Mind the Gap: A Causal Perspective on Bias Amplification in Prediction & Decision-Making

Drago Plecko and **Elias Bareinboim**
Department of Computer Science
Columbia University
dp3144@columbia.edu, eb@cs.columbia.edu

Abstract

As society increasingly relies on AI-based tools for decision-making in socially sensitive domains, investigating fairness and equity of such automated systems has become a critical field of inquiry. Most of the literature in fair machine learning focuses on defining and achieving fairness criteria in the context of prediction, while not explicitly focusing on how these predictions may be used later on in the pipeline. For instance, if commonly used criteria, such as independence or sufficiency, are satisfied for a prediction score S used for binary classification, they need not be satisfied after an application of a simple thresholding operation on S commonly used in practice. The same problem persists for numerous other statistical and causal notions of fairness. In this paper, we take an important step to address this issue. We introduce the notion of a margin complement, which measures how much a prediction score S changes due to a thresholding operation. We then demonstrate that the marginal difference in the optimal 0/1 predictor $\widehat{Y}$ between groups, written $P(\hat{y} \mid x_1) - P(\hat{y} \mid x_0)$, can be causally decomposed into the influences of X on the L_2-optimal prediction score S and the influences of X on the margin complement M, along different causal pathways (direct, indirect, spurious). When then show that under suitable causal assumptions, the influences of X on the prediction score S are equal to the influences of X on the true outcome Y. This yields a new decomposition of the disparity in the predictor $\widehat{Y}$ that allows us to disentangle causal differences inherited from the true outcome Y that exists in the real world vs. those coming from the optimization procedure itself. This observation highlights the need for more regulatory oversight due to the potential for bias amplification, and to address this issue we introduce new notions of *weak* and *strong* business necessity, together with an algorithm for assessing whether these notions are satisfied. We apply our method to three real-world datasets and derive new insights on bias amplification in prediction and decision-making.

1 Introduction

Automated systems based on machine learning and artificial intelligence are increasingly used for decision-making in a variety of real-world settings. These applications include hiring decisions, university admissions, law enforcement, credit lending and loan approvals, health care interventions, and many other high-stakes scenarios in which the automated system may significantly affect the well-being of individuals [13, 18, 5]. In this context, a pervasive question has become the following: what are the implications and consequences of using an automated system, compared to the currently implemented decision process? Various prior works highlight the potential of automated systems to perpetuate or even amplify inequities between demographic groups, with a range of examples from decision support systems for (among others) sentencing [2], face-detection [6], online advertising [29, 10], and authentication [28]. Notably, issues of unfairness and discrimination are also pervasive in settings in which decisions are made by humans. Some well-studied examples include the gender pay gap, supported by a decades-long literature [3, 4], or the racial bias in criminal sentencing [30, 21]. Therefore, AI systems designed to make decisions may often be trained with data that contains various historical biases and past discriminatory decisions against certain protected groups, constituting a large part of the underlying problem. In this work, we specifically focus on investigating when automated systems may potentially lead to an even more discriminatory process, possibly amplifying already existing differences between groups.

Within this context, it is quite useful to distinguish between different tasks appearing in the growing literature on fair machine learning. One can distinguish three specific and different tasks, namely (1) bias detection and quantification for exisiting outcomes or decision policies; (2) construction of fair predictions of an outcome; (3) construction of fair decision-making policies that are intended to be implemented in the real-world. Interestingly, a large portion of the literature in fair ML focuses on the second task of fair prediction, and what is often left unaddressed is how these predictions may be used later on in the pipeline, and what kind of consequences they may have. For instance, consider a prediction score S for a binary outcome Y that satisfies well-known fairness criteria, such as independence (demographic parity [9]) or sufficiency (calibration [8]). After a simple thresholding operation, commonly applied in settings with a binary outcome, the resulting predictor is no longer guaranteed to satisfy independence or sufficiency, and the previously provided fairness guarantees may be entirely lost. The same behavior can be observed for numerous other measures.

These difficulties do not apply only to statistical measures of fairness. Recently, a growing literature has explored causal approaches to fair machine learning [16, 14, 19, 33, 32, 31, 7, 27, 24], which have two major benefits. First, they allow for human-understandable and interpretable definitions and metrics of fairness, which are tied to the causal mechanisms transmitting the change between groups. Secondly, they offer a language that is aligned with the legal notions of discrimination, such as the disparate impact doctrine. In particular, causal approaches allow for considerations of business necessity – which aim to ellucidate which covariates may be justifiably used by decision-makers even if their usage implies a disparity between groups. However, causal approaches to fairness also suffer from the above-discussed issues – namely, a guarantee of absence of a causal influence from the protected attribute X onto a predictor S need not hold true after the predictor is thresholded [24]. Therefore, within the causal approach there is also a major need for a better understanding of how probabilistic predictions are translated into binary predictions or decisions.

