TECHNICAL REPORT
R-92
Feb, 2023

Reconciling Predictive and Statistical Parity: A Causal Approach

Drago Plečko [1] **Elias Bareinboim** [1]

Abstract

Since the rise of fair machine learning as a distinct sub-field of research, many different notions on how to quantify and measure discrimination have been proposed. Some of these notions, however, have been shown to be mutually incompatible. Such findings make it appear that numerous different kinds of fairness exist, thereby making the consensus on the appropriate notion of fairness harder to reach and hindering the progress in applying the tools of fair ML in practice. In this paper, we investigate one of these key impossibility results, namely, between the notions of statistical and predictive parity. In particular, we look at the causal decompositions of fairness measures associated with statistical and predictive parity, and obtain a novel insight into how these criteria are related through the legal doctrines of disparate treatment, disparate impact, and the notion of business necessity. Our results show that through a more careful analysis using causal lenses, the notions of statistical and predictive parity are not really mutually exclusive, but complementary and spanning a spectrum of fairness notions through the concept of business necessity. Finally, we demonstrate the importance of our findings on a real world example.

1. Introduction

As society increasingly relies on AI-based tools, an ever larger number of decisions that were once made by humans are now delegated to automated systems, and this trend is likely to accelerate in the coming years. Such automated systems may exhibit discrimination based on gender, race, religion, or other sensitive attributes, as witnessed by various examples in criminal justice (Angwin et al., 2016), facial recognition (Harwell, 2019; Buolamwini & Gebru, 2018), targeted advertising (Detrixhe & Merrill, 2019), and medical treatment allocation (Rajkomar et al., 2018), just to name a few.

In light of these challenges, a large amount of effort has been invested in attempts to detect and quantify undesired discrimination based on society's current ethical standards, and then design learning methods capable of removing such unfairness from future predictions and decisions. In this process, many different notions on how to quantify discrimination have been proposed, and the current literature is abundant with different fairness metrics, some of which are mutually incompatible (Corbett-Davies & Goel, 2018). The incompatibility of different measures can create a serious obstacle for practitioners since choosing among them, even for the system designer, is usually a non-trivial task. Moreover, it often may be desirable to simultaneously incorporate principles encoded in various different fairness measures, which may seem an impossible task.

In this paper, we focus on two prominent notions of fairness. The first notion is that of statistical or demographic parity (Darlington, 1971), which is often associated with the measure known as the parity gap or total variation (TV). The second notion we consider is that of predictive parity (Chouldechova, 2017), which can be assessed using a measure we call predictive parity measure (PPM). In particular, we leverage the known causal decompositions of the TV measure (Zhang & Bareinboim, 2018b; Plečko & Bareinboim, 2022), and obtain a new decompositions result for the PP measure. These results allow us to draw an important connection between statistical and predictive parity on the one hand, and the legal notions of disparate impact, disparate treatment, and business necessity (BN) on the other.

The disparate treatment doctrine enforces the equality of treatment of different groups, prohibiting the use of the protected attribute (e.g., race) during the decision process. One of the legal formulations for proving disparate treatment is that "a similarly situated person who is not a member of the protected class would not have suffered the same fate" (Barocas & Selbst, 2016). Disparate treatment is commonly associated with the notion of direct discrimination in the causal fairness literature. The second doctrine, known as disparate impact, focuses on *outcome fairness*, namely, the equality of outcomes among protected groups. Disparate im-

[1]Department of Computer Science, Columbia University, New York, United States. Correspondence to: Drago Plečko <dp3144@columbia.edu>.

pact discrimination occurs if a facially neutral practice has an adverse impact on members of the protected group, including cases where discrimination is unintended or implicit. In practice, the law may not necessarily prohibit the usage of all characteristics correlated with the protected attribute, due to their relevance to the business itself, legally known as "business necessity" or "job-relatedness". Therefore, some of the variables may be used to distinguish between individuals, even if they are associated with the protected attribute (Kilbertus et al., 2017). From a causal perspective, the disparate impact doctrine is closely related to indirect forms of discrimination, and taking into account considerations of business necessity is the essence of this doctrine (Barocas & Selbst, 2016).

As we demonstrate both intuitively and formally later on in the paper, the notions of statistical parity and predictive parity are relevant when assessing the legal doctrines of discrimination. In particular, we show that whenever a causal pathway between the protected attribute X and the predictor $\widehat{Y}$ does not fall under business necessity, then the causal effect transmitted along this pathway should equal 0, written (for now informally) as $\mathcal{C}(X, \widehat{Y}) = 0$, with $\mathcal{C}$ indicating the pathway. This principle represents a causal analogue of the notion of statistical parity. However, when the pathway does fall under business necessity, then the transmitted causal effect need not equal 0, i.e., $\mathcal{C}(X, \widehat{Y}) \neq 0$ may be allowed. In this case, however, the transmitted causal effect should not take an arbitrary value, but rather equal the *transmitted effect from X to the original outcome Y (observed in the real world) along the same pathway*, written as $\mathcal{C}(X, \widehat{Y}) = \mathcal{C}(X, Y)$. The latter notion, as we argue in the text, represents a causal analogue of the notion of predictive parity. Interestingly, the importance of this notion has been voiced by the legal community (Grimmelmann & Westreich, 2016), although not in a formal mathematical language.

