to denote the potential value of the outcome that would be realized if the decision is $D_i = d$ for $d = 0, 1$ (e.g., [15, 19, 31, 34]). Then the observed outcome can be written as $Y_i = Y_i(D_i)$.

Principal strata are defined as the joint potential outcome values, that is, $R_i = (Y_i(1), Y_i(0))$ [16]. Since any causal effect can be written as a function of potential outcomes, for example, $Y_i(1) - Y_i(0)$ and $Y_i(1)/Y_i(0)$, each principal stratum represents how an individual would be affected by the decision with respect to the outcome of interest. In other words, the principal strata contain all the information about how the decision impacts the outcome. Unlike the observed outcome Y_i, however, the potential outcomes, and hence principal strata, represent the predetermined characteristics of individuals and are not affected by the decision. Moreover, since we only observe one potential outcome for any individual, principal strata are not directly observable.

In the criminal justice example, the principal strata are defined by whether or not each arrestee commits a new crime under each of the two scenarios—detained or released—determined by the judge's decision. Specifically, the stratum $R_i = (0, 1)$ represents the "preventable" group of arrestees who would commit a new crime only when released, whereas the stratum $R_i = (1, 1)$ is the "dangerous" group of individuals who would commit a new crime regardless of the judge's decision. Similarly, we may refer to the stratum $R_i = (0, 0)$ as the "safe" group of arrestees who would never commit a new crime, whereas the stratum $R_i = (1, 0)$ represents the "backlash" group of individuals who would commit a new crime only when detained.[1]

Principal fairness implies that the decision is independent of the protected attribute within each principal stratum. In other words, a fair decision-maker can consider a protected attribute only so far as it relates to potential outcomes. We now give the formal definition of principal fairness.

DEFINITION 2.1 (Principal fairness). A decision-making mechanism satisfies principal fairness with respect to the outcome of interest and the protected attribute A_i if the resulting decision D_i is conditionally independent of A_i within each principal stratum R_i, that is, $\Pr(D_i \mid R_i, A_i) = \Pr(D_i \mid R_i)$.

Note that principal fairness requires one to specify the outcome of interest as well as the attribute to be protected.

[1] One could assume that an arrestee can never commit a new crime when detained, implying the absence of the backlash and dangerous groups. Here, we avoid such an assumption for the sake of generality.

TABLE 1
A numerical example that satisfies none of the statistical fairness criteria given in Definition 1.1

	Group A		Group B	
	Detained	Released	Detained	Released
$Y_i = 1$	150	100	100	100
$Y_i = 0$	100	150	120	180

TABLE 2
Numerical illustration of principal fairness that is consistent with the observed data in Table 1. Each cell represents a principal stratum defined by the values of two potential outcomes $(Y_i(1), Y_i(0))$, while two numbers within a cell represent the number of individuals detained $(D_i = 1)$ and that of those released $(D_i = 0)$, respectively. This example satisfies principal fairness because Groups A and B have the same detention rate within each principal stratum

		Group A	
		$Y_i(0) = 1$	$Y_i(0) = 0$
		Dangerous	Backlash
$Y_i(1) = 1$	Detained $(D_i = 1)$	120	30
	Released $(D_i = 0)$	30	30
		Preventable	Safe
$Y_i(1) = 0$	Detained $(D_i = 1)$	70	30
	Released $(D_i = 0)$	70	120

		Group B	
		$Y_i(0) = 1$	$Y_i(0) = 0$
		Dangerous	Backlash
$Y_i(1) = 1$	Detained $(D_i = 1)$	80	20
	Released $(D_i = 0)$	20	20
		Preventable	Safe
$Y_i(1) = 0$	Detained $(D_i = 1)$	80	40
	Released $(D_i = 0)$	80	160

As such, a decision-making mechanism that is fair with respect to one outcome may not be fair with respect to another outcome. This may be an undesirable feature if one's goal is to develop a fair decision rule that is applicable to multiple outcomes. Note that this definition is generalizable to any treatment and outcome variable types. For example, if the treatment is a continuous variable, there exist an infinite number of principal strata, but the conditional independence relation in Definition 2.1 is still well-defined.

Table 2 presents the numerical example of principal strata that is consistent with the observed data example shown in Table 1. In other words, for each group, if we compute the observed outcome $Y_i = Y_i(D_i)$ based on Table 2, then its distribution equals that of Table 1.

Recall that this example does not meet any of the three statistical fairness criteria discussed above. The example, however, satisfies principal fairness because the detention rate is identical between Groups A and B within each principal stratum. For instance, within the "dangerous" stratum, the detention rate is 80% for both groups, while it is only 20% for them within the "safe" stratum. Indeed, the decision is independent of group membership given principal strata, thereby satisfying principal fairness.

Principal fairness differs from these statistical fairness criteria in that it accounts for how the decision affects the outcome. In particular, although the accuracy criterion resembles principal fairness, the former conditions upon the observed rather than potential outcomes. This is why these two criteria are different. For example, among those who committed a new crime, the detention rate is much higher for Group A than Group B, failing to meet the accuracy criterion. The reason is that among these arrestees, the proportion of "dangerous" individuals is greater for Group A than that for Group B, and the judge is on average more likely to issue the detention decision for these individuals.

Independent of our work but closely related to the idea of principal fairness, Kallus and Zhou coarsen principal strata into two groups—the "responders" who benefit from the treatment, that is, $Y_i(1) > Y_i(0)$, and "anti-responders" who do not, that is, $Y_i(1) \leq Y_i(0)$—and consider the conditional probability of decision given a protected attribute within each of the coarsened groups [23]. Principal fairness generalizes their work by considering all principal strata.

3. RELATIONSHIP BETWEEN PRINCIPAL FAIRNESS AND STATISTICAL FAIRNESS CRITERIA

How should we reconcile this tension between principal fairness and the existing statistical fairness criteria? The tradeoffs between different fairness criteria are not new. As shown in the literature (e.g., [2, 9, 26]), it is generally impossible to simultaneously satisfy the three statistical fairness criteria introduced in Definition 1.1. In some cases, however, principal fairness implies all three statistical fairness criteria. The following theorem provides a sufficient condition.

THEOREM 3.1. *Suppose that $A_i \perp\!\!\!\perp R_i$. Then principal fairness in Definition 2.1 implies all three statistical definitions of fairness given in Definition 1.1.*

The proof is given in the Appendix.

