Fairness-Accuracy Trade-Offs: A Causal Perspective

Drago Plecko and **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
dp3144@columbia.edu, eb@cs.columbia.edu

Abstract

With the widespread adoption of AI systems, many of the decisions once made by humans are now delegated to automated systems. Recent works in the literature demonstrate that these automated systems, when used in socially sensitive domains, may exhibit discriminatory behavior based on sensitive characteristics such as gender, sex, religion, or race. In light of this, various notions of fairness and methods to quantify discrimination have been proposed, also leading to the development of numerous approaches for constructing fair predictors. At the same time, imposing fairness constraints may decrease the utility of the decision-maker, highlighting a tension between fairness and utility. This tension is also recognized in legal frameworks, for instance in the disparate impact doctrine of Title VII of the Civil Rights Act of 1964 – in which specific attention is given to considerations of *business necessity* – possibly allowing the usage of proxy variables associated with the sensitive attribute in case a high-enough utility cannot be achieved without them. In this work, we analyze the tension between fairness and accuracy from a causal lens for the first time. We introduce the notion of a path-specific excess loss (PSEL) that captures how much the predictor's loss increases when a causal fairness constraint is enforced. We then show that the total excess loss (TEL), defined as the difference between the loss of predictor fair along all causal pathways vs. an unconstrained predictor, can be decomposed into a sum of more local PSELs. At the same time, enforcing a causal constraint often reduces the disparity between demographic groups. Thus, we introduce a quantity that summarizes the fairness-utility trade-off, called the causal fairness/utility ratio, defined as the ratio of the reduction in discrimination vs. the excess in the loss from constraining a causal pathway. This quantity is particularly suitable for comparing the fairness-utility trade-off across different causal pathways. Finally, as our approach requires causally-constrained fair predictors, we introduce a new neural approach for causally-constrained fair learning. Our approach is evaluated across multiple real-world datasets, providing new insights into the tension between fairness and accuracy.

1 Introduction

Automated decision-making systems based on machine learning and artificial intelligence are now commonly implemented in various critical sectors of society such as hiring, university admissions, law enforcement, credit assessments, and health care [18, 24, 9]. These technologies now significantly influence the lives of individuals and are frequently used in high-stakes settings [41, 6, 40]. As these systems replace or augment human decision-making processes, concerns about fairness and bias based on protected attributes such as race, gender, or religion have become a prominent consideration in the ML literature. The available data used to train automated systems may contain past and present societal biases as an imprint and therefore has the potential to perpetuate or even exacerbate discrimination against protected groups. This is highlighted by reports on biases in systems for sentencing [3], facial recognition [10], online ads [38, 15], and system authentication [36], among

many others. Despite the promise of AI to enhance human decision-making, the reality is that these technologies can also reflect or worsen societal inequalities. As alluded to before, the issue does not arise uniquely from the usage of automated systems; human-driven decision-making has long been analyzed in a similar fashion. Evidence of bias in human decision-making is abundant, including studies on the gender wage gap [7, 8] and racial disparities in legal outcomes [39, 28]. Therefore, without proper care about fairness and transparency of the new generation of AI systems, it is unclear what its impact will be on the historically discriminated groups, and issues of inequity more broadly.

Within the growing literature on fair machine learning, a plethora of fairness definitions have been proposed. Commonly considered statistical criteria, among others, include demographic parity (independence [14]), equalized odds (separation [17]), and calibration (sufficiency [12]). These definitions, however, have been shown as mutually incompatible [5, 21]. Despite a number of proposals, there is still a lack of consensus on what the appropriate measures of fairness are, and how statistical notions of fairness could incorporate moral values of the society at large. For this reason, a number of works explored the causal approaches to fair machine learning [22, 19, 26, 45, 44, 42, 11, 34], and an in-depth discussion can be found in [31]. The main motivation for doing so is that the causal approach may allow the system designers to attribute the observed disparities between demographic groups to the causal mechanisms that underlie and generate them in the first place. In this way, by isolating disparities transmitted along different causal pathways, one obtains a more fine-grained analysis, and the capability to decide which causal pathways are deemed as unfair or discriminatory. More fundamentally, such considerations also form the basis of the legal frameworks for assessing discrimination in the United States and Europe. For instance, in the context of employment law, the disparate impact doctrine within the Title VII of the Civil Rights Act of 1964 [2] disallows any form of discrimination that results in a too large of a disparity between groups of interest. A core aspect of this doctrine, however, is the notion of *business necessity* (BN) or job-relatedness. 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). Often, the wording that is used is that to argue business necessity in front of a court of law, the plaintiff needs to demonstrate that "there is no practice that is less discriminatory and achieves the same utility" [1]. This concept illustrates the tension between fairness and utility, and demonstrates that we cannot be oblivious to considerations of utility from a legal standpoint.