This unification of the principles behind statistical and predictive parity through the concept of business necessity has the potential to bridge the gap between the two notions and improve the current state-of-the-art, by providing an argument against the seemingly discouraging impossibility result between the two notions. The practitioner is no longer faced with a dichotomy of choosing between statistical or predictive parity, but rather faces a spectrum of different fairness notions determined by the choice of the business necessity set (usually fixed through societal consensus or legal requirements, see Figure 1). On the one extreme of this spectrum lies statistical parity, in the case when none of the causal pathways fall under business necessity, while on the other lies predictive parity, in the case when all of the causal pathways fall under business necessity.

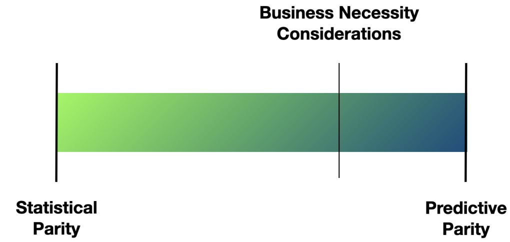

Figure 1. The spectrum between statistical and predictive parity spanned by the considerations of business necessity.

1.1. Organization & Contributions

The manuscript is organized as follows. In Section 2, we introduce important preliminary notions and results that are needed for understanding the manuscript. In particular, in Section 2.1 we introduce the key notions from causal inference (Pearl, 2000) that are needed for our framework, and in Section 2.2 we formally introduce statistical and predictive parity notions, together with the impossibility result that separates them. This section also represents where the current state-of-the-art understanding of these notions lies. In Section 3, we analyze the causal decompositions of the above fairness notions and obtain novel insights into how the notions are in fact complementary. In Section 3.1, we introduce a formal procedure that shows how to assess the legal doctrines of discrimination by leveraging both the concepts of causal predictive parity and causal statistical parity. In Section 4, we empirically demonstrate the value of our approach in the context of criminal justice with the COMPAS dataset (Angwin et al., 2016).

Our key formal contributions are the following:

- In Theorem 3.4 we provide what is, to the best of our knowledge, the first causal decomposition of the predictive parity measure. This leads naturally to the definition of causal predictive parity (Definition 3.5),
- We introduce a formal procedure (Alg. 1) for evaluating if a classifier satisfies the desired notions of causal statistical parity and causal predictive parity, which provides a unified framework for incorporating desiderata from both predictive and statistical parity.

2. Background

In this section we introduce the key preliminary results and notions in causal inference (Section 2.1) and fair machine learning (Section 2.2).

2.1. Causal Notions

In this manuscript, we use the semantics of structural causal models (SCMs), which are defined as follows:

Definition 2.1 (Structural Causal Model (SCM) (Pearl, 2000))**.** A structural causal model (SCM) is a 4-tuple $\langle V, U, \mathcal{F}, P(u)\rangle$, where

1. U is a set of exogenous variables, also called background variables, that are determined by factors outside the model;
2. $V = \{V_1, ..., V_n\}$ is a set of endogenous (observed) variables, that are determined by variables in the model (i.e. by the variables in $U \cup V$);
3. $\mathcal{F} = \{f_1, ..., f_n\}$ is the set of structural functions determining V, $v_i \leftarrow f_i(\mathrm{pa}(v_i), u_i)$, where $\mathrm{pa}(V_i) \subseteq V \setminus V_i$ and $U_i \subseteq U$ are the functional arguments of f_i;
4. $P(u)$ is a distribution over the exogenous variables U.

An important definition on top of the SCM is that of a submodel:

Definition 2.2 (Submodel (Pearl, 2000))**.** Let $\mathcal{M}$ be a structural causal model, X a set of variables in V, and x a particular value of X. A submodel $\mathcal{M}_x$ (of $\mathcal{M}$) is a 4-tuple:

$$\mathcal{M}_x = \langle V, U, \mathcal{F}_x, P(u)\rangle \tag{1}$$

where

$$\mathcal{F}_x = \{f_i : V_i \notin X\} \cup \{X \leftarrow x\}, \tag{2}$$

and all other components are preserved from $\mathcal{M}$.

The SCM $\mathcal{M}_x$ is obtained from $\mathcal{M}$ by replacing all equations in $\mathcal{F}$ related to variables X with equations that set X to a specific value x. In the fairness context, we might be interested in submodels in which the protected attribute X is set to a fixed value x. This motivates the notion of a potential response (also known as potential outcome):

Definition 2.3 (Potential Response (Pearl, 2000))**.** Let X and Y be two sets of variables in Y and $u \in \mathcal{U}$ be a unit. The potential response $Y_x(u)$ is defined as the solution for Y of the set of equations $\mathcal{F}_x$ with respect to SCM $\mathcal{M}$. That is, $Y_x(u)$ denotes the solution of Y in the submodel $\mathcal{M}_x$ of $\mathcal{M}$.

Based on the notion of a potential response, one can further define the notions of counterfactual and factual contrasts, given by:

Definition 2.4 (Contrasts (Plečko & Bareinboim, 2022))**.** Given a SCM $\mathcal{M}$, a contrast $\mathcal{C}$ is any quantity of the form

$$\mathcal{C}(C_0, C_1, E_0, E_1) = \mathbb{E}[y_{C_1} \mid E_1] - \mathbb{E}[y_{C_0} \mid E_0], \tag{3}$$

where E_0, E_1 are observed (factual) clauses and C_0, C_1 are counterfactual clauses to which the outcome Y responds. Furthermore, whenever

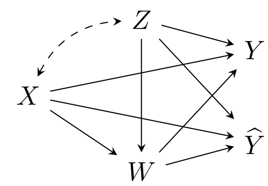

Figure 2. Standard Fairness Model.

