

Principal Fairness for Human and Algorithmic Decision-Making*

Kosuke Imai[†]

Zhichao Jiang[‡]

First Draft: May 11, 2020

This Draft: September 28, 2020

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. The key idea is 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 influenced by the decision. We introduce an axiomatic assumption that all groups are created equal once we account for relevant covariates. This assumption is motivated by a belief that protected attributes such as race and gender should not directly affect potential outcomes. Under this assumption, we show that principal fairness implies all three existing statistical fairness criteria, thereby resolving the previously recognized tradeoffs between them. Finally, we discuss how to empirically evaluate the principal fairness of a particular decision and the relationships between principal and counterfactual fairness criteria.

Keywords: algorithmic fairness, causal inference, potential outcomes, principal stratification

*We thank Hao Chen, Shizhe Chen, Christina Davis, Cynthia Dwork, Peng Ding, Robin Gong, Jim Greiner, Sharad Goel, 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.

[†]Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: <https://imai.fas.harvard.edu>

[‡]Assistant Professor, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst MA 01003.

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., Corbett-Davies and Goel, 2018; Chouldechova and Roth, 2020, 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, principal fairness incorporates causality into fairness through the key idea that one should not discriminate among individuals who would be similarly affected by the decision.¹

Consider a judge who decides, at a first appearance hearing, whether to detain or release an arrestee pending disposition of any criminal charges. 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 (Dwork et al., 2012), which demands that similar individuals should be treated similarly. The critical feature of principal fairness is that the similarity is measured based on the potential (both factual and counterfactual) outcomes.

1 Principal fairness

We begin by formally defining principal fairness. Let $D_i \in \{0, 1\}$ be the binary decision variable and $Y_i \in \{0, 1\}$ be the binary outcome variable of interest. Following the standard causal inference literature (e.g., Neyman, 1923; Fisher, 1935; Rubin, 1974; Holland, 1986), we use $Y_i(d)$ to denote the potential value of the outcome that would be realized if the decision is $D_i = d$. Then, the observed outcome can be written as $Y_i = Y_i(D_i)$.

Principal strata are defined as the joint potential outcome values, i.e., $R_i = (Y_i(1), Y_i(0))$, (Frangakis and Rubin, 2002). Since any causal effect can be written as a function of potential outcomes, e.g., $Y_i(1) - Y_i(0)$, each principal stratum represents how an individual would be affected by the decision with respect to the outcome of interest. When both the decision and outcome variables are binary, we have a total of four principal strata. Unlike the observed outcome, the potential outcomes, and hence principal strata, represent the pre-determined characteristics of individuals and are not affected by the decision. Moreover, since we do not observe $Y_i(1)$ and $Y_i(0)$ simultaneously for any individual, principal strata are not directly observable.

¹Principal fairness differs from counterfactual fairness, which is based on the potential outcomes with respect to a protected attribute rather than a decision itself (Kusner et al., 2017). Section 6 presents a detailed discussion.

		Group A		Group B	
		$Y_i(0) = 1$	$Y_i(0) = 0$	$Y_i(0) = 1$	$Y_i(0) = 0$
		Dangerous	Backlash	Dangerous	Backlash
$Y_i(1) = 1$	Detained ($D_i = 1$)	120	30	80	20
	Released ($D_i = 0$)	30	30	20	20
		Preventable	Safe	Preventable	Safe
$Y_i(1) = 0$	Detained ($D_i = 1$)	70	30	80	40
	Released ($D_i = 0$)	70	120	80	160

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

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. Let $D_i = 1$ ($D_i = 0$) represent the judge’s decision to detain (release) an arrestee, and $Y_i = 1$ ($Y_i = 0$) denote that the arrestee commits (does not commit) a new crime. Then, 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 might 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.²

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 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 , i.e., $\Pr(D_i \mid R_i, A_i) = \Pr(D_i \mid R_i)$.*

Note that principal fairness requires one to specify the outcome of interest as well as the attribute to be protected. As such, a decision-making mechanism that is fair with respect to one outcome (attribute) may not be fair with respect to another outcome (attribute).

²One could assume that an arrestee cannot 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 (see also Assumption 2 in Section 4).