Against this background, in this work we take an important step in the direction of addressing this issue. We work in a setting with a binary label Y, and the goal is to provide a binary prediction $\widehat{Y}$ or a binary decision D. Our approach is particularly suitable for settings in which the utility of the decision is monotonic with respect to the conditional probability of Y being positive, written $P(Y \mid \text{covariates})$, but the developed tools also have ramifications more broadly. Examples that fall under our scope are numerous: for instance, the utility of admitting a student to the university (D) is often monotonic in the probability that the student successfully graduates (Y). In the context of criminal justice, decisions of detention (D) are used to prevent recidivism, and the utility of the decision is monotonic in the probability that the individual recidivates (Y). Finally, various preventive measures in healthcare (D, such as vaccination, screening tests) are considered to be best applied to individuals with the highest risk of developing a target disease or suffering a negative outcome (Y). We now provide an example that sheds light on one of the key insights of our paper:

Example 1 ($\widehat{Y}$ and S Disparities in Hiring)**.** *Consider a company deciding to hire employees using an automated system for the first time. From a previous hiring cycle, the company has access to data on gender X (x_0 for female, x_1 for male) and the hiring outcome Y (y_1 for being hired, y_0 otherwise). The true underlying equations of the system are given by:*

$$X \leftarrow \textit{Bernoulli}(0.5) \tag{1}$$

$$Y \leftarrow \begin{cases} \textit{Bernoulli}(p_0) \textit{ if } X = x_0 \\ \textit{Bernoulli}(p_1) \textit{ if } X = x_1. \end{cases} \tag{2}$$

The company finds the optimal prediction score S to be $S(x) = p_x$. The optimal 0/1 predictor $\widehat{Y}$, which will also be the company's decision, is given by $\widehat{Y}(x) = \mathbb{1}(S(x) \geq \frac{1}{2})$. Note that $P(s \mid x_1) - P(s \mid x_0) = p_1 - p_0$ and $P(\widehat{y} \mid x_1) - P(\widehat{y} \mid x_0) = \mathbb{1}(p_1 \geq \frac{1}{2}) - \mathbb{1}(p_0 \geq \frac{1}{2})$. The *marginal difference between the groups can be decomposed into two parts:*

$$P(\widehat{y} \mid x_1) - P(\widehat{y} \mid x_0) = \underbrace{p_1 - p_0}_{\textit{Term I}} + \underbrace{(\mathbb{1}(p_1 \geq \frac{1}{2}) - p1) - (\mathbb{1}(p_0 \geq \frac{1}{2}) - p_0)}_{\textit{Term II}}. \tag{3}$$

For $p_0 = 0.49, p_1 = 0.51$, the disparity in S (Term I) equals 2%, and there is an additional 98% difference from the rounding of $\widehat{Y}$ (Term II). For $p_0 = 0.51, p_1 = 1$, the disparity in S equals 49% (Term I) whereas the rounding of $\widehat{Y}$ contributes -49% (Term II). □

The above example illustrates a canonical point in a very simple setting. The disparity in the optimal 0/1 predictor $\widehat{Y}$ has two constitutive elements: (1) the contribution coming from the disparity in the true outcome Y, reflected in the prediction score S, and (2) the contribution coming from thresholding the prediction score to obtain an optimal 0/1 prediction. As shown in the example, depending on the situation, a very small disparity in the outcome Y, and consequently the prediction score S, may result in a very large disparity in the optimal 0/1 predictor $\widehat{Y}$ (case of bias amplification). Complementary to this, a large disparity in S may in fact result in a very small disparity in $\widehat{Y}$ (case of bias amelioration).

In the remainder of the manuscript, our goal is to provide a decomposition of the disparity in a thresholded predictor $\widehat{Y}$ into the disparity in true outcome Y and the disparity originating from optimization procedure, but *along each causal pathway* between the protected attribute X and the predictor $\widehat{Y}$. In particular, our contributions are the following:

(1) We introduce the notion of margin complement (Def. 2), and provide a path-specific decomposition of the disparity in the 0/1 predictor $\widehat{Y}$ into its contributions from the optimal score predictor S and the margin complement M (Thm. 1),

(2) We prove that under suitable causal assumptions, the causal decomposition of the optimal prediction score S is equivalent with the causal decomposition of the true outcome Y (Thm. 2). This allows us to obtain a new decomposition of the disparity in $\widehat{Y}$ into contributions from Y and the margin complement M (Cor. 3),

(3) Motivated by the above decompositions, we introduce a new concept of weak and strong business necessity (Def. 4), highlighting a new need for regulatory instructions in the context of automated systems. We provide an algorithm for assessing fairness under considerations of weak and strong business necessity (Alg. 1),

(4) We provide identification, estimation, and sample influence results for all of the quantities relevant to the above framework (Props. 4-6). We evaluate our approach on three real-world examples (Ex. 2-4) and provide new empirical insights into the question of bias amplification.

Our work is related to the previous literature on causal fairness and the causal decompositions appearing in this literature [33, 24]. However, our approach offers an entirely new causal decomposition into contributions from the true outcome Y and the margin complement M. Our work is also related to recent results showing that focusing purely on prediction, and ignoring decision-making aspects, may lead to inequitable outcomes and cause harm to marginalized groups [24, 20, 25], highlighting a need to expand focus from narrow statistical definitions of fair predictions to a more comprehensive understanding of equity in algorithmic decisions. Finally, our work is also related to the literature audit and assess the fairness of decisions made by humans [23, 15], and understading how AI systems may help humans overcome their biases [11].