(a) $E_0 = E_1$, the contrast $\mathcal{C}$ is said to be counterfactual;

(b) $C_0 = C_1$, the contrast $\mathcal{C}$ is said to be factual.

For instance, the contrast $(C_0 = \{x_0\}, C_1 = \{x_1\}, E_0 = \emptyset, E_1 = \emptyset)$ corresponds to the *average treatment effect (ATE)* $\mathbb{E}[y_{x_1} - y_{x_0}]$. Similarly, the contrast $(C_0 = \{x_0\}, C_1 = \{x_1\}, E_0 = \{x_0\}, E_1 = \{x_0\})$ corresponds to the *effect of treatment on the treated (ETT)* $\mathbb{E}[y_{x_1} - y_{x_0} \mid x_0]$. Many other important causal quantities can be represented as contrasts, as exemplified in later sections.

An important further observation is that the knowledge contained in the SCM is almost never available to the practitioner. Therefore, instead of trying to learn the SCM, a common way of encoding assumptions about the underlying SCM is through an object called a causal diagram. We describe below the constructive procedure that allows one to articulate a diagram from a coarse understanding of the SCM.

Definition 2.5 (Causal Diagram (Pearl, 2000; Bareinboim et al., 2022))**.** Let $\mathcal{M} = \langle V, U, \mathcal{F}, P(u)\rangle$ be an SCM. A graph $\mathcal{G}$ is said to be a *causal diagram* (of $\mathcal{M}$) if:

1. there is a vertex for every endogenous variable $V_i \in V$,
2. there is an edge $V_i \rightarrow V_j$ if V_i appears as an argument of $f_j \in \mathcal{F}$,
3. there is a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \subset U$ are correlated or the corresponding functions f_i, f_j share some $U_{ij} \in U$ as an argument.

Throughout this manuscript, we assume the causal diagram $\mathcal{G}_{\text{SFM}}$ known as the standard fairness model (SFM) (Plečko & Bareinboim, 2022) over endogenous variables $\{X, Z, W, Y, \widehat{Y}\}$ shown in Figure 2, where the nodes represent:

- the *protected attribute*, labeled X (e.g., gender, race, religion), which is assumed to be binary,
- the set of *confounding* variables Z, which are not causally influenced by the attribute X (e.g., demographic information, zip code),

- the set of *mediator* variables W that are possibly causally influenced by the attribute (e.g., educational level or other job-related information),
- the *outcome* variable Y (e.g., GPA, salary),
- the *predictor* of the outcome $\widehat{Y}$ (e.g., predicted GPA, predicted salary).

We next introduce the key notions and results from the fair ML literature needed for our discussion.

2.2. Statistical & Predictive Parity Notions

The notion of statistical parity (also known as demographic parity), going back to (Darlington, 1971), is defined as follows: [1]

Definition 2.6 (Statistical Parity (Darlington, 1971))**.** Let X be the protected attribute, and let $\widehat{Y}$ be the predictor of the outcome. We say that $\widehat{Y}$ satisfies statistical parity (SP) with respect to X if

$$P(\widehat{y} \mid x_1) = P(\widehat{y} \mid x_0). \tag{4}$$

The SP criterion can also be written as a conditional independence statement,

$$\widehat{Y} \perp\!\!\!\perp X. \tag{5}$$

Finally, the total variation (TV) measure is defined as:

$$\text{TV}_{x_0,x_1}(\widehat{y}) = P(\widehat{y} \mid x_1) - P(\widehat{y} \mid x_0). \tag{6}$$

The intuition behind the notion of statistical parity is that the predictor $\widehat{Y}$ should not contain information about the protected attribute X.

In contrast to this idea, the notion of predictive parity, introduced by (Chouldechova, 2017), is defined as follows:

Definition 2.7 (Predictive Parity (Chouldechova, 2017))**.** Let X be the protected attribute, and Y the outcome. Let $\widehat{Y}$ be the predictor of Y. We say that $\widehat{Y}$ satisfies predictive parity (PP) with respect to X, Y if

$$P(y \mid x_1, \widehat{y}) = P(y \mid x_0, \widehat{y}) \quad \forall \widehat{y}. \tag{7}$$

Alternatively, the PP criterion can also be written as a conditional independence statement, i.e.,

$$Y \perp\!\!\!\perp X \mid \widehat{Y}. \tag{8}$$

Finally, define the predictive parity measure to be

$$\text{PPM}_{x_0,x_1}(y \mid \widehat{y}) = P(y \mid x_1, \widehat{y}) - P(y \mid x_0, \widehat{y}). \tag{9}$$

[1] We do not consider here the quantity known as equality of odds (EO) (Hardt et al., 2016). Still, for further discussions on its causal interpretation, refer to (Zhang & Bareinboim, 2018a).

An important known result when trying to understand the PP criterion is the following:

Proposition 2.8 (PP and Efficient Learning)**.** *Let $\mathcal{M}$ be an SCM compatible with the Standard Fairness Model (SFM). Suppose that the predictor $\widehat{Y}$ is based on the features X, Z, W. Suppose also that $\widehat{Y}$ is an efficient learner, meaning that:*

$$\widehat{Y}(x, z, w) = P(y \mid x, z, w). \tag{10}$$

Then, it follows that $\widehat{Y}$ satisfies predictive parity w.r.t. X and Y.