The question of fairness-utility trade-offs has been explored in the fairness literature [13], with the essential argument that an unconstrained predictor always achieves a greater or equal utility than a constrained one. Despite this, the literature seems to be divided on this issue. For instance, some works argue that fairness and utility trade-offs are negligible in practice [35], while others argue that such trade-offs need not even exist [25, 16]. In this paper, we directly refute such arguments and demonstrate that, from a causal viewpoint, fairness and utility are always in a trade-off. The key subtlety in the works that find no trade-off [35, 25, 16] are the fairness metrics used to impose constraints. Naturally, the implications on the predictor's utility will strongly depend on the exact type of the fairness constraint that is enforced[1], and we demonstrate that imposing causal fairness constraints along different causal pathways almost always reduces predictive power.

Interestingly, the tension between fairness and utility has been largely unexplored in the causal fairness literature (with some exceptions such as [27, 32, 31]). Our main aim of this paper is to fill in this gap, and provide a systematic way of analyzing the fairness-accuracy trade-off from a causal lens. We illustrate our approach in a simple linear setting:

Example 1 (Linear Fairness-Accuracy Causal Trade-Offs)**.** *Consider variables $X; W; Y$ behaving according to the following system of equations:*

$$X \quad Bernoulli(0.5) \tag{1}$$

$$W \quad X + {}_{w} \tag{2}$$

$$Y \quad X + W + {}_{y}; \tag{3}$$

where ${}_{w} \quad N(0; {}^{2}_{w})$; ${}_{y} \quad N(0; {}^{2}_{y})$. Variable X is the protected attribute, and Y is the outcome of interest. The causal diagram of Eqs. 1 -3 is shown in Fig. 2 (with the Z set empty). Attribute

[1]Works that discuss how fairness-accuracy trade-offs do not exist or are insignificant in practice often focus on equality of odds [17] ($\hat{Y} \perp X \mid Y$). Notably, this metric always allows for the perfect predictor $\hat{Y} = Y$, and thus in settings with good predictive power, the cost of enforcing this constraint may indeed be negligible.

X can influence Y along two different pathways: the direct path $X \to Y$, and the indirect path $X \to W \to Y$. Therefore, we consider fair predictors $\hat{Y}^S$ of the form $\hat{Y}^S = \hat{\alpha}_S X + \hat{\alpha}_S W$, where the predictor $\hat{Y}^S$ removes effects in the set S, with S ranging in $\{\emptyset, \text{DE}, \text{IE}, \{\text{DE}, \text{IE}\}\}$ (DE, IE stand for direct and indirect effects). For instance, the optimal predictor $\hat{Y}$ has $\hat{\alpha}_\emptyset = \beta_{Y|X}$, $\hat{\alpha}_\emptyset = \beta_{Y|W \cdot X}$, and therefore its mean-squared error (MSE) equals $E[Y - \hat{Y}]^2 = \sigma^2_{Y|X,W}$. The DE-fair predictor has $\hat{\alpha}_{\text{DE}} = 0, \hat{\alpha}_{\text{DE}} = \beta_{Y|W \cdot X}$, and the fully-fair predictor $\hat{Y}^{\{\text{DE}, \text{IE}\}}$ has $\hat{\alpha}_{\{\text{DE}, \text{IE}\}} = 0, \hat{\alpha}_{\{\text{DE}, \text{IE}\}} = 0$. Thus, the corresponding MSE values can be written as (see Appendix A for details):