The condition states that different protected groups have the same distribution of principal strata R_i. In the criminal justice example, this means that no group is inherently more dangerous than the other. This independence differs from the equal base rate condition, that is, $Y_i \perp\!\!\!\perp A_i$, that has been identified in the literature as a sufficient condition for simultaneously satisfying the three statistical existing fairness criteria [26]. The equal base

rate condition is based on observed outcomes, which may be affected by the decision under consideration. In contrast, our sufficient condition, $A_i \perp\!\!\!\perp R_i$, is about the independence between the protected attribute and principal strata. Principal strata are based on potential outcomes, which cannot be affected by the decision, and hence are considered as the characteristics of arrestees. As a result, $A_i \perp\!\!\!\perp R_i$ does not necessarily imply the equal base rate condition, or vice versa. It can be shown, however, that if principal fairness holds, $A_i \perp\!\!\!\perp R_i$ also implies the equal base rate condition. In other words, according to Theorem 3.1, principal fairness represents an alternative condition under which statistical fairness criteria hold simultaneously.

In many settings, it is reasonable to assume that the protected attribute does not directly affect *potential* outcomes. In the criminal justice example, being a member of a particular racial group should not make one inherently more dangerous. The protected attribute can, however, affect potential outcomes through other mediating variables. In particular, the existence of racial discrimination can yield an association between race and various socioeconomic variables, which in turn generates the dependence between race and potential outcomes. For this reason, the independence condition in Theorem 3.1 is likely to be violated in many real-world applications.

Thus, we further investigate the connection between principal fairness and the statistical fairness criteria in more general settings without requiring the independence condition $A_i \perp\!\!\!\perp R_i$. Consider the following monotonicity assumption.

ASSUMPTION 1 (Monotonicity).

$$Y_i(1) \leq Y_i(0) \quad \text{for all } i.$$

Assumption 1 is plausible in many applications when the effect of the decision on the outcome is nonpositive for all individuals. In our criminal justice example, the assumption implies that detention makes it no more likely for an arrestee to commit a new crime in comparison to release. Our theoretical results in the remainder of this paper critically depend on Assumption 1 (we develop a sensitivity analysis in Section 7 to assess the robustness of the results to the potential violation of this assumption). Generalization of our results to nonbinary outcomes would require an alternative assumption (see [33] for example).

The following theorem establishes the exact relationship between $\Pr(D_i \mid R_i, A_i)$ with $\Pr(D_i, Y_i \mid A_i)$ under Assumption 1.

THEOREM 3.2. *Under Assumption 1, we have*

$$\begin{aligned}
&\Pr\big(D_i = 1 \mid R_i = (0,0), A_i\big) \\
&\quad = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid A_i)}{\Pr(R_i = (0,0) \mid A_i)}, \\
&\Pr\big(D_i = 1 \mid R_i = (0,1), A_i\big) \\
&\quad = \frac{\Pr(Y_i = 0 \mid A_i)}{\Pr(R_i = (0,1) \mid A_i)} - \frac{\Pr(R_i = (0,0) \mid A_i)}{\Pr(R_i = (0,1) \mid A_i)}, \\
&\Pr\big(D_i = 1 \mid R_i = (1,1), A_i\big) \\
&\quad = \frac{\Pr(D_i = 1, Y_i = 1 \mid A_i)}{\Pr(R_i = (1,1) \mid A_i)}.
\end{aligned}$$

The proof is given in the Appendix (see [23] which derived these identification results by combining $R_i = (0,0)$ and $R_i = (1,1)$ into one group).

Theorem 3.2 shows that the conditional probability of principal strata given the protected attribute, that is, $\Pr(R_i \mid A_i)$, is the key factor in relating principal fairness to the statistical fairness criteria. If A_i is not independent of R_i, principal fairness and the statistical fairness definitions do not imply each other. When $A_i \perp\!\!\!\perp R_i$ holds, however, principal fairness is equivalent to the statistical fairness criteria under the monotonicity assumption. This result is stated as the following corollary.

COROLLARY 1. *Suppose that $A_i \perp\!\!\!\perp R_i$ holds. Then, under Assumption 1, principal fairness is equivalent to the three statistical fairness criteria given in Definition 1.1.*

The proof is given in the Appendix.

4. COMPARISON WITH THE EXISTING CAUSALITY-BASED FAIRNESS CRITERIA

We are not the first one to incorporate causality into the study of algorithmic fairness. In this section, we explain how principal fairness differs from the existing causality-based fairness criteria.

4.1 Counterfactual Equalized Odds Criterion

The explicit conditioning of potential outcomes in fairness criteria is not new. Independent of our work, Coston et al. propose the following counterfactual equalized odds criterion [12]:

$$\Pr\big(D_i \mid Y_i(0), A_i\big) = \Pr\big(D_i \mid Y_i(0)\big). \tag{1}$$

The authors justify conditioning on the potential outcome under the control condition, $Y_i(0)$, by arguing that it represents a "natural baseline" in most risk assessment settings. Indeed, if researchers have historical data where the outcome under the baseline condition is observed for all units, that is, $Y_i = Y_i(0)$, then this criterion is equivalent to the statistical accuracy criterion described above. Based on such historical data, it is also easy to use the counterfactual equalized odds criteria in fairness evaluation.

Unlike the counterfactual equalized odds criterion, principal fairness conditions on principal strata, which is

defined by all potential outcomes rather than a baseline potential outcome alone. This means that in the case of a binary treatment, principal fairness includes $Y_i(1)$ as well as $Y_i(0)$. The key idea is that principal fairness considers how the decision impacts individuals, requiring the comparison of all potential outcomes. In contrast, the counterfactual equalized odds criterion focuses on the assessment of risk, which is defined as the outcome in the absence of an intervention.

The difference between the two criteria can be illustrated via the numerical example in Table 2. As explained earlier, this example satisfies principal fairness, and yet it fails to meet the counterfactual equalized odds criterion. For example, among those who would commit a crime if released, the detention rate is higher for Group A (19/29) than Group B (16/26). The reason is that those who would commit a crime if released include both the "dangerous" and "preventable" individuals. The proportion of "dangerous" individuals is larger for Group A than that for Group B, and the judge is more likely to impose a detention decision for these individuals.

The counterfactual equalized odds criterion could be viewed as a special case of principal fairness when the decision is binary and the potential outcome under the treatment condition $Y_i(1)$ is constant across individuals (this is different from Assumption 1). For example, if no individual can commit a new crime when detained, the two criteria are equivalent. In our empirical application, however, we find that a new crime can be committed even when an arrestee is detained [20]. In addition, there are many settings where such an assumption is not appropriate. They include the impacts of lending decisions on household finance, and the effects of admissions decisions on future wages.

In general, the following theorem establishes a sufficient condition under which principal fairness implies the counterfactual equalized odds criterion.

THEOREM 4.1. *Suppose that* $Y_i(1) \perp\!\!\!\perp A_i \mid Y_i(0)$. *Then principal fairness implies the counterfactual equalized odds criterion, that is,* $\Pr(D_i \mid Y_i(0), A_i) = \Pr(D_i \mid Y_i(0))$.

The proof is given in the Appendix.