1.1 Preliminaries

We use the language of structural causal models (SCMs) as our basic semantical framework [22]. A structural causal model (SCM) is a tuple $\mathcal{M} := \langle V, U, \mathcal{F}, P(u) \rangle$, where V, U are sets of endogenous (observables) and exogenous (latent) variables, respectively, $\mathcal{F}$ is a set of functions f_{V_i}, one for each $V_i \in V$, where $V_i \leftarrow f_{V_i}(\mathrm{pa}(V_i), U_{V_i})$ for some $\mathrm{pa}(V_i) \subseteq V$ and $U_{V_i} \subseteq U$. $P(u)$ is a strictly positive probability measure over U. Each SCM $\mathcal{M}$ is associated to a causal diagram $\mathcal{G}$ [22] over the node set V where $V_i \rightarrow V_j$ if V_i is an argument of f_{V_j}, and $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding U_{V_i}, U_{V_j} are not independent. An instantiation of the exogenous variables $U = u$ is called a *unit*. By $Y_x(u)$ we

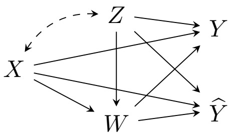

Figure 1: Standard Fairness Model.

denote the potential response of Y when setting $X = x$ for the unit u, which is the solution for $Y(u)$ to the set of equations obtained by evaluating the unit u in the submodel $\mathcal{M}_x$, in which all equations in $\mathcal{F}$ associated with X are replaced by $X = x$. Building on the notion of a potential response, one can further define the notions of counterfactual and factual contrasts, given by:

Definition 1 (Contrasts [24])**.** *Given an SCM $\mathcal{M}$, a contrast $\mathcal{C}$ is any quantity of the form*

$$\mathcal{C}(C_0, C_1, E_0, E_1) = \mathbb{E}[y_{C_1} \mid E_1] - \mathbb{E}[y_{C_0} \mid E_0], \tag{4}$$

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

(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 later on.

Throughout this manuscript, we assume a specific cluster causal diagram $\mathcal{G}_{\text{SFM}}$ known as the standard fairness model (SFM) [24] over endogenous variables $\{X, Z, W, Y, \widehat{Y}\}$ shown in Fig. 1. The SFM consists of the following: *protected attribute*, labeled X (e.g., gender, race, religion), 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). The SFM also encodes the assumptions typically used in the causal inference literature about the lack of hidden confounding[1].

2 Margin Complements

We begin by introducing a quantity that plays a key role in the results of this paper.

Definition 2 (Margin Complement)**.** *Let $U = u$ be a unit, and let S denote a prediction score for a binary outcome Y. Let the subscript C denote a counterfactual clause, so that Z_C denotes a potential response. The margin complement M of the score S for the unit $U = u$ and threshold t is defined as:*

$$M(u) = \mathbb{1}(S(u) \geq t) - S(u). \tag{5}$$

Following this, a potential response of a margin complement M, labeled M_C, is given by $M_C(u) = \mathbb{1}(S_C(u) \geq t) - S_C(u)$.

The definition of a margin complement has a straightforward intuition: given a prediction score S and a threshold t, the margin complement M tells us in which direction the thresholded version $\mathbb{1}(S(u) \geq t)$ moves compared to the score $S(u)$. A positive margin complement indicates that a thresholded predictor is larger than the probability prediction, and a negative margin complement the opposite. A similar reasoning holds for the potential responses of the margin complement M_C: we are interested in what the margin complement *would have been* for an individual $U = u$ under possibly different, counterfactual conditions described by C. As we demonstrate shortly, margin complements (and their potential responses) play a major role in explaining how inequities are generated between

[1] Partial identification techniques for bounding effects can be used for relaxing these assumptions [34].

groups at the time of decision making. In this section, our key aim is to analyze the optimal 0/1 predictor $\widehat{Y}$ and provide a decomposition of its total variation measure (TV, for short), defined as

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

In particular, when working with the causal diagram in Fig. 1, we can notice that TV measure is comprised of three types of variations coming from X: the direct effect $X \to \widehat{Y}$, the mediated effect $X \to W \to \widehat{Y}$, and the confounded effect $X \dashleftarrow\dashrightarrow Z \to \widehat{Y}$. Our goal is to construct a decomposition of the TV measure that allow us to distinguish how much of each of the causal effects is due to a difference in the prediction score S, and how much due to margin complements M.

To investigate this, we first introduce the known definitions of direct, indirect, and spurious effects from the causal fairness literature:

Definition 3 (x-specific Causal Measures [33, 24])**.** *The x-specific {direct, indirect, spurious} effects of X on a random variable Y are defined as follows:*

$$\begin{aligned} x\text{-}DE_{x_0,x_1}(y \mid x) &= P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x) & (7)\\ x\text{-}IE_{x_1,x_0}(y \mid x) &= P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_1} \mid x) & (8)\\ x\text{-}SE_{x_0,x_1}(y) &= P(y_{x_0} \mid x_1) - P(y_{x_0} \mid x_0). & (9)\end{aligned}$$

Armed with these definitions, we can prove the following theorem:

Theorem 1 (Causal Decomposition of Optimal 0/1 Predictor)**.** *Let $\widehat{Y}$ be the optimal predictor with respect to the 0/1-loss based on covariates X, Z, W. Let S denote the optimal predictor with respect to the L_2 loss. The total variation (TV, for short) measure of the predictor $\widehat{Y}$, written as $P(\hat{y} \mid x_1) - P(\hat{y} \mid x_0)$, can be decomposed into direct, indirect, and spurious effects of X on the score S and the margin complement M as follows:*

$$\begin{aligned} TV_{x_0,x_1}(\widehat{y}) &= x\text{-}DE_{x_0,x_1}(s \mid x_0) + x\text{-}DE_{x_0,x_1}(m \mid x_0) & (10)\\ &\quad - \big(x\text{-}IE_{x_1,x_0}(s \mid x_0) + x\text{-}IE_{x_1,x_0}(m \mid x_0)\big) & (11)\\ &\quad - \big(x\text{-}SE_{x_1,x_0}(s) + x\text{-}SE_{x_1,x_0}(m)\big). & (12)\end{aligned}$$

The above theorem is the first key result of this paper. The disparity between groups with respect to the optimal 0/1-loss predictor, measured by $\text{TV}_{x_0,x_1}(\hat{y})$ can be decomposed into direct, indirect, and spurious contributions coming from (i) the optimal L_2-loss predictor S, and (ii) the margin complement M. This provides us with a unique capability: for each causal pathway (direct, indirect, spurious), we can disentangle the contribution coming from the probability prediction S vs. the contribution coming from the optimization procedure itself (i.e., the rounding of the predictor). The former, as we will see shortly, is simply a representation of the bias already existing in the true outcome Y, whereas the latter represents a newly introduced difference that is the result of using an automated system. The contribution of the margin complement may act to both ameliorate or amplify an existing disparity, a point we investigate across a range of real-world datasets in Sec. 5.

We next move onto a crucial point mentioned above: the disparity observed in the optimal L_2 score S is simply a representation of the disparity existing in the real-world, under suitable causal assumptions. This is demonstrated in the following theorem:

Theorem 2 (Symmetry of L_2-score S and Outcome Y Decompositions)**.** *Let $\mathcal{M}$ be an SCM compatible with the Standard Fairness Model, and let S be the optimal L_2 prediction score. Then, the causal decompositions of the score S and the true outcome Y are symmetric, meaning that*

$$\begin{aligned} x\text{-}DE_{x_0,x_1}(s \mid x_0) &= x\text{-}DE_{x_0,x_1}(y \mid x_0) & (13)\\ x\text{-}IE_{x_1,x_0}(s \mid x_0) &= x\text{-}IE_{x_1,x_0}(y \mid x_0) & (14)\\ x\text{-}SE_{x_1,x_0}(s) &= x\text{-}SE_{x_1,x_0}(y). & (15)\end{aligned}$$

Based on this result, we can in fact see that the causal influences from X on the true outcome Y are equivalent to the causal influences from X on the optimal prediction score S. Combining this insight with Thm. 1 gives us the following corollary:

Corollary 3 (Causal Decomposition of Optimal 0/1 Predictor)**.** *Under the Standard Fairness Model, the TV measure of the optimal 0/1-loss predictor $\widehat{Y}$ can be decomposed as:*

$$\begin{aligned} \textit{TV}_{x_0,x_1}(\widehat{y}) = x\textit{-DE}_{x_0,x_1}(y \mid x_0) + x\textit{-DE}_{x_0,x_1}(m \mid x_0) \quad (16) \\ - \big(x\textit{-IE}_{x_1,x_0}(y \mid x_0) + x\textit{-IE}_{x_1,x_0}(m \mid x_0)\big) \quad (17) \\ - \big(x\textit{-SE}_{x_1,x_0}(y) + x\textit{-SE}_{x_1,x_0}(m)\big). \quad (18) \end{aligned}$$

This corollary provides a very important step compared to Thm. 1: we can now decompose the disparity in the optimal predictor $\widehat{Y}$ into the contribution inherited from the true outcome Y and the contribution that arises from the optimization procedure (rounding of the predictor). Multiple important questions arise from this fundamental observation – and we begin by discussing its relation with the concept of business necessity.

3 Weak and Strong Business Necessity

From a legal standpoint, questions of fairness and discrimination can be interpreted based on the disparate treatment and disparate impact doctrines of Title VII of the Civil Rights Act of 1964 [1]. The disparate treatment doctrine disallows the effect of the protected attribute *ceteris paribus* – meaning that similarly situated individuals who differ with respect to the protected characteristic should not have disparate outcomes. When interpreted causally, the disparate treatment doctrine can be seen as disallowing a direct type of effect from X onto the outcome. The disparate impact doctrine, however, is more broad and may prohibit any from of discrimination (be it direct, indirect, or spurious) that results in a large disparity between groups. The core of this doctrine is the notion of *business necessity* (BN). Considerations of business necessity may allow variables correlated with the protected attribute to act as a proxy, and the law does not necessarily prohibit their usage due to their relevance to the business itself (or more broadly the utility of the decision-maker). Based on the decomposition from Cor. 3, new considerations in the context of business necessity emerge:

Definition 4 (Weak and Strong Business Necessity)**.** *Let $\mathcal{M}$ be an SCM compatible with the Standard Fairness Model. Let CE denote a causal pathway (DE, IE, or SE). If a causal pathway does not fall under business necessity, then we require:*

$$x\textit{-CE}_{x,x'}(s \mid x'') = x\textit{-CE}_{x,x'}(m \mid x'') = 0. \quad (19)$$