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

Table 2: Observed data calculated from Table 1. None of the statistical fairness criteria given in Definition 2 is met.

Table 1 presents a numerical illustration, in which the detention rate is identical between Groups A and B within each principal stratum. For example, 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.

2 Comparison with the statistical fairness criteria

How does principal fairness differ from the existing definitions of statistical fairness? We consider the following criteria (see e.g., Corbett-Davies and Goel, 2018; Chouldechova and Roth, 2020, for reviews).

DEFINITION 2 (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. 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 example, 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 (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 (did not commit) a new crime, the same proportion of arrestees had been detained across racial groups.

Principal fairness differs from these statistical fairness criteria in that it accounts for the possibility of the decision affecting the outcome. In particular, although the accuracy criterion resembles principal fairness, the former conditions upon the observed rather than potential outcomes. Table 2 presents the observed data consistent with the numerical example shown in Table 1. Although this example satisfies principal fairness, it fails to meet the accuracy criterion as well as the other two statistical fairness criteria. For example, among those who committed a new crime, the detention

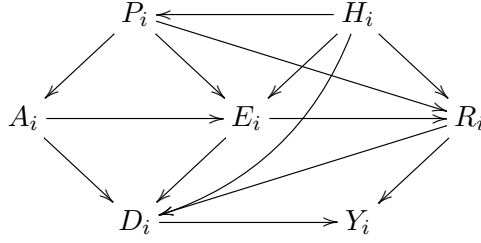


Figure 1: Direct acyclic graph for the relationship between the protected attribute A_i and principal strata R_i . In the criminal justice application, A_i represents the race of an arrestee, R_i is their risk category (safe, preventable, dangerous, and backlash), D_i represents the decision of judge, P_i represents parents’ characteristics including 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 indicator of committing a new crime, which is a deterministic function of judge’s decision D_i and risk category R_i . Assumption 1 holds with $\mathbf{W}_i = (H_i, P_i, S_i)$, i.e., $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$.

rate is much higher for Group A than Group B. 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.

3 All groups are created equal

How should we reconcile this tension between principal fairness and the existing statistical fairness criteria? The tradeoffs between different fairness criteria are not new. Chouldechova (2017) and Kleinberg et al. (2017) show that it is generally impossible to simultaneously satisfy the three statistical fairness criteria introduced in Definition 2. Below, we establish that the “*all groups are created equal*” assumption, which underlies the notion of principal fairness itself, can resolve these tradeoffs. To motivate this axiomatic assumption, we introduce a causal model in the context of criminal justice example. Under this model, the assumption implies that no racial group is inherently more dangerous than other groups once we account for relevant factors.

Figure 1 shows this causal model as a directed acyclic graph, where an arrow represents a causal relationship. The race of an arrestee, A_i , is affected by his/her parents’ characteristics including 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, and the historical processes such as slavery and Jim Crow laws, H_i , which also affect the parents’ characteristics, P_i .

Under this 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, experiences, and risk category of arrestee as well as 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 an arrow between these two variables. As a result, under this model, the arrestee’s race is conditionally independent of risk category, i.e., $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$, 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.

We now formalize and generalize this axiomatic assumption.³

ASSUMPTION 1 (ALL GROUPS ARE CREATED EQUAL) *There exist a set of covariates \mathbf{W}_i such that the principal strata are conditionally independent of the protected attribute given \mathbf{W}_i , i.e., $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$.*

In general, to ensure the validity of Assumption 1, the conditioning set \mathbf{W}_i should include the common causes of A_i and R_i as well as all the mediators on the causal pathway from A_i to R_i while excluding the covariates that are affected by both A_i and R_i . For example, conditioning on the outcome will violate the assumption. Since the likelihood of committing a new crime may be affected by both the race of arrestee (through the judge’s decision) and the risk category, this outcome variable represents a collider that induces the dependence between them when included in the conditioning set.⁴ This discussion demonstrates that a causal model is essential for guiding an appropriate choice of conditioning variables.