This conditional independence relation implies, in our example, that among those who exhibit the same behavior under the release decision, the crime rate under the detention decision is identical for Groups A and B. This condition is violated in many settings where the protected attribute is associated with $Y_i(1)$ through variables other than $Y_i(0)$. Thus, it is important to consider the joint potential outcomes as done in principal fairness rather than the baseline potential outcome alone.

4.2 Counterfactual Fairness

In the algorithmic fairness literature, *counterfactual fairness* represents one prominent fairness criterion that builds upon the causal inference framework. Kusner et al. define the counterfactual fairness by considering the potential decision when the protected attributes are set to a fixed value [28]. Under their definition, a decision is counterfactually fair if a protected attribute does not have a causal effect on the decision. In the criminal justice example, counterfactual fairness implies that the decision an arrestee would receive if he/she were white should be similar to the decision that would be given if the arrestee were black.

Formally, we can write this criterion as

$$\Pr\{D_i(a) = 1\} = \Pr\{D_i(a') = 1\}$$

for any $a \neq a'$ where $D_i(a)$ represents the potential decision when the protected attribute A_i takes the value a. Below, we briefly compare principal fairness with counterfactual fairness.

First, while principal fairness considers the potential outcomes with respect to different decisions, counterfactual fairness is based on the potential outcomes regarding different values of the protected attribute. In the causal inference literature, some advocated the mantra "no causation without manipulation" by pointing out the difficulty of imagining a hypothetical intervention of altering one's immutable characteristics such as race and gender (e.g., [19]). Addressing this issue often requires one to consider alternative causal quantities such as the causal effects of perceived attributes [17] and stochastic intervention of mediators [21]. In contrast, principal fairness avoids these conceptual issues and can be evaluated under the widely used unconfoundedness and monotonicity assumptions.

Second, while principal fairness is based on the conditional independence between the *realized* decision D_i and the protected attribute A_i, counterfactual fairness requires the distribution of *potential* decision to be equal across the values of the protected attribute. Counterfactual fairness can be defined at an individual level, that is, $D_i(a) = D_i(a')$ and can also be aggregated to any group. Counterfactual fairness demands that, for example, an arrestee should receive the same decision even if he/she were to belong to a different racial group. In contrast, principal fairness, like existing statistical fairness criteria, is fundamentally a group-level notion and cannot be defined at an individual level. An important limitation, therefore, is that ensuring group-level fairness may not guarantee individual-level fairness.

Finally, recall that as shown in Corollary 1, principal fairness implies all other statistical fairness criteria under the assumption of $A_i \perp\!\!\!\perp R_i$. However, even under this assumption, principal fairness neither implies nor is implied

by counterfactual fairness. As the following example illustrates, a decision rule that directly depends on the protected attribute can satisfy principal fairness while failing to meet counterfactual fairness. Alternatively, a decision rule that does not depend on the protected attribute can meet counterfactual fairness but may fail to meet principal fairness.

EXAMPLE. Consider a population characterized by the following distributions of principal strata $R \in \{(0,0), (1,0), (0,1), (1,1)\}$, the protected attribute $A \in \{0,1\}$, and the covariate $X \sim \text{Unif}(0,1)$,

$$\begin{aligned}
&\Pr(A = 1 \mid X) = X,\\
&\Pr\big(R = (1,1) \mid A = a, X\big)\\
&\quad = \Pr\big(R = (0,0) \mid A = a, X\big) = 0.3,\\
&\Pr\big(R = (1,0) \mid A = a, X\big)\\
&\quad = \Pr\big(R = (0,1) \mid A = a, X\big) = 0.2
\end{aligned}$$

for $a = 0, 1$. This implies $A \perp\!\!\!\perp R$. Consider the decision rule of the following form, $D = \mathbf{1}\{\alpha X + \beta A \geq 1\}$. Suppose $\alpha = 5/2$ and $\beta = -1$. Then we have

$$\begin{aligned}
&\Pr\big\{D(1) = 1\big\} = 0.2,\\
&\Pr\big\{D(0) = 1\big\} = 0.6,\\
&\Pr(D = 1 \mid R = r, A = 1) = \Pr(X \geq 0.8 \mid A = 1) = 0.36,\\
&\Pr(D = 1 \mid R = r, A = 0) = \Pr(X \geq 0.4 \mid A = 0) = 0.36.
\end{aligned}$$

Thus, the decision rule violates counterfactual fairness while satisfying principal fairness. Moreover, the three statistical fairness criteria also hold since $A \perp\!\!\!\perp R$. In contrast, consider $\alpha = 5/2$ and $\beta = 0$. Then we have

$$\begin{aligned}
&\Pr\big\{D(1) = 1\big\} = \Pr\big\{D(0) = 1\big\} = 0.6,\\
&\Pr(D = 1 \mid R = r, A = 1) = \Pr(X \geq 0.4 \mid A = 1) = 0.84,\\
&\Pr(D = 1 \mid R = r, A = 0) = \Pr(X \geq 0.4 \mid A = 0) = 0.36.
\end{aligned}$$

Thus, the decision rule violates principal fairness while satisfying counterfactual fairness.

5. CONDITIONAL FAIRNESS CRITERIA

Although we have so far focused on fairness criteria based on marginal distributions, policy makers and researchers may be interested in evaluating fairness within each subpopulation defined by a set of pretreatment covariates. The importance of such conditioning covariates has been recognized in the algorithmic fairness literature. Specifically, even when a statistical fairness criterion holds conditional on a set of covariates, the same criterion may not be satisfied without those conditioning covariates. The reason is that these conditioning covariates may be correlated with the protected attribute itself. This problem is called inframarginality in the literature and applies to all statistical fairness criteria including principal fairness [11]. The inframarginality problem simply reflects an unavoidable fact that conditional independence does not necessarily imply marginal independence and vice versa.

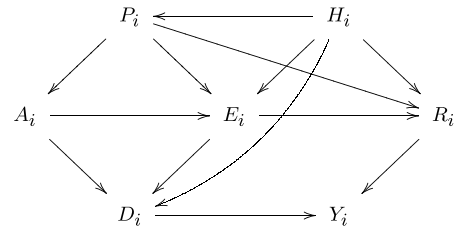

FIG. 1. *Direct acyclic graph for the relationship between the protected attribute A_i and principal strata R_i. In the criminal justice application, A_i represents the race of an arrestee, R_i is their risk category (safe, preventable, dangerous and backlash), D_i represents the decision of judge, P_i represents parents' characteristics, which include their attributes and socioeconomic status (SES), E_i represents arrestee's own experiences such as SES and H_i represents historical processes. Finally, Y_i is the indicator of committing a new crime, which is a deterministic function of judge's decision D_i and risk category R_i. The conditional independence $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$ holds with $\mathbf{W}_i = (H_i, P_i, E_i)$.*

The following theorem shows that if the conditioning covariates eliminate the dependence between the protected attribute and principal stratum, then conditional on these covariates, principal fairness implies all three statistical definitions of fairness and the counterfactual equalized odds criterion.