We say that a pathway satisfies weak business necessity if:

$$x\textit{-CE}_{x,x'}(s \mid x'') = x\textit{-CE}_{x,x'}(y \mid x''),\ x\textit{-CE}_{x,x'}(m \mid x'') = 0. \quad (20)$$

We say that a pathway satisfies strong business necessity if:

$$x\textit{-CE}_{x,x'}(s \mid x'') = x\textit{-CE}_{x,x'}(y \mid x''),\ x\textit{-CE}_{x,x'}(m \mid x'') \textit{ unconstrained}. \quad (21)$$

The above definition distinguishes between three very important cases, and sheds light on a new aspect of the concept of business necessity. According to the definition, there are three versions of business necessity considerations:

(1) A causal pathway is not in the business necessity set, and is considered discriminatory. In this case, both the contribution of the prediction score S and the margin complement M need to be equal to 0 (i.e., no discrimination is allowed along the pathway),

(2) A causal pathway satisfies weak business necessity, and is not considered discriminatory. In this case, the effect of X on the prediction score S needs to equal the effect of X onto the true outcome Y along the same pathway [26]. However, the contribution of the margin complement M along the pathway needs to equal 0.

(3) A causal pathway satisfies strong business necessity, and is not considered discriminatory. Similarly as for weak necessity, the effect of X on S needs to equal the effect of X on Y, but in this case, the contribution of the margin complement M is unconstrained.

The distinction between cases (2) and (3) opens the door for new regulatory requirements and specifications. In particular, whenever a causal effect is considered non-discriminatory, the attribute X needs to affect S to the extent to which it does in the real world. However, the system designer

```
Algorithm 1: Auditing Weak & Strong Business Necessity
Input: data $\mathcal{D}$, BN-Set BN $\subseteq \{\text{DE}, \text{IE}, \text{SE}\}$, BN-Strength, predictor $\widehat{Y}$, prediction score $S$
Output: SUCCESS or FAIL of ensuring that disparate impact and treatment hold under BN
foreach CE $\in \{DE, IE, SE\}$ do
    Compute the effects $x$-CE($y$), $x$-CE($m$), $x$-CE($s$), $x$-CE($\widehat{y}$)
    if CE $\in$ Strong-BN then
        Assert that $x$-CE($s$) $=$ $x$-CE($\widehat{y}$), otherwise FAIL
    else if CE $\in$ Weak-BN then
        Assert that $x$-CE($s$) $=$ $x$-CE($\widehat{y}$) $\wedge$ $x$-CE($m$) $= 0$, otherwise FAIL
    else
        Assert that $x$-CE($\widehat{s}$) $=$ $x$-CE($\widehat{m}$) $= 0$, otherwise FAIL
if not FAIL then
    return SUCCESS
```

also needs to decide whether a difference existing in the predicted probabilities S is allowed to be amplified (or ameliorated) by means of rounding. The latter point distinguishes between weak and strong business necessity, and should be a consideration of any system designer issuing binary decisions. In Alg. 1, we propose a formal approach for evaluating considerations of weak and strong business necessity for any input of a predictor $\widehat{Y}$ and a prediction score S. We next turn to the practical aspects of the procedure described in Alg. 1.

4 Identification, Estimation, Sample Influence

In Thm. 1 and Cor. 3 we decomposed the observed disparity in the TV measure of the optimal 0/1 predictor into its various components. The quantities appearing in the decomposition are counterfactuals, and therefore we need to address the question of *identification* of these quantities. In other words, we need to understand whether these quantities of interest can be uniquely computed based on the available data and the causal assumptions. We can provide a positive answer to this question:

Proposition 4 (Identification of Causal Measures)**.** *Let $\mathcal{M}$ be an SCM compatible with the Standard Fairness Model, and let $P(V)$ be the associated observational distribution. The x-specific causal effects of X on the outcome Y, predictor $\widehat{Y}$, prediction score S, and the margin complement M are identifiable (uniquely computable) from $P(V)$ and the SFM.*

The proof of the proposition, together with the identification expressions for the different quantities can be found in Appendix B. Importantly, we next provide estimation expressions that allow us to compute the desired quantities from the observational data:

Proposition 5 (Estimation Expressions)**.** *Let $f(x, z, w)$ denote the estimator of the expected value $\mathbb{E}[T \mid x, z, w]$. Let $\hat{P}(x \mid v')$ denote an estimator for the probability $P(x \mid v')$ for different choices of v'. Under the Standard Fairness Model, the x-specific direct, indirect, and spurious effects of X on a variable T can be estimated as:*

$$x\text{-}DE^{\text{est}}_{x_0,x_1}(t \mid x_0) = \frac{1}{n}\sum_{i=1}^{n}[f(x_1, w_i, z_i) - f(x_0, w_i, z_i)]\frac{\hat{P}(x_0 \mid w_i, z_i)}{\hat{P}(x_0)} \tag{22}$$

$$x\text{-}IE^{\text{est}}_{x_1,x_0}(t \mid x_0) = \frac{1}{n}\sum_{i=1}^{n} f(x_1, w_i)\Big[\frac{\hat{P}(x_0 \mid w_i, z_i)}{\hat{P}(x_0)} - \frac{\hat{P}(x_1 \mid w_i, z_i)}{\hat{P}(x_1 \mid z_i)}\frac{\hat{P}(x_0 \mid z_i)}{\hat{P}(x_0)}\Big] \tag{23}$$

$$x\text{-}SE^{\text{est}}_{x_1,x_0}(t) = \frac{1}{n}\sum_{i=1}^{n} f(x_1, w_i)\Big[\frac{\hat{P}(x_1 \mid w_i, z_i)}{\hat{P}(x_1 \mid z_i)}\frac{\hat{P}(x_0 \mid z_i)}{\hat{P}(x_0)} - \frac{\hat{P}(x_1 \mid w_i, z_i)}{\hat{P}(x_1)}\Big]. \tag{24}$$

