Principal Fairness for Human and Algorithmic Decision-Making*

Kosuke Imai[†] Zhichao Jiang[‡]

First Draft: May 11, 2020
This Draft: May 22, 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 motivate principal fairness by the belief that all people are created equal, implying that the potential outcomes should not depend on protected attributes such as race and gender once we adjust for relevant covariates. 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.

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

*We thank Christina Davis, Cynthia Dwork, Jim Greiner, Gary King, and Jamie Robins for comments and discussions.

[†]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 based on the 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, i.e., 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. It is important to note that 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.

[1] Principal fairness differs from counterfactual fairness, which considers the potential outcomes with respect to a protected attribute rather than a decision itself (Kusner et al., 2017).

		Group A		Group B	
		$Y_i(0)=1$	$Y_i(0)=0$	$Y_i(0)=1$	$Y_i(0)=0$
		Dangerous	Preventable	Dangerous	Preventable
$Y_i(1)=1$	Detained ($D_i=1$)	120	70	80	100
	Released ($D_i=0$)	30	70	20	100
		Backlash	Safe	Backlash	Safe
$Y_i(1)=0$	Detained ($D_i=1$)	30	30	20	40
	Released ($D_i=0$)	30	120	20	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 — released or detained — 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.

Principal fairness implies that the decision is independent of protected attribute within each principal stratum. 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).

Table 1 presents a numerical illustration in the context of the criminal justice example, 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 the two groups, while it is 20% for both groups within the "safe" stratum. Indeed, the decision is independent of group membership

	Group A		Group B	
	Detained	Released	Detained	Released
$Y_i = 1$	190	60	180	140
$Y_i = 0$	60	190	60	260

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

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 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 people 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 show how the assumption, which underlies the notion of principal fairness, can be used to resolve these tradeoffs.

Principal fairness is motivated by the belief that all people are created equal with respect to the potential outcomes of interest. Recall that the principal strata represent the pre-determined characteristics of individuals, which are not affected by the decision. In our application, for example, we may justify principal fairness by assuming that no racial group has an innate tendency to be dangerous. In other words, once we adjust for economic and social factors that are relevant for committing a new crime, all racial groups should have the same proportion of dangerous arrestees. Similarly, the remaining three principal strata should be equally distributed across the racial groups conditional on these covariates. We formalize this assumption as follows.

Assumption 1 (All people 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.,* $R_i \perp\!\!\!\perp A_i \mid \mathbf{W}_i$.

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 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 shows that if all people are assumed to be created equal, principal fairness can resolve the tradeoffs between the competing definitions of statistical fairness.

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, it is important to discuss an additional assumption 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 that is widely invoked 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.[2] 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 $\mathbf{W}_i$.

[2] 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).

6 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. 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.

References

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.

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.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association 81*, 945–960.

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.

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.

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.

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

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

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

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

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

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

and

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

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

S1.2 Proof of Theorem 2

We need the following lemma.

Lemma S1 *Suppose Assumption 2 holds. Then, for any covariates $\boldsymbol{V}_i$, we have*

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

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

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

Therefore, we obtain

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

and

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

□

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 $\boldsymbol{V}_i = \mathbf{W}_i$, 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 $\boldsymbol{V}_i = \emptyset$, we have

$$\begin{aligned}
\Pr(D_i = 1 \mid R = (0,0), A_i) &= 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid A_i)}{\Pr(R_i = (0,0) \mid A_i)}, && \text{(S9)}\\
\Pr(D_i = 1 \mid R = (0,1), A_i) &= \frac{\Pr(Y_i = 0 \mid A_i) - \Pr(R_i = (0,0) \mid A_i)}{\Pr(R_i = (0,1) \mid A_i)}, && \text{(S10)}\\
\Pr(D_i = 1 \mid R = (1,1), A_i) &= \frac{\Pr(D_i = 1, Y_i = 1 \mid A_i)}{\Pr(R_i = (1,1) \mid A_i)}. && \text{(S11)}
\end{aligned}$$

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. □

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

$$\begin{aligned}
\Pr(D_i = 1 \mid R = (0,0), \mathbf{W}_i, A_i) &= 1 - \frac{\Pr(D_i = 0, Y_i = 0 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0,0) \mid \mathbf{W}_i, A_i)}, && \text{(S12)}\\
\Pr(D_i = 1 \mid R = (0,1), \mathbf{W}_i, A_i) &= \frac{\Pr(Y_i = 0 \mid A_i) - \Pr(R_i = (0,0) \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (0,1) \mid \mathbf{W}_i, A_i)}, && \text{(S13)}\\
\Pr(D_i = 1 \mid R = (1,1), \mathbf{W}_i, A_i) &= \frac{\Pr(D_i = 1, Y_i = 1 \mid \mathbf{W}_i, A_i)}{\Pr(R_i = (1,1) \mid \mathbf{W}_i, A_i)}. && \text{(S14)}
\end{aligned}$$

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

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

□