THEOREM 5.1. *Suppose that there exist a set of variables $\mathbf{W}_i$ such that $A_i \perp\!\!\!\perp R_i \mid \mathbf{W}_i$ holds. Then, conditional on $\mathbf{W}_i$, principal fairness implies the counterfactual equalized odds criterion and all three statistical definitions of fairness. That is,* $\Pr(D_i \mid R_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid R_i, \mathbf{W}_i)$ *implies* $\Pr(D_i \mid Y_i(0), \mathbf{W}_i, A_i) = \Pr(D_i \mid Y_i(0), \mathbf{W}_i)$, $\Pr(D_i \mid \mathbf{W}_i, A_i) = \Pr(D_i \mid \mathbf{W}_i)$, $\Pr(Y_i \mid D_i, \mathbf{W}_i, A_i) = \Pr(Y_i \mid D_i, \mathbf{W}_i)$, *and* $\Pr(D_i \mid Y_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid Y_i, \mathbf{W}_i)$. *Moreover, if Assumption 1 also holds, then principal fairness is equivalent to all three statistical definitions of fairness conditional on $\mathbf{W}_i$.*

The proof is omitted because it is similar to those of Theorems 3.1, 3.2 and 4.1 except that we condition on $\boldsymbol{W}_i$.

The conditional independence $A_i \perp\!\!\!\perp R_i \mid \mathbf{W}_i$ means that no racial group is inherently more dangerous than other groups once we account for relevant factors $\mathbf{W}_i$. In a causal model, the absence of direct effect of A_i on R_i implies the existence of $\mathbf{W}_i$ that satisfies $A_i \perp\!\!\!\perp R_i \mid \mathbf{W}_i$ where $\mathbf{W}_i$ can include mediators as well as common causes. The lack of the direct effect of race can be viewed as an axiomatic assumption that belonging to a particular racial group does not make one inherently more dangerous than members of other racial groups.

For illustration, consider the causal model, shown as a directed acyclic graph in Figure 1, in the context of

the criminal justice example. The race of an arrestee, A_i, is affected by his/her parents' characteristics, which include their attributes and social economic status (SES), P_i. The arrestee's own experiences, E_i, are influenced by their race, A_i, their parents' characteristics, P_i, and the historical processes such as slavery and Jim Crow laws, H_i, which also affect the parents' characteristics.

Under this causal model, all of these three covariates affect the risk category of arrestee (principal strata; i.e., safe, preventable, dangerous and backlash), R_i, whereas the judge's decision, D_i, is affected by the race, the experiences and the historical processes. The key assumption of the model is that the arrestee's race does not *directly* affect their risk category, as indicated by the absence of a direct arrow between these two variables. As a result, under this model, the arrestee's race is conditionally independent of risk category, that is, $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$, where $\mathbf{W}_i = (H_i, P_i, E_i)$. In other words, once we account for these factors, no racial group has an innate tendency to be dangerous relative to the other groups.

Theorem 5.1 shows that once we condition on $\mathbf{W}_i$ that satisfies $A_i \perp\!\!\!\perp R_i \mid \mathbf{W}_i$, principal fairness implies the counterfactual equalized odds criterion and all statistical fairness criteria. However, this result should not be used to justify the appropriateness of conditioning on $\mathbf{W}_i$. The reason is that the inclusion of conditioning covariates in fairness criteria can lead to discrimination based on those variables. If the conditioning covariates are good proxy variables for the protected attribute, then any conditional fairness criteria could lead to discrimination against those groups who should be protected. Thus, the choice of conditioning covariates must be made with special care [6, 25].

Finally, the conditioning covariates also play an important role in counterfactual fairness as well but for a different reason. Conditioning on covariates that are affected by the protected attribute need to be done carefully to avoid inducing a post-treatment bias (see, e.g., [25, 27]). To address this issue, researchers have considered path-specific effects through the framework of causal mediation analysis (e.g., [8, 30, 32, 37]). In such an analysis, a key question is which mediators should be included.

To further illustrate the difference between counterfactual fairness and principal fairness with conditioning, we again consider the causal model shown in Figure 1. Suppose we would like to condition on E_i. Then counterfactual fairness requires that the race has no effect on the decision other than through this variable. Because the race can only affect the decision directly or through E_i, counterfactual fairness is violated conditional on E_i due to the existence of the direct effect. In contrast, principal fairness may still hold conditional on E_i if the association from the direct effect of A_i on D_i cancels out with the association from the common cause H_i. Consistent with Example 4.2, a decision rule that directly depends on the protected attribute can satisfy principal fairness.

6. EMPIRICAL EVALUATION AND POLICY LEARNING UNDER PRINCIPAL FAIRNESS

Finally, we discuss how to use the above theoretical results in empirical studies. We first show how to empirically assess the independence conditions in Theorems 4.1 and 5.1, that is, $Y_i(1) \perp\!\!\!\perp A_i \mid Y_i(0)$ and $A_i \perp\!\!\!\perp R_i$. To do this, we must identify the distribution of the principal stratum within each group defined by the protected attribute. We begin by introducing the following unconfoundedness assumption, which is widely used in the causal inference literature.

ASSUMPTION 2 (Unconfoundedness). $Y_i(d) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$ for any d.

Assumption 2 holds if $\mathbf{X}_i$ contains all the information used for decision-making. In practice, if we are unsure about whether the protected attribute is used for decision-making, we may still include it in $\mathbf{X}_i$ to make the unconfoundedness assumption more plausible [35].

The next theorem shows that under Assumptions 1 and 2, the evaluation of the independence relations, $Y_i(1) \perp\!\!\!\perp A_i \mid Y_i(0)$ and $A_i \perp\!\!\!\perp R_i$, reduces to the estimation of conditional probability, $\Pr(Y_i = 1 \mid D_i, \mathbf{X}_i)$, from the observed data.

THEOREM 6.1. *Under Assumptions 1 and 2, we have*

$$\Pr\{Y_i(1) = 1 \mid A_i, Y_i(0) = 1\} = \frac{m_1(A_i)}{m_0(A_i)},$$
$$\Pr(R_i = (0,0) \mid A_i) = 1 - m_0(A_i),$$
$$\Pr(R_i = (0,1) \mid A_i) = m_0(A_i) - m_1(A_i),$$
$$\Pr(R_i = (1,1) \mid A_i) = m_1(A_i),$$

where $m_d(A_i) = \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = d, \mathbf{X}_i) \mid A_i\}$.

The proof is given in the Appendix.

Theorem 6.1 shows that we can empirically evaluate the validity of $Y_i(1) \perp\!\!\!\perp A_i \mid Y_i(0)$ and $A_i \perp\!\!\!\perp R_i$ by checking whether the distribution of principal strata R_i depends on the protected attribute A. The result also holds conditional on any covariates that are included in $\mathbf{X}_i$.