Assumption 1 motivates the consideration of principal fairness *conditional* on the same set of covariates \mathbf{W}_i , i.e., $\Pr(D_i \mid A_i, R_i, \mathbf{W}_i) = \Pr(D_i \mid R_i, \mathbf{W}_i)$. Once we account for these covariates, the assumption that no racial group is inherently dangerous suggests that a fair decision should not take into account the arrestee’s race within risk category. Most importantly, by conditioning on \mathbf{W}_i that satisfies Assumption 1, principal fairness resolves the tradeoffs between the competing definitions of statistical fairness.⁵ The following theorem shows that under Assumption 1, principal fairness implies all three statistical fairness criteria, conditional on the relevant covariates.

THEOREM 1 (PRINCIPAL FAIRNESS IMPLIES STATISTICAL FAIRNESS) *Suppose that Assumption 1 holds. Then, conditional on \mathbf{W}_i , principal fairness in Definition 1 implies all three statistical definitions of*

³Friedler et al. (2016) introduces a related “we’re all equal” assumption under a general but non-causal framework. The main difference between our assumption and theirs lies in the consideration of principal strata.

⁴As in the existing literature, we do not explicitly consider the possible racial bias in arrest. If the race of an individual affects the likelihood of their arrest, however, the analysis of arrestees may induce the dependence between A_i and R_i even conditional of \mathbf{W}_i (see Knox et al., 2019). If this is the case, one possible solution is to measure and condition on the variables that mediate the effect of A_i or that of the arrest.

⁵Assumption 1 also eliminates the problem of infra-marginality discussed by Corbett-Davies and Goel (2018) because the distribution of potential outcomes is identical between protected groups.

fairness given in Definition 2. That is, under Assumption 1, $\Pr(D_i \mid R_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid R_i, \mathbf{W}_i)$ implies $\Pr(D_i \mid \mathbf{W}_i, A_i) = \Pr(D_i \mid \mathbf{W}_i)$, $\Pr(Y_i \mid D_i, \mathbf{W}_i, A_i) = \Pr(Y_i \mid D_i, \mathbf{W}_i)$, and $\Pr(D_i \mid Y_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid Y_i, \mathbf{W}_i)$.

Proof is given in Appendix S1.1. Theorem 1 emphasizes the essential role of conditioning on appropriate covariates in fairness criteria. The result also highlights a primary difficulty of various statistical definitions of fairness including principal fairness — criteria that hold conditionally may not hold marginally or vice versa.

4 Equivalence between principal fairness and statistical fairness

Theorem 1 shows that under Assumption 1, principal fairness represents a stronger notion of fairness than the existing statistical fairness definitions. We next show that principal definition is equivalent to these statistical fairness criteria under the additional assumption of monotonicity.

ASSUMPTION 2 (MONOTONICITY)

$$Y_i(1) \leq Y_i(0)$$

for all i .

Assumption 2 is plausible in many applications. In our criminal justice example, the assumption implies that being detained does not make it easier to commit a new crime than being released. The following theorem establishes the equivalence relationship between principal fairness and statistical fairness under this additional assumption.

THEOREM 2 (EQUIVALENCE BETWEEN PRINCIPAL FAIRNESS AND STATISTICAL FAIRNESS) *Suppose that Assumptions 1 and 2 hold. Then, conditional on \mathbf{W}_i , principal fairness is equivalent to the three statistical fairness criteria given in Definition 2.*

Proof is given in Appendix S1.2.

5 Empirical evaluation of principal fairness

Since principal strata are not directly observable, an additional assumption is required for empirically evaluating the principal fairness of particular decision. In particular, we must identify the conditional distribution of the decision given the principal stratum and some observed covariates \mathbf{X}_i , i.e., $\Pr(D_i \mid R_i, \mathbf{X}_i)$. We introduce the following unconfoundedness assumption widely used in the causal inference literature.

ASSUMPTION 3 (UNCONFOUNDEDNESS) $Y_i(d) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$.

Assumption 3 holds if \mathbf{X}_i contains all the information used for decision-making which may include the protected attribute. 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 (VanderWeele and Shpitser, 2011). The next theorem shows that under Assumptions 2 and 3, the evaluation of principal fairness reduces to the estimation of regression function, $\Pr(Y_i = 1 \mid D_i, X_i)$.