Proof. Notice that for any $X = x$, $P(y \mid x, \widehat{y})$ equals

$$\sum_{\substack{z,w:\\ \widehat{Y}(x,z,w)=\widehat{y}}} P(y \mid x, z, w, \widehat{y}) P(z, w \mid x, \widehat{y}) \tag{11}$$

$$= \sum_{\substack{z,w:\\ \widehat{Y}(x,z,w)=\widehat{y}}} P(y \mid x, z, w) P(z, w \mid x, \widehat{y}) \tag{12}$$

$$= \widehat{y} \sum_{\substack{z,w:\\ \widehat{Y}(x,z,w)=\widehat{y}}} P(z, w \mid x, \widehat{y}) = \widehat{y}. \tag{13}$$

The first step (Eq. 11) follows from the law of total probability, the second (Eq. 12) from noting that $\widehat{Y} \perp\!\!\!\perp Y \mid X, Z, W$, and the third (Eq. 13) from the efficiency of the learner (Eq. 10). Therefore, it follows that $P(y \mid x_1, \widehat{y}) = P(y \mid x_0, \widehat{y})$, meaning that $\widehat{Y}$ satisfies PP. □

Proposition 2.8 shows that PP is expected to hold for an efficient learner $\widehat{Y}$. In some sense, this means that $\widehat{Y}$ should "exhaust", or capture all the variations coming into the outcome Y in the current real world. In particular, $\widehat{Y}$ should also capture all the variations of X coming into Y.

2.3. Statistical & Predictive Parity: An Impossibility?

We now draw the attention of the reader to the stark contrast between the SP notion and the PP notion introduced in the previous section. For the TV measure to be 0, the predictor $\widehat{Y}$ should not be associated at all with the attribute X, whereas for PP to hold, $\widehat{Y}$ should contain all variations coming from X. Perhaps unsurprisingly after this discussion, the following theorem holds:

Theorem 2.9 (SP and PP impossibility (Kleinberg et al., 2016))**.** *The fairness criteria of predictive parity and statistical parity,*

$$Y \perp\!\!\!\perp X \mid \widehat{Y}, \tag{14}$$

$$\widehat{Y} \perp\!\!\!\perp X, \tag{15}$$

are mutually exclusive except in degenerate cases, when $Y \perp\!\!\!\perp X$.

The above theorem might lead the reader to believe that SP and PP criteria come from different planets, bearing no relation to each other. After all, the theorem states that it is not possible for the predictor $\widehat{Y}$ to include *all variations* of X coming into Y and simultaneously include *no variations* of X coming into Y. This realization is the starting point of our discussion in the rest of the manuscript.

3. Causal Decompositions

In this section, we look at the causal decompositions of the measures associated with statistical parity and predictive parity. We start by analyzing the decompositions of the total variation (TV) measure, which is most commonly used to determine if statistical parity is satisfied. The causal decomposition of the TV measure requires the usage of causal measures known as counterfactual direct, indirect, and spurious effects, which are defined as:

Definition 3.1 (Counterfactual Causal Measures)**.** The counterfactual-{direct, indirect, spurious} effects of X on $\widehat{Y}$ are defined as follows

$$\text{Ctf-DE}_{x_0,x_1}(\widehat{y} \mid x) = P(\widehat{y}_{x_1,W_{x_0}} \mid x) - P(\widehat{y}_{x_0} \mid x) \quad (16)$$
$$\text{Ctf-IE}_{x_1,x_0}(\widehat{y} \mid x) = P(\widehat{y}_{x_1,W_{x_0}} \mid x) - P(\widehat{y}_{x_1} \mid x) \quad (17)$$
$$\text{Ctf-SE}_{x_0,x_1}(\widehat{y}) = P(\widehat{y}_{x_0} \mid x_1) - P(\widehat{y}_{x_0} \mid x_0). \quad (18)$$

The measures capture the variations from X to $\widehat{Y}$ going through: (i) the direct mechanism $X \to \widehat{Y}$; (ii) the indirect mechanism $X \to W \to \widehat{Y}$; (iii) the spurious mechanisms $X \dashleftarrow\dashrightarrow Z \to \widehat{Y}, X \dashleftarrow\dashrightarrow Z \to W \to \widehat{Y}$, respectively. Based on the defined measures, the $\text{TV}_{x_0,x_1}(\widehat{y})$ measure admits and additive decomposition, first obtained in (Zhang & Bareinboim, 2018b) and then extended in (Plečko & Bareinboim, 2022), given in the following theorem[2]:

Theorem 3.2 (Causal Decomposition of Statistical Parity (Zhang & Bareinboim, 2018b))**.** *The total variation measure admits a decomposition into its direct, indirect, and spurious variations:*

$$\begin{aligned} \textit{TV}_{x_0,x_1}(\widehat{y}) = \textit{Ctf-DE}_{x_0,x_1}(\widehat{y} \mid x_0) - & \quad (19)\\ \textit{Ctf-IE}_{x_1,x_0}(\widehat{y} \mid x_0) - & \quad (20)\\ \textit{Ctf-SE}_{x_1,x_0}(\widehat{y}). & \quad (21)\end{aligned}$$

The above theorem shows how we can disentangle direct, indirect, and spurious variations within the TV measure. We emphasize the importance of this result in the context of assessing the legal doctrines of discrimination. If a causal pathway (direct, indirect, or spurious) does not lie in the business necessity set, then the corresponding counterfactual measure (Ctf-DE, IE, or SE) needs to equal 0. To formalize this notion, we can now introduce the criterion of *causal statistical parity*:

[2]We note that the same decomposition can also be applied for the true outcome Y instead of the predictor $\widehat{Y}$, as will be relevant in later sections.