Second, we consider the empirical evaluation of principal fairness. Combining Theorems 3.2 and 6.1, the following corollary shows that the same assumptions used in Theorem 6.1 are sufficient for identifying the conditional distribution of decision D_i given the principal strata and the protected attribute. Using this conditional distribution, one can empirically assess the principal fairness of the decision.

COROLLARY 2. *Under Assumptions 1 and 2, we have*

$$\Pr\{D_i = 1 \mid R_i = (0,0), A_i\} = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid A_i)}{1 - m_0(A_i)},$$

$$\Pr\{D_i = 1 \mid R_i = (0,1), A_i\} = \frac{m_0(A_i) - \Pr(Y_i = 1 \mid A_i)}{m_0(A_i) - m_1(A_i)},$$
$$\Pr\{D_i = 1 \mid R_i = (1,1), A_i\} = \frac{\Pr(D_i = 1, Y_i = 1 \mid A_i)}{m_1(A_i)}.$$

The proof is given in the Appendix. The formulas also hold conditional on any covariates that are included in $\mathbf{X}_i$, and thus allow for the evaluation of conditional principal fairness.

Finally, we consider policy learning under principal fairness. For simplicity, we focus on a deterministic policy $D_i = \delta(\boldsymbol{V}_i)$, where $\boldsymbol{V}_i$ represents the covariates used for making decisions. Suppose that the protected attribute is binary. Then principal fairness requires the decision rule $\delta(\boldsymbol{V}_i)$ to satisfy the following equality constraint, $\Pr\{\delta(\boldsymbol{V}_i) \mid R_i, A_i = 1\} = \Pr\{\delta(\boldsymbol{V}_i) \mid R_i, A_i = 0\}$. This constraint may be difficult to satisfy due to the fact that R_i is an unobserved variable. The following theorem expresses this probability, $\Pr\{\delta(\boldsymbol{V}_i) \mid R_i, A_i = 1\}$, in a different form that only depends on observed variables.

THEOREM 6.2. *Suppose that Assumptions 1 holds and the decision is a function of $\boldsymbol{V}_i$, that is, $D_i = \delta(\boldsymbol{V}_i)$. Then we have*

$$\Pr\{\delta(\boldsymbol{V}_i) = 1 \mid R_i = r, A_i\} = \mathbb{E}\left[\frac{e_r(\boldsymbol{V}_i, A_i)}{\mathbb{E}\{e_r(\boldsymbol{V}_i, A_i) \mid A_i\}}\delta(\boldsymbol{V}_i) \mid A_i\right],$$

for $r = (0,0), (0,1)$ and $(1,1)$ where $e_r(\boldsymbol{V}_i, A_i) = \Pr(R_i = r \mid \boldsymbol{V}_i, A_i)$.

The proof is given in the Appendix.

The identification formulas for $e_r(W_i, A_i)$ are given in Theorem 6.1. When we know which covariates are used in the decision D_i, these identification formulas provide an alternative way to evaluate principal fairness in addition to Corollary 2. Specifically, Theorem 6.2 shows that $\Pr(D_i \mid R_i = r, A_i)$ is equal to the decision probability within each protected group in a weighted population. The weights depend on the proportions of principal strata given the covariates and the protected attribute. Therefore, to learn a policy that satisfies principal fairness, one could first estimate $e_r(\boldsymbol{V}_i, A_i)$ using Theorem 6.1 and then use these estimated weights to augment the existing fairness-aware policy learning approaches with the principal fairness constraints (e.g., [1, 7, 24]).

7. SENSITIVITY ANALYSIS FOR MONOTONICITY

Without monotonicity, the proportions of principal strata are not identifiable. This is why all results in Section 6 rely on Assumption 1. In this section, we propose a sensitivity analysis for this key identification assumption. As the sensitivity parameter, we use the ratio between the proportion of stratum $(1,0)$ and stratum $(0,1)$ conditional on the covariates,

$$\xi = \frac{\Pr(R_i = (1,0) \mid \mathbf{X}_i)}{\Pr(R_i = (0,1) \mid \mathbf{X}_i)}. \tag{2}$$

The sensitivity parameter characterizes the deviation from monotonicity. If $\xi = 0$, then monotonicity holds. Kallus and Zhou develop a similar sensitivity analysis but their sensitivity parameter is defined on the difference scale [23]. The following theorem generalizes Theorems 3.2 and 6.1 and Corollary 2 using this sensitivity parameter.

THEOREM 7.1. *Suppose $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i$ holds, which is a stronger version of Assumption 2. Using the sensitivity parameter defined in equation (2) with a known $\xi \neq 1$, we can write*

$$\begin{aligned}
&\Pr(D_i = 1 \mid R_i = r, A_i) \\
&= \frac{\Pr(R_i = r \mid D_i = 1, A_i)\Pr(D_i = 1 \mid A_i)}{\Pr(R_i = r \mid A_i)} \\
&= \frac{\mathbb{E}\{\Pr(R_i = r \mid D_i = 1, \mathbf{X}_i) \mid D_i = 1, A_i\}\Pr(D_i = 1 \mid A_i)}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) \mid A_i\}} \\
&= \frac{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) \mid D_i = 1, A_i\}\Pr(D_i = 1 \mid A_i)}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) \mid A_i\}}
\end{aligned}$$

for $r \in \{(0,0), (0,1), (1,1), (1,0)\}$, where

$$\begin{aligned}
&\Pr(R_i = (0,0) \mid \mathbf{X}_i) \\
&= 1 - \frac{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi}, \\
&\Pr(R_i = (0,1) \mid \mathbf{X}_i) \\
&= \frac{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi}, \\
&\Pr(R_i = (1,1) \mid \mathbf{X}_i) \\
&= \frac{\Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i)}{1 - \xi}, \\
&\Pr(R_i = (1,0) \mid \mathbf{X}_i) \\
&= \frac{\xi \Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi}.
\end{aligned}$$

The proof is given in the Appendix. For policy learning under principal fairness, the formula given in Theorem 6.2 still holds without monotonicity. Therefore, under equation (2), we can estimate $e_r(\boldsymbol{V}_i, A_i)$ using the formulas of $\Pr(R_i = r \mid \mathbf{X}_i)$ given in Theorem 7.1.

8. CONCLUDING REMARKS

To assess the fairness of human and algorithmic decision-making, one must consider how the decisions themselves affect individuals. Such consideration requires the notion of fairness to be placed in the causal inference framework. In a separate work, we apply the idea of principal fairness to the common settings, in which humans

make decisions partly based on algorithmic recommendations [20]. Since human decision-makers rather than algorithms ultimately impact individuals, one must assess whether algorithmic recommendations improve the fairness of human decisions. We empirically examine this issue through the experimental evaluation of the pretrial risk assessment instrument widely used in the US criminal justice system.