THEOREM 3 (IDENTIFICATION) *Under Assumptions 2 and 3, we have*

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

In Appendix S1.3, we prove this theorem and generalize it to the evaluation of principal fairness conditional on relevant covariates \mathbf{W}_i .

6 Comparison with counterfactual fairness

As shown above, the key difference between principal fairness and existing statistical fairness criteria is that the former considers how decisions affect individuals. In the literature, *counterfactual fairness* represents one prominent fairness criterion that similarly builds upon the causal inference framework (Kusner et al., 2017). According to this criterion, 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 is 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 is based on the statistical 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, i.e., $D_i(a) = D_i(a')$, which 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. Ensuring group-level fairness may not guarantee individual-level fairness, and vice versa.

Second, the covariate adjustment requires care for both principal and counterfactual fairness criteria. For principal fairness, choosing an appropriate set of conditioning variables resolves the conflict between various definitions of group-level fairness. Yet, the challenge is how to appropriately choose conditioning variables such that potential outcomes become statistically independent of the protected attribute (i.e., Assumption 1 holds). For counterfactual fairness, one cannot simply condition on covariates that are affected by the protected attribute because this would induce a post-treatment bias (see e.g., Kilbertus et al., 2017; Knox et al., 2019). To address this issue, researchers have considered path-specific effects through the framework of causal mediation analysis (e.g., Nabi and Shpitser, 2018; Chiappa, 2019). In such an analysis, a key question for analysts is which mediators should be included. For both principal and counterfactual fairness, therefore, a careful consideration of underlying causal assumptions is required for covariate adjustment.

Finally, while principal fairness considers the potential outcomes with respect to different decisions, counterfactual fairness is based on the potential outcomes with respect to different values of 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., Holland, 1986). In addition, causal mediation analysis relies upon the so-called “cross-world” independence assumption that cannot be satisfied even when the randomization of mediators is possible (Richardson and Robins, 2013). Addressing these issues often requires one to consider alternative causal quantities such as the causal effects of perceived attributes (Greiner and Rubin, 2011) and stochastic intervention of mediators (Jackson and VanderWeele, 2018). In contrast, principal fairness avoids these conceptual and identifiability issues and can be evaluated under the widely used unconfoundedness assumption (see Section 5).

7 Concluding Remarks

To assess the fairness of human and algorithmic decision-making, we must consider how the decisions themselves affect individuals. This requires the notion of fairness to be placed in the causal inference framework. In ongoing work, we extend principal fairness to the common settings, in which humans make decisions partly based on the recommendations produced by algorithms (Imai et al., 2020).

Since human decision-makers rather than algorithms ultimately impact individuals, the fairness of algorithmic recommendations critically depends on how they can improve the fairness of human decisions. We empirically examine this issue through the experimental evaluation of the pre-trial risk assessment instrument widely used in the US criminal justice system.

Finally, 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. Another possible direction is the extension of principal fairness to a dynamic system. As pointed out by D’Amour et al. (2020) and Chouldechova and Roth (2020), 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.

References

- Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 7801–7808.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2), 153–163.
- Chouldechova, A. and A. Roth (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM* 63(5), 82–89.
- Corbett-Davies, S. and S. Goel (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. Technical report, arXiv:1808.00023.
- D’Amour, A., H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern (2020, January). 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*, pp. 525–534.
- Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). Fairness through awareness. In *ITCS ’12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226.
- Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver and Boyd.

- Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics* 58(1), 21–29.
- Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian (2016). On the (im)possibility of fairness. Technical report, arXiv:1609.07236.
- Greiner, D. J. and D. B. Rubin (2011, August). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics* 93(3), 775–785.
- Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81, 945–960.
- Imai, K., Z. Jiang, D. J. Greiner, R. Halen, and S. Shin (2020). Experimental evaluation of computer-assisted human decision-making: Application to pretrial risk assessment instrument. Technical report, Harvard University.
- Jackson, J. W. and T. J. VanderWeele (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* 29(6), 825–835.
- Kilbertus, N., M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666.
- Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitrou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43, pp. 1–23.
- Knox, D., W. Lowe, and J. Mummolo (2019). Administrative records mask racially biased policing. *American Political Science Review*, 1–19.
- Kusner, M., J. Loftus, C. Russell, and R. Silva (2017). Counterfactual fairness. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA.
- Nabi, R. and I. Shpitser (2018). Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science* 5, 465–480.

- Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical report, University of Washington.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66, 688–701.
- VanderWeele, T. J. and I. Shpitser (2011). A new criterion for confounder selection. *Biometrics* 67(4), 1406–1413.

Supplementary Appendix

S1 Proofs

S1.1 Proof of Theorem 1

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

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

where the third equality follows from principal fairness and Assumption 1. Similarly, we can show

$$\Pr(D_i = d, Y_i = y \mid \mathbf{W}_i, A_i) = \Pr(D_i = d, Y_i = y \mid \mathbf{W}_i) \quad (\text{S1})$$

for $d, y = 0, 1$. Therefore, we have

$$\begin{aligned}
\Pr(D_i \mid \mathbf{W}_i, A_i) &= \Pr(D_i, Y_i = 1 \mid \mathbf{W}_i, A_i) + \Pr(D_i, Y_i = 0 \mid \mathbf{W}_i, A_i) \\
&= \Pr(D_i, Y_i = 1 \mid \mathbf{W}_i) + \Pr(D_i, Y_i = 0 \mid \mathbf{W}_i) \\
&= \Pr(D_i \mid \mathbf{W}_i),
\end{aligned} \quad (\text{S2})$$

and

$$\begin{aligned}
\Pr(Y_i \mid \mathbf{W}_i, A_i) &= \Pr(D_i = 1, Y_i \mid \mathbf{W}_i, A_i) + \Pr(D_i = 0, Y_i \mid \mathbf{W}_i, A_i) \\
&= \Pr(D_i = 1, Y_i \mid \mathbf{W}_i) + \Pr(D_i = 0, Y_i \mid \mathbf{W}_i) \\
&= \Pr(Y_i \mid \mathbf{W}_i).
\end{aligned} \quad (\text{S3})$$

Then, from (S1) and (S2), we have $\Pr(Y_i \mid D_i, \mathbf{W}_i, A_i) = \Pr(Y_i \mid D_i, \mathbf{W}_i)$, and from (S1) and (S3), we have $\Pr(D_i \mid Y_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid Y_i, \mathbf{W}_i)$. \square

S1.2 Proof of Theorem 2

We need the following lemma.

LEMMA S1 *Suppose Assumption 2 holds. Then, for any covariates \mathbf{V}_i , we have*

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

Proof of Lemma S1. Under Assumption 2, we can write

$$\begin{aligned}\Pr(R_i = (0, 0) \mid \mathbf{V}_i, A_i) &= \Pr(Y_i(0) = 0 \mid \mathbf{V}_i, A_i), \\ \Pr(R_i = (0, 1) \mid \mathbf{V}_i, A_i) &= \Pr(Y_i(0) = 1 \mid \mathbf{V}_i, A_i) - \Pr(Y_i(1) = 1 \mid \mathbf{V}_i, A_i), \\ \Pr(R_i = (1, 1) \mid \mathbf{V}_i, A_i) &= \Pr(Y_i(1) = 1 \mid \mathbf{V}_i, A_i).\end{aligned}\tag{S4}$$

Therefore, we obtain

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

and

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

□

We now prove Theorem 2. From Theorem 1, it suffices to show that the three statistical fairness criteria imply principal fairness. From the three statistical fairness criteria, we have

$$\Pr(D_i, Y_i \mid \mathbf{W}_i, A_i) = \Pr(D_i, Y_i \mid \mathbf{W}_i).\tag{S5}$$

Applying Lemma S1 with $\mathbf{V}_i = \mathbf{W}_i$, we have

$$\Pr(D_i = 1 \mid R = (0, 0), \mathbf{W}_i, A_i) = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0, 0) \mid \mathbf{W}_i, A_i)},\tag{S6}$$

$$\Pr(D_i = 1 \mid R = (0, 1), \mathbf{W}_i, A_i) = \frac{\Pr(Y_i = 0 \mid \mathbf{W}_i, A_i) - \Pr(R_i = (0, 0) \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0, 1) \mid \mathbf{W}_i, A_i)},\tag{S7}$$

$$\Pr(D_i = 1 \mid R = (1, 1), \mathbf{W}_i, A_i) = \frac{\Pr(D_i = 1, Y_i = 1 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (1, 1) \mid \mathbf{W}_i, A_i)}.\tag{S8}$$