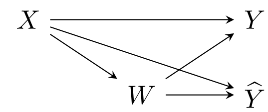

Figure 3. Standard Fairness Model with $Z = \emptyset$ from Thm. 3.4, extended with the predictor $\widehat{Y}$.

Definition 3.3 (Causal Statistical Parity)**.** We say that $\widehat{Y}$ satisfies causal statistical parity with respect to the protected attribute X if

$$\begin{aligned} \text{Ctf-DE}_{x_0,x_1}(\widehat{y} \mid x_0) &= \text{Ctf-IE}_{x_1,x_0}(\widehat{y} \mid x_0) \quad (22)\\ &= \text{Ctf-SE}_{x_1,x_0}(\widehat{y}) = 0. \quad (23)\end{aligned}$$

In practice, causal statistical parity can be a strong requirement, but we note that the notion can be easily relaxed to include only a subset of the Ctf-{DE, IE, or SE} measures, under business necessity requirements.

After decomposing the TV measure, we proceed along similar lines with the aim of obtaining a causal understanding of the predictive parity criterion, $Y \perp\!\!\!\perp X \mid \widehat{Y}$. The formal decomposition result of the PP measure is shown in the following theorem:

Theorem 3.4 (Causal Decomposition of Predictive Parity)**.** *Let $\mathcal{M}$ be an SCM compatible with the causal graph in Fig. 3 (i.e., SFM with $Z = \emptyset$). Then, it follows that the $PPM_{x_0,x_1}(y \mid \widehat{y}) = P(y \mid x_1, \widehat{y}) - P(y \mid x_0, \widehat{y})$ can be decomposed into its causal and spurious anti-causal variations as follows:*

$$\begin{aligned} PPM_{x_0,x_1}(y \mid \widehat{y}) =& P(y_{x_1} \mid x_1, \widehat{y}) - P(y_{x_0} \mid x_1, \widehat{y}) \quad (24)\\ &+ P(y_{x_0} \mid \widehat{y}_{x_1}) - P(y_{x_0} \mid \widehat{y}_{x_0}). \quad (25)\end{aligned}$$

Under the additional assumptions that (i) the SCM $\mathcal{M}$ is linear and Y is continuous; (ii) the learner $\widehat{Y}$ is efficient, we have that:

$$\mathbb{E}(y_{x_1} \mid x_1, \widehat{y}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}) = \alpha_{XW}\alpha_{WY} + \alpha_{XY} \quad (26)$$
$$\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0}) = -(\alpha_{XW}\alpha_{WY} + \alpha_{XY}), \quad (27)$$

where $\alpha_{V_iV_j}$ is the linear coefficient between variables V_i, V_j.

Proof. Note that $\mathbb{E}(y \mid x_1, \widehat{y}) - \mathbb{E}(y \mid x_0, \widehat{y})$ equals

$$\mathbb{E}(y_{x_1} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_0, \widehat{y}_{x_0}) \tag{28}$$
$$= \underbrace{\mathbb{E}(y_{x_1} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1})}_{\text{Term (I)}} \tag{29}$$
$$+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0})}_{\text{Term (II)}} \tag{30}$$
$$+ \underbrace{\mathbb{E}(y_{x_0} \mid x_1, \widehat{y}_{x_0}) - \mathbb{E}(y_{x_0} \mid x_0, \widehat{y}_{x_0})}_{\text{Term (III)}}. \tag{31}$$

Since there are no backdoor paths between X and $Y, \widehat{Y}$, Term (III) vanishes. By noting that $\mathbb{E}(y_x \mid x_1, \widehat{y}_{x_1}) = \mathbb{E}(y_x \mid x_1, \widehat{y}) \;\; \forall x$ by consistency axiom (Pearl, 2000, Ch.7), and applying the observation to Term (I) gives us the first part of the theorem. For the second part, we further assume that the SCM is linear, and that the predictor $\widehat{Y}$ is efficient, i.e., $\widehat{Y}(x, w) = \mathbb{E}[Y \mid x, w]$. In this case, the efficiency simply translates to the fact that

$$\alpha_{W\widehat{Y}} = \alpha_{WY}, \tag{32}$$
$$\alpha_{X\widehat{Y}} = \alpha_{XY}. \tag{33}$$

Due to linearity, for every unit u, we have that

$$y_{x_1}(u) - y_{x_0}(u) = \alpha_{XW}\alpha_{WY} + \alpha_{XY}, \tag{34}$$

and since Term (I) can be written as $\sum_u [y_{x_1}(u) - y_{x_0}(u)] P(u \mid x_1, \widehat{y})$ using the unit-level expansion of counterfactual distributions (Tian & Pearl, 2000; Bareinboim et al., 2022), Eq. 26 follows. Similarly, Term (II) can be expanded as

$$\sum_u \widehat{y}_{x_0}(u)[P(u \mid \widehat{y}_{x_1}) - P(u \mid \widehat{y}_{x_0})]. \tag{35}$$