The difference between principal fairness and counterfactual equalized odds criterion sheds light on the predictive performance evaluation of algorithmic risk assessments. The current literature focuses on the prediction accuracy of $Y_i(0)$ when evaluating algorithmic fairness under the counterfactual equalized odds criterion (e.g., [12, 29]). However, in general, $Y_i(0)$ alone does not fully characterize counterfactual outcomes: individuals with the same value of $Y_i(0)$ may differ in the value of $Y_i(d)$ where $d \neq 0$. Principal fairness generalizes counterfactual equalized odds criterion by considering principal strata which depend on all potential outcomes. In particular, the evaluation of algorithmic decision or recommendation requires one to condition on principal strata rather than the observed outcome or a single potential outcome.

Although this paper focuses on the introduction of principal fairness as a new fairness concept, much work remains to be done. In particular, future work should consider the development of algorithms that achieve principal fairness. In a separate paper, we consider a methodological framework for policy learning that involves the joint potential outcomes [4]. Incorporating principal fairness as a constraint within this framework may enable us to learn fair policies from data.

Another possible direction is the extension of principal fairness to a dynamic decision-making system. As previously pointed out [10, 13], real-world algorithmic systems operate in complex environments that are constantly changing, often due to the actions of algorithms themselves. Therefore, an explicit consideration of the dynamic causal interactions between algorithms and human decision-makers can help us develop long-term fairness criteria.

Finally, but importantly, principal fairness does not solve the big data's disparate impact problem pointed out by Barocas and Selbst [3]. Historical biases can affect principal fairness through potential outcomes and data used to estimate them. In particular, by conditioning on potential outcomes, a principal fairness criterion may end up inheriting historical biases that have existed in the world and data derived from it. One way to address this issue is to further adjust for such biases as discussed in Section 5. The tasks of identifying and measuring these historical factors, however, remain challenging and are likely to require a better understanding of the underlying causal structure. We leave this and other open problems to future work.

APPENDIX: PROOFS

Proof of Theorem 3.1

We prove a more general version of Theorem 3.1 with any variables $\boldsymbol{V}_i$ in the conditioning set. That is, under $A_i \perp\!\!\!\perp R_i \mid \boldsymbol{V}_i$, principal fairness in implies all three statistical definitions of fairness conditional on $\boldsymbol{V}_i$.

Because the observed stratum $(D_i = 1, Y_i = 1)$ is a mixture of principal strata $R_i = (1,0), (1,1)$, we have

$$
\begin{aligned}
&\Pr(D_i = 1, Y_i = 1 \mid \boldsymbol{V}_i, A_i) \\
&\quad = \Pr\big(D_i = 1, R_i = (1,0) \mid \boldsymbol{V}_i, A_i\big) \\
&\qquad + \Pr\big(D_i = 1, R_i = (1,1) \mid \boldsymbol{V}_i, A_i\big) \\
&\quad = \Pr\big(D_i = 1 \mid R_i = (1,0), \boldsymbol{V}_i, A_i\big) \\
&\qquad \times \Pr\big(R_i = (1,0) \mid \boldsymbol{V}_i, A_i\big) \\
&\qquad + \Pr\big(D_i = 1 \mid R_i = (1,1), \boldsymbol{V}_i, A_i\big) \\
&\qquad \times \Pr\big(R_i = (1,1) \mid \boldsymbol{V}_i, A_i\big) \\
&\quad = \Pr\big(D_i = 1 \mid R_i = (1,0), \boldsymbol{V}_i\big) \Pr\big(R_i = (1,0) \mid \boldsymbol{V}_i\big) \\
&\qquad + \Pr\big(D_i = 1 \mid R_i = (1,1), \boldsymbol{V}_i\big) \Pr\big(R_i = (1,1) \mid \boldsymbol{V}_i\big) \\
&\quad = \Pr\big(D_i = 1, R_i = (1,0) \mid \boldsymbol{V}_i\big) \\
&\qquad + \Pr\big(D_i = 1, R_i = (1,1) \mid \boldsymbol{V}_i\big) \\
&\quad = \Pr(D_i = 1, Y_i = 1 \mid \boldsymbol{V}_i),
\end{aligned}
$$

where the third equality follows from principal fairness and the assumption $A_i \perp\!\!\!\perp R_i \mid \boldsymbol{V}_i$. Similarly, we can show

$$
\Pr(D_i = d, Y_i = y \mid \boldsymbol{V}_i, A_i) = \Pr(D_i = d, Y_i = y \mid \boldsymbol{V}_i)
$$

for $d, y = 0, 1$. This implies the three statistical definitions of fairness.

Proof of Theorem 3.2

We prove a more general version of Theorem 3.2 with any variables $\boldsymbol{V}_i$ in the conditioning set. From Assumption 1, we obtain

$$
\begin{aligned}
&\Pr\big(D_i = 1 \mid R_i = (0,0), \boldsymbol{V}_i, A_i\big) \\
&\quad = 1 - \frac{\Pr(D_i = 0, R_i = (0,0) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,0) \mid \boldsymbol{V}_i, A_i)} \\
&\quad = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,0) \mid \boldsymbol{V}_i, A_i)}, \\
&\Pr\big(D_i = 1 \mid R_i = (1,1), \boldsymbol{V}_i, A_i\big) \\
&\quad = \frac{\Pr(D_i = 1, R_i = (1,1) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (1,1) \mid \boldsymbol{V}_i, A_i)} \\
&\quad = \frac{\Pr(D_i = 1, Y_i = 1 \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (1,1) \mid \boldsymbol{V}_i, A_i)},
\end{aligned}
$$

and

$$
\Pr\big(D_i = 1 \mid R_i = (0,1), \boldsymbol{V}_i, A_i\big)
$$

$$
\begin{aligned}
&= \frac{\Pr(D_i = 1, R_i = (0,1) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,1) \mid \boldsymbol{V}_i, A_i)} \\
&= \frac{\Pr(D_i = 1 \mid \boldsymbol{V}_i, A_i) - \Pr(D_i = 1, R_i = (1,1) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,1) \mid \boldsymbol{V}_i, A_i)} \\
&\quad - \frac{\Pr(D_i = 1, R_i = (0,0) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,1) \mid \boldsymbol{V}_i, A_i)} \\
&= \frac{\Pr(D_i = 1, Y_i = 0 \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,1) \mid \boldsymbol{V}_i, A_i)} \\
&\quad - \frac{\Pr(R_i = (0,0) \mid \boldsymbol{V}_i, A_i) - \Pr(D_i = 0, Y_i = 0 \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = (0,1) \mid \boldsymbol{V}_i, A_i)} \\
&= \frac{\Pr(Y_i = 0 \mid \boldsymbol{V}_i, A_i) - \Pr(R_i = (0,0) \mid \boldsymbol{V}_i, A_i)}{\Pr(R_i = 1 \mid \boldsymbol{V}_i, A_i)}.
\end{aligned}
$$

Proof of Corollary 1

From Theorem 3.2, under $A_i \perp\!\!\!\perp R_i$, principal fairness is equivalent to

$$
\begin{aligned}
\Pr(D_i = 0, Y_i = 0 \mid A_i) &= \Pr(D_i = 0, Y_i = 0), \\
\Pr(Y_i = 0 \mid A_i) &= \Pr(Y_i = 0), \\
\Pr(D_i = 1, Y_i = 1 \mid A_i) &= \Pr(D_i = 0, Y_i = 0),
\end{aligned}
$$

which are equivalent to the three statistical fairness criteria.