The proof of the proposition can be found in Appendix B. We next define the sample influences for the different estimators:

Definition 5 (Sample Influence)**.** *The sample influence of the i-th sample on the estimator x-CE$^{\text{est}}$ of the causal effect CE is given by corresponding term in the summations in Eqs. 22-24. For instance,*

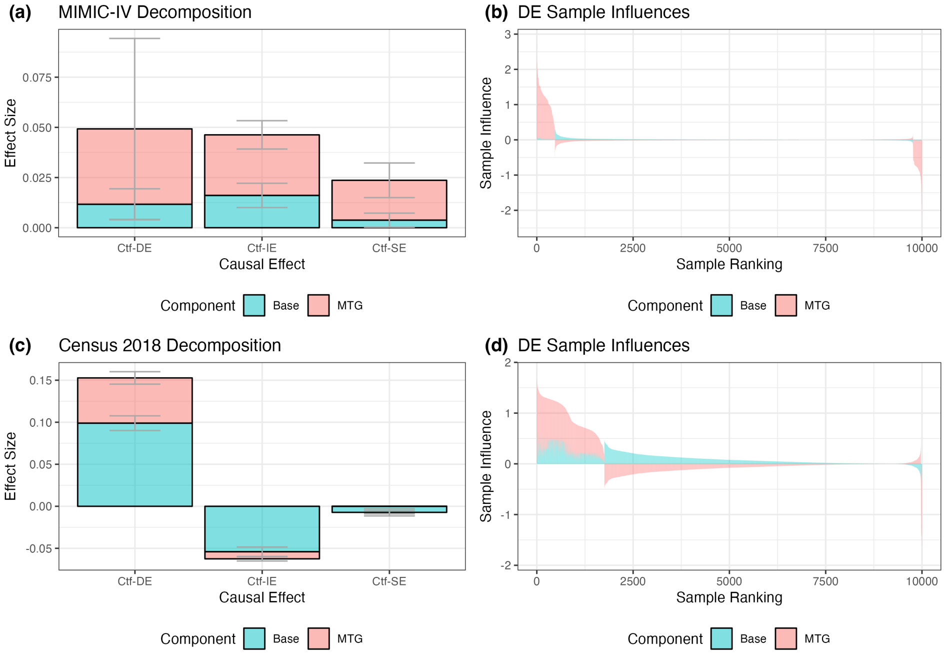

Figure 2: Causal decomposition from Cor. 3 and sample influence on the MIMIC-IV and Census 2018 datasets.

the i-th sample influence on x-DE$^{\text{est}}_{x_0,x_1}(t \mid x_0)$ is given by

$$\textit{SI-DE}(i) = [f(x_1, w_i, z_i) - f(x_0, w_i, z_i)]\frac{\hat{P}(x_0 \mid w_i, z_i)}{\hat{P}(x_0)}, \tag{25}$$

and analogously for the indirect and spurious effects.

The sample influences tell us how each of the samples contributes to the overall estimator of the quantity. These sample-level contributions may be interesting to investigate from the point of view of the system designer, to identify any subpopulations who are discriminated against. For direct sample influences, we prove the following proposition:

Proposition 6 (Direct Effect Sample Influence)**.** *The direct effect sample influence SI-DE(i) in Eq. 25 is an estimator of*

$$(x_0, z, w)\text{-}\textit{DE}_{x_0,x_1}(t \mid x_0, z_i, w_i)\frac{P(w_i, z_i \mid x_0)}{P(w_i, z_i)}, \tag{26}$$

where (x_0, z, w)-DE$_{x_0,x_1}(t \mid x, z, w) = \mathbb{E}[T_{x_1, W_{x_0}} \mid x_0, z, w] - \mathbb{E}[T_{x_0} \mid x_0, z, w]$ is the so-called (x_0, z, w)-specific direct effect of X on T.

Prop. 6 demonstrates an important point – namely that the sample influences along the direct path are not just quantities of a statistical interest, but also *causally* meaningful quantities. In particular, the influence of the i-th sample is proportional to the direct effect of the $x_0 \to x_1$ transition for the group of units u compatible with the event x_0, z_i, w_i. The influence is further proportional to $P(w_i, z_i \mid x_0)/P(w_i, z_i)$ that measures how much more likely the covariates z_i, w_i of the i-th sample are in the $X = x_0$ group (for which the discrimination is quantified) vs. the overall population. Therefore, practitioners may have both a valid statistical and a causal reason for investigating these sample influences. In this section, we demonstrate our approach on three real-world datasets, starting with an example from the medical domain:

5 Experiments

Example 2 (Acute Care Triage on MIMIC-IV Dataset [12])**.** *Clinicians in the Beth Israel Deaconess Medical Center in Boston, Massachusetts treat critically ill patients admitted to the intensive care unit*