From Assumption 1 and (S5), all terms on the right-hand sides of (S6), (S7), (S8) do not depend on A_i . As a result, we have $\Pr(D_i \mid R_i, \mathbf{W}_i, A_i) = \Pr(D_i \mid R_i, \mathbf{W}_i)$. □

S1.3 Proof of Theorem 3

Applying Lemma S1 with $\mathbf{V}_i = \emptyset$, we have

$$\Pr(D_i = 1 \mid R = (0, 0), A_i) = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid A_i)}{\Pr(R_i = (0, 0) \mid A_i)}, \quad (\text{S9})$$

$$\Pr(D_i = 1 \mid R = (0, 1), A_i) = \frac{\Pr(Y_i = 0 \mid A_i) - \Pr(R_i = (0, 0) \mid A_i)}{\Pr(R_i = (0, 1) \mid A_i)}, \quad (\text{S10})$$

$$\Pr(D_i = 1 \mid R = (1, 1), A_i) = \frac{\Pr(D_i = 1, Y_i = 1 \mid A_i)}{\Pr(R_i = (1, 1) \mid A_i)}. \quad (\text{S11})$$

From (S4), we have

$$\begin{aligned} \Pr(R_i = (0, 0) \mid A_i) &= \Pr(Y_i(0) = 0 \mid A_i) \\ &= \mathbb{E}\{\Pr(Y_i(0) = 0 \mid \mathbf{X}_i) \mid A_i\} \\ &= \mathbb{E}\{\Pr(Y_i = 0 \mid D_i = 0, \mathbf{X}_i) \mid A_i\}, \end{aligned}$$

where the second equality follows from the law of total probability and the third equality follows from Assumption 3. Similarly, we can obtain

$$\Pr(R_i = (0, 1) \mid A_i) = \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) \mid A_i\} - \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i) \mid A_i\},$$

and

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

Plugging the expressions for $\Pr(R_i \mid A_i)$ into (S9) to (S11) yields the formulas in Theorem 3. \square

We generalize Theorem 3 to the identification of $\Pr(D_i \mid R, \mathbf{W}_i, A_i)$. Applying Lemma S1 with $\mathbf{V}_i = \mathbf{W}_i$, we have

$$\Pr(D_i = 1 \mid R = (0, 0), \mathbf{W}_i, A_i) = 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0, 0) \mid \mathbf{W}_i, A_i)}, \quad (\text{S12})$$

$$\Pr(D_i = 1 \mid R = (0, 1), \mathbf{W}_i, A_i) = \frac{\Pr(Y_i = 0 \mid A_i) - \Pr(R_i = (0, 0) \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0, 1) \mid \mathbf{W}_i, A_i)}, \quad (\text{S13})$$

$$\Pr(D_i = 1 \mid R = (1, 1), \mathbf{W}_i, A_i) = \frac{\Pr(D_i = 1, Y_i = 1 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (1, 1) \mid \mathbf{W}_i, A_i)}. \quad (\text{S14})$$

Similarly, under Assumption 3, we have

$$\begin{aligned} \Pr(R_i = (0, 0) \mid \mathbf{W}_i, A_i) &= \mathbb{E}\{\Pr(Y_i = 0 \mid D_i = 0, \mathbf{X}_i) \mid \mathbf{W}_i, A_i\}, \\ \Pr(R_i = (0, 1) \mid \mathbf{W}_i, A_i) &= \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = 0, \mathbf{X}_i) \mid \mathbf{W}_i, A_i\} - \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i) \mid \mathbf{W}_i, A_i\}, \\ \Pr(R_i = (1, 1) \mid \mathbf{W}_i, A_i) &= \mathbb{E}\{\Pr(Y_i = 1 \mid D_i = 1, \mathbf{X}_i) \mid \mathbf{W}_i, A_i\}, \end{aligned}$$

where we assume \mathbf{X}_i contains (\mathbf{W}_i, A_i) . Plugging these into (S12) to (S14) yields

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

\square