We now look at units u which are compatible with $\widehat{Y}_{x_1}(u) = \widehat{y}$ and $\widehat{Y}_{x_0}(u) = \widehat{y}$. We can expand $\widehat{Y}_{x_1}(u)$ as

$$\widehat{Y}_{x_1}(u) = \alpha_{X\widehat{Y}} + \alpha_{XW}\alpha_{W\widehat{Y}} + \alpha_{W\widehat{Y}} u_W. \tag{36}$$

Thus, we have that

$$\widehat{Y}_{x_1}(u) = \widehat{y} \implies \alpha_{W\widehat{Y}} u_W = \widehat{y} - \alpha_{X\widehat{Y}} + \alpha_{XW}\alpha_{W\widehat{Y}}. \tag{37}$$

Similarly, we also obtain that

$$\widehat{Y}_{x_0}(u) = \widehat{y} \implies \alpha_{W\widehat{Y}} u_W = \widehat{y}. \tag{38}$$

Due to the efficiency of learning which implies that $\alpha_{W\widehat{Y}} = \alpha_{WY}$ and $\alpha_{X\widehat{Y}} = \alpha_{XY}$, Eq. 37 and 38 imply

$$y_{x_0}(u) = \widehat{y} - (\alpha_{XY} + \alpha_{XW}\alpha_{WY}) \;\; \forall u \text{ s.t. } \widehat{Y}_{x_1}(u) = \widehat{y}, \tag{39}$$

$$y_{x_0}(u) = \widehat{y} \;\; \forall u \text{ s.t. } \widehat{Y}_{x_0}(u) = \widehat{y}, \tag{40}$$

which in turn shows that

$$\mathbb{E}(y_{x_0} \mid \widehat{y}_{x_1}) - \mathbb{E}(y_{x_0} \mid \widehat{y}_{x_0}) = -\alpha_{XY} - \alpha_{XW}\alpha_{WY}. \tag{41}$$

□

We now carefully unpack the key insight from Theorem 3.4. In particular, we showed that in the case of an SFM with $Z = \emptyset$[3] the predictive parity measure can be written as:

$$\text{PPM} = \underbrace{P(y_{x_1} \mid x_1, \widehat{y}) - P(y_{x_0} \mid x_1, \widehat{y})}_{\text{causal}} + \tag{42}$$
$$\underbrace{P(y_{x_0} \mid \widehat{y}_{x_1}) - P(y_{x_0} \mid \widehat{y}_{x_0})}_{\text{reverse-causal spurious}}. \tag{43}$$

In words, the first term of the decomposition measures the causal variations induced from a transition $x_0 \to x_1$, for a fixed set of units. Interestingly, in the linear case, *this effect does not depend on the constructed predictor* $\widehat{Y}$, but only on the underlying system, i.e., it is not under the control of the predictor designer. To achieve the criterion PPM $= 0$, the second term needs to be exactly the reverse of the causal effect, captured by the spurious variations induced by changing $\widehat{y}_{x_0} \to \widehat{y}_{x_1}$ in the selection of units. The second term, which is in the control of the predictor $\widehat{Y}$ designer, needs to cancel out the causal effect measured by the first term. This key observation motivates a novel definition that we call *causal predictive parity*:

Definition 3.5 (Causal Predictive Parity)**.** Let $\widehat{Y}$ be a predictor of the outcome Y, and let X be the protected attribute. Then we say that $\widehat{Y}$ satisfies causal predictive parity (CPP) with respect to a counterfactual contrast (C_0, C_1, E, E) if

$$\mathbb{E}[y_{C_1} \mid E] - \mathbb{E}[y_{C_0} \mid E] = \mathbb{E}[\widehat{y}_{C_1} \mid E] - \mathbb{E}[\widehat{y}_{C_0} \mid E]. \tag{44}$$

Furthermore, we say that $\widehat{Y}$ satisfies CPP with respect to a factual contrast (C, C, E_0, E_1) if

$$\mathbb{E}[y_C \mid E_1] - \mathbb{E}[y_C \mid E_0] = \mathbb{E}[\widehat{y}_C \mid E_1] - \mathbb{E}[\widehat{y}_C \mid E_0]. \tag{45}$$

The intuition behind the notion of causal predictive parity captures the intuition behind predictive parity. If a contrast $\mathcal{C}$ describes some amount of variation in the outcome Y, then it should describe the same amount of variation in the predicted outcome $\widehat{Y}$. For any of the contrasts Ctf-DE, IE, SE corresponding to a causal pathway, causal predictive parity would require that

$$\mathcal{C}(X, \widehat{Y}) = \mathcal{C}(X, Y). \tag{46}$$

Algorithm 1 Business Necessity Cookbook

Input: data $\mathcal{D}$, BN-Set BN $\subseteq \{\text{DE, IE, SE}\}$
for CE $\in \{\text{DE, IE, SE}\}$ **do**
  if CE $\in$ BN **then**
    Compute the effects Ctf-CE(y), Ctf-CE$(\widehat{y})$
    Assert that Ctf-CE(y) $=$ Ctf-CE$(\widehat{y})$, otherwise FAIL
  else
    Compute the effect Ctf-CE$(\widehat{y})$
    Assert that Ctf-CE$(\widehat{y}) = 0$, otherwise FAIL
  end if
end for
if not FAIL **then**
  return SUCCESS
end if
Output: SUCCESS or FAIL of ensuring that disparate impact and treatment hold under business necessity.