Proof of Theorem 4.1

By the law of total probability, we have

$$
\begin{aligned}
&\Pr\{D_i \mid Y_i(0), A_i\} \\
&\quad = \sum_{y_1 = 0,1} \Pr\{D_i \mid Y_i(1) = y_1, Y_i(0), A_i\} \\
&\qquad \times \Pr\{Y_i(1) = y_1 \mid Y_i(0), A_i\} \\
&\quad = \sum_{y_1 = 0,1} \Pr\{D_i \mid Y_i(1) = y_1, Y_i(0)\} \\
&\qquad \times \Pr\{Y_i(1) = y_1 \mid Y_i(0)\} \\
&\quad = \Pr\{D_i \mid Y_i(0)\},
\end{aligned}
$$

where the second equality follows from principal fairness and $Y_i(1) \perp\!\!\!\perp A_i \mid Y_i(0)$.

Proof of Theorem 6.1

Under Assumption 1, we have

$$
\begin{aligned}
&\Pr(R_i = (0,0) \mid A_i) \\
&\quad = \Pr(Y_i(0) = 0 \mid A_i), \qquad (3) \\
&\Pr(R_i = (0,1) \mid A_i) \\
&\quad = \Pr(Y_i(0) = 1 \mid A_i) - \Pr(Y_i(1) = 1 \mid A_i), \qquad (4) \\
&\Pr(R_i = (1,1) \mid A_i) \\
&\quad = \Pr(Y_i(1) = 1 \mid A_i). \qquad (5)
\end{aligned}
$$

Under Assumption 2, we have

$$
\Pr\{Y_i(d) = y \mid A_i\} = \mathbb{E}\{\Pr(Y_i = y \mid D_i = d, \mathbf{X}_i) \mid A_i\},
$$

where we assume $\mathbf{X}_i$ contains A_i. Plugging this into equations (3) to (5) yields the formulas in Theorem 6.1.

Proof of Corollary 2

This corollary follows from Theorem 3.2 and equation (8).

Proof of Theorem 6.2

From the law of total probability, we have

$$
\begin{aligned}
&\Pr\{\delta(\boldsymbol{V}_i) = 1 \mid R_i = r, A_i\} \\
&\quad = \mathbb{E}\{\Pr(D_i = 1 \mid \boldsymbol{V}_i, R_i = r, A_i) \mid R_i = r, A_i\} \\
&\quad = \sum_{\boldsymbol{v}} \Pr(\delta(\boldsymbol{V}_i) = 1 \mid \boldsymbol{V}_i = \boldsymbol{v}, A_i) \\
&\qquad \times \Pr(\boldsymbol{V}_i = \boldsymbol{v} \mid R_i = r, A_i) \\
&\quad = \sum_{\boldsymbol{v}} \Bigg[\Pr\{\delta(\boldsymbol{V}_i) = 1 \mid \boldsymbol{V}_i = \boldsymbol{v}, A_i\} \\
&\qquad \times \frac{\Pr(R_i = r \mid \boldsymbol{V}_i = \boldsymbol{v}, A_i) \Pr(\boldsymbol{V}_i = \boldsymbol{v} \mid A_i)}{\Pr(R_i = r \mid A_i)} \Bigg] \\
&\quad = \frac{\sum_{\boldsymbol{v}} \mathbb{E}\{e_r(\boldsymbol{V}_i, A_i)\delta(\boldsymbol{V}_i) \mid \boldsymbol{V}_i = \boldsymbol{v}, A_i\} \Pr(\boldsymbol{V}_i = \boldsymbol{v} \mid A_i)}{\Pr(R_i = r \mid A_i)} \\
&\quad = \frac{\mathbb{E}\{e_r(\boldsymbol{V}_i, A_i)\delta(\boldsymbol{V}_i) \mid A_i\}}{\mathbb{E}\{e_r(\boldsymbol{V}_i, A_i) \mid A_i\}} \\
&\quad = \mathbb{E}\left[\frac{e_r(\boldsymbol{V}_i, A_i)}{\mathbb{E}\{e_r(\boldsymbol{V}_i, A_i) \mid A_i\}} \delta(\boldsymbol{V}_i) \mid A_i \right],
\end{aligned}
$$

where can replace the summation with integral for continuous $\boldsymbol{V}_i$.

Proof of Theorem 7.1

By Bayes' rule, we have

$$
\begin{aligned}
&\Pr(D_i = 1 \mid R_i = r, A_i) \\
&\quad = \frac{\Pr(R_i = r \mid D_i = 1, A_i) \Pr(D_i = 1 \mid A)i)}{\Pr(R_i = r \mid A_i)} \\
&\quad = \frac{\mathbb{E}\{\Pr(R_i = r \mid D_i = 1, \mathbf{X}_i) D_i = 1, A_i\} \Pr(D_i = 1 \mid A)i)}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) \mid A_i\}} \\
&\quad = \frac{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) D_i = 1, A_i\} \Pr(D_i = 1 \mid A)i)}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i) \mid A_i\}},
\end{aligned}
$$

where the second equality follows from the law of total probability and the third equality follows from $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i$. We then derive the formula for $\Pr(R_i = r \mid \mathbf{X}_i)$. Under equation (2) and Assumption 2, we have

$$
\begin{aligned}
&\Pr(Y_i = 0 \mid D_i = 1, \mathbf{X}_i) \\
&\quad = \Pr(R_i = (0,0) \mid \mathbf{X}_i) + \Pr(R_i = (0,1) \mid \mathbf{X}_i), \\
&\Pr(Y_i = 0 \mid D_i = 0, \mathbf{X}_i) \\
&\quad = \Pr(R_i = (0,0) \mid \mathbf{X}_i) + \Pr(R_i = (1,0) \mid \mathbf{X}_i) \\
&\quad = \Pr(R_i = (0,0) \mid \mathbf{X}_i) + \xi \Pr(R_i = (0,1) \mid \mathbf{X}_i).
\end{aligned}
$$

Solving the equations, we obtain

$$\Pr(R_i = (0,1) \mid \mathbf{X}_i) = \frac{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi},$$

$$\Pr(R_i = (0,0) \mid \mathbf{X}_i) = 1 - \frac{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi}.$$

Therefore,

$$\Pr(R_i = (1,0) \mid \mathbf{X}_i) = \frac{\xi \Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i)}{1 - \xi},$$

$$\Pr(R_i = (1,1) \mid \mathbf{X}_i) = \frac{\Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i) - \xi \Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i)}{1 - \xi}.$$

ACKNOWLEDGMENTS

We thank Elias Bareinboim, Hao Chen, Shizhe Chen, Christina Davis, Cynthia Dwork, Peng Ding, Robin Gong, Jim Greiner, Sharad Goel, Nathan Kallus, Gary King, Jamie Robins and Pragya Sur for comments and discussions. We also thank anonymous reviewers of the Alexander and Diviya Magaro Peer Pre-Review Program at IQSS for valuable feedback.