(ICU). For all patients, 24 hours after admission various physiological and treatment information is collected. In particular, the information available to the clinicians consists of following (grouped into the Standard Fairness model):

- *protected atrribute X, in this case race (x_0 African-American, x_1 White),*
- *set of confounders Z consisting of {sex, age, chronic health status},*
- *set of mediators W consisting of {lactate, SOFA score, admission diagnosis, PaO$_2$/FiO$_2$ ratio, aspartate aminotransferase}.*

The clinicians are interested in which of the patients require further close monitoring. They want to determine the top half of the patients who are the most likely to (i) die during their hospital stay; (ii) have an ICU stay longer than 10 days, known as persistent critical illness. This combined outcome is labeled Y. These high risk patients will remain in the most acute care unit. To be able to better predict the outcome, clinicians use the electronic health records (EHR) data of the hospital, to construct a predictor of the outcome Y based on covariates X, Z, W. The clinicians are particularly interested in the fairness implications of using this automated system.

To investigate this, after constructing the score predictions S and a binary predictor $\widehat{Y} = \mathbb{1}(S(x, z, w) > \text{Quant}(0.5; S))$ that selects the top half of the patients, they use the decomposition described in Cor. 3 to investigate different contributions to the resulting disparity. The decomposition is shown in Fig. 2(a), and uncovers a number of important effects. Firstly, along the direct effect, x-DE$_{x_0,x_1}(y)$ and x-DE$_{x_0,x_1}(m)$ are larger than 0, meaning that minority group individuals have a lower chance of receiving acute care (purely based on race). Along the indirect and spurious effects, the situation is different: x-IE$_{x_1,x_0}(y)$ and x-SE$_{x_1,x_0}(y)$ and their respective margin complement contributions are different than 0 and negative – implying that minority group individuals have a larger probability of being given acute care as a result of confounding and mediating variables. Finally, the direct effect sample influences (2(b)) highlight that the margin complements are large for a small minority of individuals, requiring further subgroup investigation by the hospital team. □

We next move onto an application in the context of labor and salary decisions:

Example 3 (Salary Increase on the Census 2018 Dataset)**.** *The United States Census of 2018 collected broad information about the US Government employees, including demographic information Z (Z_1 for age, Z_2 for race, Z_3 for nationality), gender X (x_0 female, x_1 male), marital and family status M, education information L, and work-related information R. The US Government wishes to use this data prospectively to decide whether a new employee should receive a bonus starting package upon signing the employment contract. To determine which employees should receive such a bonus, a Government's department decides to predict which of the employees should earn a salary above the median (which is \$50,000/year). They construct a machine learning prediction score S that predicts above-median earnings (Y), and the department decides to allocate the bonus to all employees who are predicted to be above-median earners with a probability greater than 50% (i.e., $\widehat{Y} = \mathbb{1}(S \geq 0.5)$).*

A team of investigators in a different department within the Government is in charge of assessing what the impacts of this AI system are on the gender pay gap. They collect the required data, and perform the decomposition from Cor. 3, shown in Fig. 2(c). The decomposition indicates strong direct effects x-DE$_{x_0,x_1}(y)$ and x-DE$_{x_0,x_1}(m)$, which imply that men are more likely to receive a starting bonus than women. The indirect effects x-IE$_{x_1,x_0}(y)$, x-IE$_{x_1,x_0}(m)$ are both negative, with latter effect of the margin complement not being significant. These effects, however, also mean that men are more likely to receive a bonus due to mediating variables. Finally, for the spurious effects, the effects are not significantly different from 0. Based on these findings, the team decide to return the predictions to the original department, with the requirement that the direct effect of gender on the margin complement, x-DE$_{x_0,x_1}(m)$, must be reduced to 0, to avoid any possibility of bias amplification. □

Finally, we apply Alg. 1 to a well-known example from criminal justice:

Example 4 (Recidivism Prevention on the COMPAS Dataset [17])**.** *Courts in Broward County, Florida use machine learning algorithms, developed by a private company called 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*

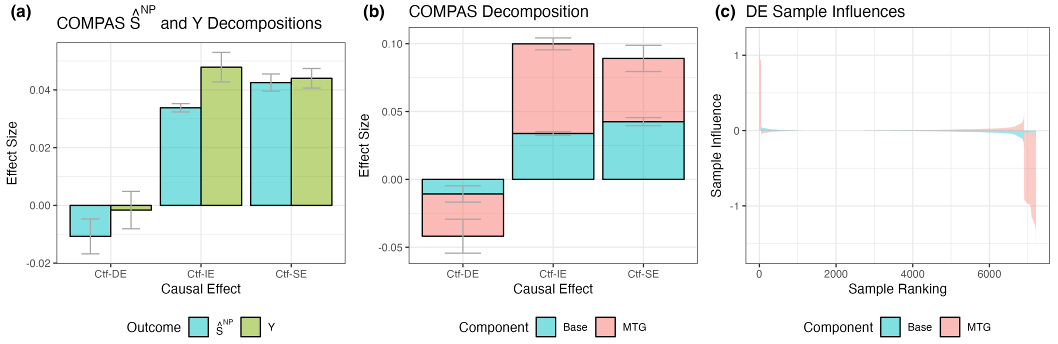

Figure 3: Application of Alg. 1 on the COMPAS dataset.

charge D. The courts wish to know which individuals are highly likely to recidivate, such that their probability of recidivism is above 50%. The company constructs a prediction scores $\hat{S}^{NP}$ and the court subsequently uses this for deciding whether to detain individuals at high risk of re-offending.