3.1. Combining Statistical and Predictive Parity

We now tie the notions of statistical and predictive parity through the concept of *business necessity*. In particular, if a contrast $\mathcal{C}$ is associated with variations that are not in the business necessity set, then the value of this contrast should be $\mathcal{C}(X, \widehat{Y}) = 0$, following the intuition of causal statistical parity from Definition 3.3. However, if the variations associated with the contrast *are* in the business necessity set, then the value of that contrast should be equal for the predictor to the value for the true outcome

$$\mathcal{C}(X, \widehat{Y}) = \mathcal{C}(X, Y), \tag{47}$$

following the intuition of causal predictive parity. Combining these two notions through business necessity results in Algorithm 1. The algorithm requires the user to compute the measures

$$\text{Ctf-\{DE, IE, SE\}}(y), \text{Ctf-\{DE, IE, SE\}}(\widehat{y}). \tag{48}$$

Furthermore, for each of the DE, IE, and SE effects, the user needs to determine whether the causal effect (CE) in question falls into the business necessity set. If yes, then the algorithm asserts that

$$\text{Ctf-CE}(y) = \text{Ctf-CE}(\widehat{y}). \tag{49}$$

In the other case, when the causal effect is not in the business necessity set, the algorithm asserts that

$$\text{Ctf-CE}(\widehat{y}) = 0. \tag{50}$$

We remark that Algorithm 1 is written in its population level version, in which the causal effects are estimated perfectly with no uncertainty. In the finite sample case, one needs to compute the confidence intervals of different measures, and check whether they overlap. If one is also interested in constructing a new fair predictor before using Algorithm 1 (instead of testing an existing one), one may use tools for causally removing discrimination, for example the `fairadapt` R-package (Plečko & Meinshausen, 2020; Plečko et al., 2021) .

[3]We remark that the essence of the argument is unchanged in the case with $Z \neq \emptyset$, but handling this case limits the clarity of presentation.

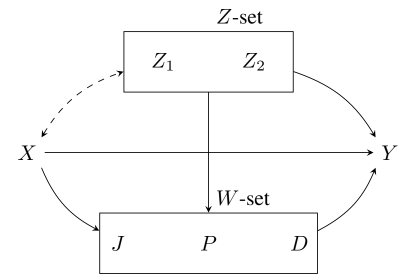

Figure 4. SFM for the COMPAS dataset.

4. Experiment

We now apply Algorithm 1 to the COMPAS dataset (Angwin et al., 2016), as described in the following example.

Courts in Broward County, Florida use machine learning algorithms, developed by Northpointe, to predict whether individuals released on parole are at high risk of re-offending within 2 years (Y). The algorithm is based on the demographic information Z (Z_1 for gender, Z_2 for age), race X (x_0 denoting White, x_1 Non-White), juvenile offense counts J, prior offense count P, and degree of charge D.

We construct the standard fairness model (SFM) for this example, which is shown in Figure 4. The bidirected arrow between X and $\{Z_1, Z_2\}$ indicates possible co-variations of race with age and sex, which may not be causal in nature[4]. Furthermore, $\{Z_1, Z_2\}$ are the confounders, not causally affected by race X. The set of mediators $\{J, P, D\}$, however, may be affected by race, due to an existing societal bias in policing and criminal justice. Finally, all of the above mentioned variables may influence the outcome Y.

Having access to data from Broward County, and equipped with Algorithm 1, we want to prove that the recidivism predictions produced by Northpointe (labeled $\widehat{Y}^{NP}$) violate legal doctrines of anti-discrimination. Suppose that in an initial hearing, the Broward County district court determines that the direct and indirect effects are not in the business necessity set, while the spurious effect is. In words, gender (Z_1) and age (Z_2) are allowed to be used to distinguish between the minority and majority groups when predicting recidivism, while other variables are not. If Northpointe's

[4]The causal model is non-committal regarding the complex historical/social processes that lead to such co-variations.

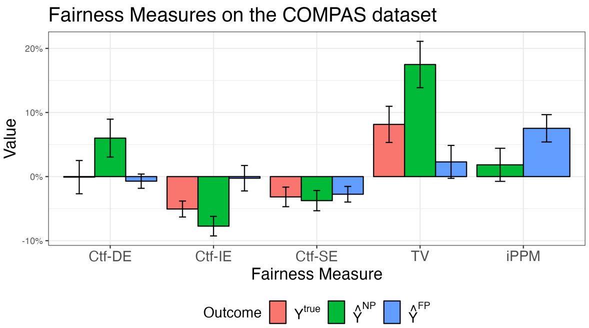

Figure 5. Fairness measures for the COMPAS example.

predictions are found to be discriminatory, we are required by the court to produce better, non-discriminatory predictions.