FUNDING

We acknowledge the partial support by the National Science Foundation (SES–2051196) and Cisco Systems, Inc. (CG# 2370386).

REFERENCES

[1] AGARWAL, A., BEYGELZIMER, A., DUDÍK, M., LANGFORD, J. and WALLACH, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning* 60–69. PMLR.

[2] BAROCAS, S., HARDT, M. and NARAYANAN, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. Available at http://www.fairmlbook.org.

[3] BAROCAS, S. and SELBST, A. D. (2016). Big data's disparate impact. *California Law Review* **104** 671–732.

[4] BEN-MICHAEL, E., IMAI, K. and JIANG, Z. (2022). Policy learning with assymetric utilities. Technical Report. ArXiv Preprint. Available at https://arxiv.org/pdf/2206.10479.pdf.

[5] BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M. and ROTH, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **50** 3–44. MR4198551 https://doi.org/10.1177/0049124118782533

[6] BEUTEL, A., CHEN, J., DOSHI, T., QIAN, H., WOODRUFF, A., LUU, C., KREITMANN, P., BISCHOF, J. and CHI, E. H. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society. AIES'*19 453–459. Association for Computing Machinery, New York, NY, USA.

[7] CELIS, L. E., HUANG, L., KESWANI, V. and VISHNOI, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* 319–328.

[8] CHIAPPA, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 7801–7808.

[9] CHOULDECHOVA, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **5** 153–163. https://doi.org/10.1089/big.2016.0047

[10] CHOULDECHOVA, A. and ROTH, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* **63** 82–89.

[11] CORBETT-DAVIES, S. and GOEL, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Technical Report. Available at arXiv:1808.00023.

[12] COSTON, A., MISHLER, A., KENNEDY, E. H. and CHOULDECHOVA, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency* 582–593.

[13] D'AMOUR, A., SRINIVASAN, H., ATWOOD, J., BALJEKAR, P., SCULLEY, D. and HALPERN, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *FAT**'20: *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency* 525–534.

[14] DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O. and ZEMEL, R. (2012). Fairness through awareness. In *Proceedings of the* 3*rd Innovations in Theoretical Computer Science Conference* 214–226. ACM, New York. MR3388391

[15] FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

[16] FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x

[17] GREINER, D. J. and RUBIN, D. B. (2011). Causal effects of perceived immutable characteristics. *Rev. Econ. Stat.* **93** 775–785.

[18] HARDT, M., PRICE, E. and SREBRO, N. (2016). Equality of opportunity in supervised learning. Technical Report. Available at arXiv:1610.02413.

[19] HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

[20] IMAI, K., JIANG, Z., GREINER, D. J., HALEN, R. and SHIN, S. (2022). Experimental evaluation of computer-assisted human decision-making: Application to pretrial risk assessment instrument (with discussions). *J. Roy. Statist. Soc. Ser. A* To appear.

[21] JACKSON, J. W. and VANDERWEELE, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* **29** 825–835. https://doi.org/10.1097/EDE.0000000000000901

[22] JOHNDROW, J. E. and LUM, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Ann. Appl. Stat.* **13** 189–220. MR3937426 https://doi.org/10.1214/18-AOAS1201

[23] KALLUS, N. and ZHOU, A. (2019). Assessing disparate impact of personalized interventions: Identifiability and bounds. In 33*rd Conference on Neural Information Processing Systems*.

[24] KAMISHIMA, T., AKAHO, S. and SAKUMA, J. (2011). Fairness-aware learning through regularization approach. In 2011 *IEEE* 11*th International Conference on Data Mining Workshops* 643–650.

[25] KILBERTUS, N., CARULLA, M. R., PARASCANDOLO, G., HARDT, M., JANZING, D. and SCHÖLKOPF, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* 656–666.

[26] KLEINBERG, J., MULLAINATHAN, S. and RAGHAVAN, M. (2017). Inherent trade-offs in the fair determination of risk scores. In 8*th Innovations in Theoretical Computer Science Conference* C. H. Papadimitrou, ed.). *LIPIcs. Leibniz Int. Proc. Inform.* **67** Art. No. 43. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3754967

[27] KNOX, D., LOWE, W. and MUMMOLO, J. (2022). Administrative records mask racially biased policing. *Am. Polit. Sci. Rev.* **114** 619–637.

[28] KUSNER, M., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. In *Proceedings of the* 31*st Conference on Neural Information Processing Systems*.

[29] MITCHELL, S., POTASH, E., BAROCAS, S., D'AMOUR, A. and LUM, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Appl.* **8** 141–163. MR4243544 https://doi.org/10.1146/annurev-statistics-042720-125902

[30] NABI, R. and SHPITSER, I. (2018). Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[31] NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. (Translated in 1990). *Statist. Sci.* **5** 465–480. MR1092986

[32] PLECKO, D. and BAREINBOIM, E. (2022). Causal Fairness Analysis. Technical Report. ArXiv Preprint. Available at https://arxiv.org/abs/2207.11385.

[33] PLEČKO, D. and MEINSHAUSEN, N. (2020). Fair data adaptation with quantile preservation. *J. Mach. Learn. Res.* **21** Paper No. 242. MR4209528

[34] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* **66** 688–701.

[35] VANDERWEELE, T. J. and SHPITSER, I. (2011). A new criterion for confounder selection. *Biometrics* **67** 1406–1413. MR2872391 https://doi.org/10.1111/j.1541-0420.2011.01619.x

[36] ZAFAR, M. B., VALERA, I., GOMEZ RODRIGUEZ, M. and GUMMADI, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the* 26*th International Conference on World Wide Web. WWW'*17 1171–1180. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

[37] ZHANG, J. and BAREINBOIM, E. (2018). Fairness in decision-making—the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'*18*/IAAI'*18*/EAAI'*18. AAAI Press, Menlo Park, CA.