$$E[Y - \hat{Y}^{\text{DE}}]^2 = \sigma^2_{Y|X,W} + \frac{\sigma^2_{W|X}}{2}, \quad E[Y - \hat{Y}^{\{\text{DE}, \text{IE}\}}]^2 = \sigma^2_{Y|X,W} + \frac{\beta^2_{Y|W \cdot X}\sigma^2_{W|X} + \beta^2_{W|X}\sigma^2_{Y|X,W} + \beta^2_{Y|W \cdot X}\beta^2_{W|X}\sigma^2_{W|X}}{2} + \beta^2_{Y|W \cdot X}\beta^2_{W|X}\sigma^2_{W|X}. \quad (4)$$

Our goal is to decompose the total excess loss originating from imposing the fairness constraints into the path-specific contributions:

$$\underbrace{E[Y - \hat{Y}^{\{\text{DE}, \text{IE}\}}]^2}_{\text{fully-fair predictor's loss}} - \underbrace{E[Y - \hat{Y}^{\emptyset}]^2}_{\text{unconstrained loss}} = \underbrace{\frac{\beta^2_{Y|W \cdot X}\sigma^2_{W|X}}{2}}_{\text{Term I}} + \underbrace{\frac{\beta^2_{Y|W \cdot X}\sigma^2_{W|X} + \beta^2_{W|X}\sigma^2_{Y|X,W} + \beta^2_{Y|W \cdot X}\beta^2_{W|X}\sigma^2_{W|X}}{2} + \beta^2_{Y|W \cdot X}\beta^2_{W|X}\sigma^2_{W|X}}_{\text{Term II}} \quad (5)$$

$$= \underbrace{E[Y - \hat{Y}^{\text{DE}}]^2 - E[Y - \hat{Y}^{\emptyset}]^2}_{\text{Term I = excess DE loss}} \quad (6)$$

$$+ \underbrace{E[Y - \hat{Y}^{\{\text{DE}, \text{IE}\}}]^2 - E[Y - \hat{Y}^{\text{DE}}]^2}_{\text{Term II = excess IE loss}}. \quad (7)$$

At the same time, for each pathway, we can also observe the decrease in group disparity associated with removing the effect along the path, by computing

$$\underbrace{(E[\hat{Y}^{\{\text{DE}, \text{IE}\}} \mid x_1] - E[\hat{Y}^{\{\text{DE}, \text{IE}\}} \mid x_0])}_{\text{disparity after removing DE, IE}} \quad \underbrace{(E[Y^{\emptyset} \mid x_0] - E[Y^{\emptyset} \mid x_0])}_{\text{disparity before removing DE, IE}} = \quad ; \quad (8)$$

which again decomposes into the contributions along direct and indirect effects:

$$(E[\hat{Y}^{\text{DE}} \mid x_0] - E[\hat{Y}^{\text{DE}} \mid x_0]) \quad (E[\hat{Y}^{\emptyset} \mid x_0] - E[\hat{Y}^{\emptyset} \mid x_0]) = \quad (9)$$

$$(E[\hat{Y}^{\{\text{DE}, \text{IE}\}} \mid x_1] - E[\hat{Y}^{\{\text{DE}, \text{IE}\}} \mid x_0]) \quad (E[\hat{Y}^{\text{DE}} \mid x_0] - E[\hat{Y}^{\text{DE}} \mid x_0]) = \quad : \quad (10)$$