In light of this information, we proceed as follows (see source code). We first obtain a causally fair predictor $\widehat{Y}^{FP}$ using the `fairadapt` package. Then, we compute the counterfactual causal measures of fairness for the true outcome Y, Northpointe's predictions $\widehat{Y}^{NP}$, and the fair predictions $\widehat{Y}^{FP}$ (see Figure 5). For the direct effect, we have:

$$\text{Ctf-DE}_{x_0,x_1}(y \mid x_0) = -0.08\% \pm 2.59\%, \tag{51}$$
$$\text{Ctf-DE}_{x_0,x_1}(\widehat{y}^{NP} \mid x_0) = 6\% \pm 2.96\%, \tag{52}$$
$$\text{Ctf-DE}_{x_0,x_1}(\widehat{y}^{FP} \mid x_0) = -0.72\% \pm 1.11\%. \tag{53}$$

Since the direct effect is not in the business necessity set, Northpointe's predictions clearly violate the disparate treatment doctrine (green bar for the Ctf-DE measure in Figure 5). Our predictions, however, do not exhibit a statistically significant direct effect of race on the outcome, so they do not violate the criterion (blue bar). Next, for the indirect effects, we obtain:

$$\text{Ctf-IE}_{x_1,x_0}(y \mid x_0) = -5.06\% \pm 1.24\%, \tag{54}$$
$$\text{Ctf-IE}_{x_1,x_0}(\widehat{y}^{NP} \mid x_0) = -7.73\% \pm 1.53\%, \tag{55}$$
$$\text{Ctf-IE}_{x_1,x_0}(\widehat{y}^{FP} \mid x_0) = -0.25\% \pm 1.98\%. \tag{56}$$

Once again, the indirect effect, which is in the business necessity set, is different from 0 for the Northpointe's predictions (violating disparate impact, see green bar for Ctf-IE in Figure 5), but not statistically different from 0 for our predictions (blue bar). Interestingly, the indirect effect is different from 0 for the true outcome (red bar), indicating a bias in the current real world. Finally, for the spurious effects, we obtain

$$\text{Ctf-SE}_{x_1,x_0}(y) = -3.17\% \pm 1.53\%, \tag{57}$$
$$\text{Ctf-SE}_{x_1,x_0}(\widehat{y}^{NP}) = -3.75\% \pm 1.58\%, \tag{58}$$
$$\text{Ctf-SE}_{x_1,x_0}(\widehat{y}^{FP}) = -2.75\% \pm 1.22\%. \tag{59}$$

Since the spurious effect is in the business necessity set and the confidence intervals of all three measures overlap, no violations with respect to spurious effects are found. The estimated effects for the outcome $Y, \widehat{Y}^{NP}$, and $\widehat{Y}^{FP}$ are also shown graphically in Figure 5. We conclude that Northpointe's predictions $\widehat{Y}^{NP}$ violate the legal doctrines of fairness, while our predictions $\widehat{Y}^{FP}$ do not. Importantly, tying back to the original discussion that motivated our approach, Northpointe's predictions $\widehat{Y}^{NP}$ are further away from statistical parity than our predictions $\widehat{Y}^{FP}$ according to the TV measure (see TV column in Fig. 5), while at the same time better calibrated according to the integrated PP measure (iPPM) that averages the PP measures from Eq. 9 across different values of $\widehat{y}$ (see iPPM column in the figure). This observation demonstrates the trade-off between statistical and predictive parity through business necessity.

5. Conclusions

The literature in fair machine learning is abundant with fairness measures (Corbett-Davies & Goel, 2018), many of which are mutually incompatible. Nonetheless, it is doubtable that each of these measures corresponds to a fundamentally different ethical conception of fairness. The multitude of possible approaches to quantifying discrimination makes the consensus on an appropriate notion of fairness unattainable. Furthermore, the impossibility results between different measures may be discouraging to data scientists who wish to quantify and remove discrimination, but are almost immediately faced with a choice of which measure they wish to subscribe to.

In this work, we attempt to remedy a part of this issue by focusing on the impossibility of simultaneously achieving statistical and predictive parity. As our discussion shows, the guiding idea behind statistical parity is that variations transmitted along causal pathways from the protected attribute to the predictor should equal 0, i.e., the decision should not depend on the protected attribute through the causal pathway in question (Definition 3.3). Complementary to this notion, the guiding principle behind predictive parity is that the variations transmitted along a causal pathway should be equal for the predictor as they are for the outcome *in the real world* (Definition 3.5). Statistical parity will therefore be satisfied when the BN set includes all variations coming from X to Y, while predictive parity will be satisfied when the BN set is empty. The choice of the BN set, therefore, interpolates between statistical and predictive parity, in a way that can be formally assessed based on Algorithm 1.

Therefore, instead of being viewed as mutually exclusive notions, tied through the impossibility result from Theorem 2.9, statistical and predictive parity can be viewed as the extremes of a spectrum which is spanned by the different choices of the BN set, as visualized in Figure 1.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 5 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, NY, USA, 2018.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Technical Report arXiv:1703.00056, arXiv.org, 2017.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Darlington, R. B. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971.

Detrixhe, J. and Merrill, J. B. The fight against financial advertisers using facebook for digital redlining, 11 2019.

Grimmelmann, J. and Westreich, D. Incomprehensible discrimination. *Calif. L. Rev. Circuit*, 7:164, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Harwell, D. Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/, 12 2019.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Plečko, D. and Bareinboim, E. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022. *(To appear in Foundations and Trends in Machine Learning)*.

Plečko, D. and Meinshausen, N. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21:242, 2020.

Plečko, D., Bennett, N., and Meinshausen, N. fairadapt: Causal reasoning for fair data pre-processing. *arXiv preprint arXiv:2110.10200*, 2021.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169 (12):866–872, 2018.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Zhang, J. and Bareinboim, E. Equality of opportunity in classification: A causal approach. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3671–3681, Montreal, Canada, 2018a. Curran Associates, Inc.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.