After a court hearing in which it was decided that the indirect and spurious effects fall under business necessity requirements, a team from ProPublica wishes to investigate the implications of using the automated predictions $\hat{S}^{NP}$. They obtain the relevant data and apply Alg. 1, with the results shown in Fig. 3. The team first compares the decompositions of the true outcome Y and the predictor $\hat{S}^{NP}$ (Fig. 3(a)). For the spurious effect, they find that x-SE$_{x_1,x_0}(y)$ is not statistically different from x-SE$_{x_1,x_0}(\hat{s}^{NP})$, in line with BN requirements. For the indirect effects, they find that the indirect effect is lower for the predictor $\hat{S}^{NP}$ compared to the true outcome Y, indicating no concerning violations. However, for the direct effect, while the x-DE$_{x_0,x_1}(y)$ is not statistically different from 0, the predictor $\hat{S}^{NP}$ has a significant direct effect of X, i.e., x-DE$_{x_0,x_1}(\hat{s}^{NP}) \neq 0$. This indicates a violation of the fairness requirements determined by the court.

After comparing the decompositions of $\hat{S}^{NP}$ and Y, the team moves onto understanding the contributions of the margin complements (Fig. 3b). As it turns out, for each of the three effects, there is a pronounced effect of the margin complements. For the direct effect, which does not fall under business necessity, the non-zero margin complement contribution x-DE$_{x_1,x_0}(m) \neq 0$ represents another violation of fairness requirements. However, for the indirect and spurious effects, the ProPublica team realizes the court did not specify anything about margin complement contributions – based on this, for the next court hearing they are preparing an argument showing that the effects x-IE$_{x_1,x_0}(m)$ and x-SE$_{x_1,x_0}(m)$ are significantly different from 0, thereby exacerbating the differences between groups. Finally, based on sample influences (Fig. 3c), they realize that the direct effect is driven by a small minority of individuals, and they decide to investigate this through a further analysis. □

6 Conclusion

In this paper, we developed new tools for understanding the fairness impacts of transforming a continuous prediction score S into binary predictions $\widehat{Y}$ or binary decisions D. In Thm. 1 and Cor. 3 we showed that the TV measure of the optimal 0/1 predictor decomposes into direct, indirect, and spurious contributions that are inherited from the true outcome Y in the real world, and also contributions from the margin complement M (Def. 2) arising from the automated optimization procedure. This observation motivated new notions of *weak* and *strong* business necessity (BN) – in the former case, differences inherited from the true outcome Y are allowed to be propagated into predictions or decision, while any differences resulting from the optimization procedure are disallowed. In contrast to this, strong BN allows both of these differences, and does not prohibit possible instances of disparity amplification. In Alg. 1 we introduced a formal procedure for assessing weak and strong BN. Finally, a range of real-world examples we investigated demonstrated that the considerations in this manuscript are of genuine importance in practice – even though the transformation of continuous predictions into binary decisions could in principle result either in bias amelioration or amplification (as illustrated in Ex. 1), we have observed empirically that it almost always leads to bias amplification – highlighting the need for performing this type of analysis, and the importance of regulatory oversight in this context.

References

[1] C. R. Act. Civil rights act of 1964. *Title VII, Equal Employment Opportunities*, 1964.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 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`.

[3] F. D. Blau and L. M. Kahn. The gender earnings gap: learning from international comparisons. *The American Economic Review*, 82(2):533–538, 1992.

[4] F. D. Blau and L. M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865, 2017.

[5] T. Brennan, W. Dieterich, and B. Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

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

[7] S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

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

[9] R. B. Darlington. Another look at "cultural fairness" 1. *Journal of educational measurement*, 8 (2):71–82, 1971.

[10] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015 (1):92–112, Apr. 2015. doi: 10.1515/popets-2015-0007.

[11] K. Imai, Z. Jiang, D. J. Greiner, R. Halen, and S. Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189, 2023.

[12] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L.-w. H. Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[13] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

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

[15] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[16] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[17] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[18] J. F. Mahoney and J. M. Mohen. Method and system for loan origination and underwriting, Oct. 23 2007. US Patent 7,287,008.

[19] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

[21] D. Pager. The mark of a criminal record. *American journal of sociology*, 108(5):937–975, 2003.

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

[23] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[24] D. Plečko and E. Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

[25] D. Plecko and E. Bareinboim. Causal fairness for outcome control. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] D. Plecko and E. Bareinboim. Reconciling predictive and statistical parity: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 (13), pages 14625–14632, 2024.

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

[28] J. Sanburn. Facebook thinks some native american names are inauthentic. *Time*, Feb. 14 2015. URL `http://time.com/3710203/facebook-native-american-names/`.

[29] L. Sweeney. Discrimination in online ad delivery. Technical Report 2208240, SSRN, Jan. 28 2013. URL `http://dx.doi.org/10.2139/ssrn.2208240`.

[30] L. T. Sweeney and C. Haney. The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, 10(2):179–195, 1992.

[31] Y. Wu, L. Zhang, X. Wu, and H. Tong. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in neural information processing systems*, 32, 2019.

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

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

[34] J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Acknowledgements This research was supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.