Prototypical fairness/utility trade-offs for predictors $\hat{Y}^S$ are visualized in Fig. 1, for three randomly drawn triples $(\beta_{Y|X}; \beta_{Y|W \cdot X}; \beta_{W|X})$. Predictor $\hat{Y}^{\emptyset}$ is optimal and thus always has 0 excess loss (and lies on the vertical axis). Fully-fair predictor $\hat{Y}^{\{\text{DE}, \text{IE}\}}$ removes both direct and indirect effects, and thus always has 0 TV measure (and lies on the vertical axis). The slopes in the plot between $\hat{Y}^{\emptyset}$, $\hat{Y}^{\text{DE}}$, and $\hat{Y}^{\text{DE}, \text{IE}}$ (indicated by arrows), geometrically capture the tension between excess loss and reducing discrimination upon imposing a constraint. These slopes are computed as $\frac{\text{TV difference}}{\text{Excess Loss}}$, a quantity that we call the causal fairness-utility ratio (CFUR). Based on Eqs. 6-7 and 9-10 we can compute that

$$\text{CFUR}(\text{DE}) = \frac{1}{\quad}; \quad \text{CFUR}(\text{IE}) = \frac{2}{\beta^2_{Y|W \cdot X}\sigma^2_{W|X} + \beta^2_{W|X}\sigma^2_{Y|X,W} + 2\beta^2_{Y|W \cdot X}\beta^2_{W|X}\sigma^2_{W|X}}; \quad (11)$$

summarizing the fairness-utility trade-off for each pathway.

The above example illustrates how in the simple linear case we can attribute the increased loss from imposing fairness constraints to the specific causal pathway in question. It also shows we can compute the associated change in the disparity between groups, measured by the so-called TV measure $E[\hat{Y} \mid x_1] - E[\hat{Y} \mid x_0]$ (also known as the *disparity gap*), a result known under the rubric of causal decompositions [45, 31] (this connection is explored explicitly in Appendix C). With the above example in mind, we list below the key contributions of this manuscript:

Figure 1: TV vs. Excess Loss.

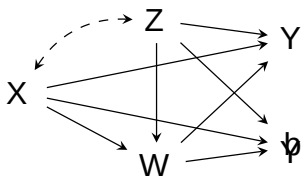


Figure 2: Standard Fairness Model.

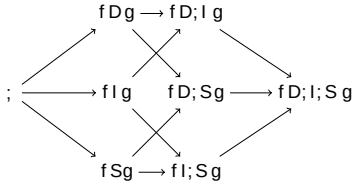


Figure 3: Graphical representation of G_{PSEL}.

(i) We introduce the notion of path-specific excess loss (PSEL) associated with imposing a fairness constraint along a causal path (Def. 3), and we prove how the total excess loss (TEL) can be decomposed into a sum of path-specific excess losses (Thm. 1),

(ii) We develop an algorithm for attributing path-specific excess losses to different causal paths (Alg. 1), allowing the system designer to explain how the total excess loss is affected by different fairness constraints. In this context, we show the equivalence of Alg. 1 with a Shapley value [37] approach (Prop 3),

(iii) For purposes of applying Alg. 1, a key requirement is the construction of causally-fair predictors $\hat{Y}_S$ that remove effects along pathways S. We introduce a novel Lagrangian formulation of the optimization problem for such $\hat{Y}_S$ (Def. 5) accompanied with a training procedure for learning the predictor (Alg. 2),

(iv) We introduce the causal fairness/utility ratio (CFUR, Def. 4) that summarizes how much the group disparity can be reduced per unit fixed cost in terms of excess loss. We compute CFURs on a range of real-world datasets, and demonstrate that from a causal viewpoint fairness and utility are almost always in tension.

1.1 Preliminaries

We use the language of structural causal models (SCMs) [29]. An SCM is a tuple $M := \langle V, U, F, P(u) \rangle$, where V, U are sets of endogenous (observable) and exogenous (latent) variables, respectively, F is a set of functions $\{f_{V_i}\}$, one for each $V_i \in V$, where $V_i = f_{V_i}(pa(V_i), U_{V_i})$ for some $pa(V_i) \subseteq V$ and $U_{V_i} \subseteq U$. $P(u)$ is a strictly positive probability measure over U. Each SCM M is associated to a causal diagram G [4] over the node set V where $V_i \to V_j$ 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 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 M_x, in which all equations in F associated with X are replaced by $X = x$. Throughout the paper, we assume a specific cluster causal diagram G_{SFM} known as the standard fairness model (SFM) [31] over endogenous variables $\{X, Z, W, Y, \hat{Y}\}$ shown in Fig. 2. 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 $\hat{Y}$ (e.g., predicted GPA, predicted salary). The SFM encodes the assumptions typically used in the causal inference literature about the lack of hidden confounding. The availability of the SFM and the implied assumptions are a possible limitation of the paper, while we note that partial identification techniques for bounding effects can be used for relaxing them [46]. Based on the SFM, we will use the following causal fairness measures:

Definition 1 (Population-level Causal Fairness Measures [30, 31]). The natural direct, indirect, and spurious effects are defined as

$$NDE_{x_0,x_1}(y) = P(y_{x_1,w_{x_0}}) - P(y_{x_0}) \quad (12)$$

$$NIE_{x_1,x_0}(y) = P(y_{x_1,w_{x_0}}) - P(y_{x_1}) \quad (13)$$

$$NSE_x(y) = P(y \mid x) - P(y_x). \quad (14)$$

A causally-fair predictor for a subset S of the above measures is defined as:

De nition 2 (Causally Fair Predictor [33]). The causally S-fair predictor $\hat{Y}^S$ with respect to a loss function L and pathways in S is the solution to the following optimization problem:

$$\hat{Y}^S = \arg\min_f \; E\,L(Y; f(X, Z, W)) \tag{15}$$

$$\text{subject to} \quad NDE_{x_0,x_1}(f) = NDE_{x_0,x_1}(y) \;\; \mathbb{1}(DE \not\subseteq S) \tag{16}$$

$$NIE_{x_1,x_0}(f) = NIE_{x_1,x_0}(y) \;\; \mathbb{1}(IE \not\subseteq S) \tag{17}$$

$$NSE_{x_0}(f) = NSE_{x_0}(y) \;\; \mathbb{1}(SE \not\subseteq S) \tag{18}$$

$$NSE_{x_1}(f) = NSE_{x_1}(y) \;\; \mathbb{1}(SE \not\subseteq S). \tag{19}$$

The de nition of $\hat{Y}^S$ has a straightforward interpretation. For any pathway in the set S, the corresponding causal effect should be the same as proposed in the path-speci c causal fairness literature [26, 11]. However, importantly, pathways that are not in S also need to be constrained – the effect of X on $\hat{Y}$ along these paths should not change compared to the true outcome Y [33]. For instance, if the direct path is not in S, then we expect to have $NDE_{x_0,x_1}(\hat{y}) = NDE_{x_0,x_1}(y)$ (and similarly for IE, SE).

2 Path-Speci c Excess Loss

In this section, we introduce the concept of a path-speci c excess loss, and then demonstrate how the total excess loss can be decomposed into path-speci c excess losses.

De nition 3 (Path-Speci c Excess Loss). Let $L(\hat{Y}, Y)$ be a loss function and $\hat{Y}^S$ the optimal S-fair predictor with respect to L. De ne the path-speci c excess loss (PSEL) of a pair S, S' as:

$$\text{PSEL}(S \to S') = E[L(\hat{Y}^{S'}, Y)] \quad E[L(\hat{Y}^S, Y)]. \tag{20}$$

The quantity $\text{PSEL}(\emptyset \to \{D, I, S\})$ is called the total excess loss (TEL).

The total excess loss computes the increase in the loss for the totally constrained predictor $\hat{Y}^{\{D,I,S\}}$ with direct, indirect, and spurious effects removed compared to the unconstrained predictor $\hat{Y}^{\{\}} = \hat{Y}$. The total excess loss can be decomposed as a sum of path-speci c excess losses (all proofs are provided in Appendix B):

Theorem 1 (Total Excess Loss Decomposition). The total excess loss $\text{PSEL}(\emptyset \to \{D, I, S\})$ can be decomposed into a sum of path-speci c excess losses as follows:

$$\text{PSEL}(\emptyset \to \{D, I, S\}) = \text{PSEL}(\emptyset \to \{D\}) + \text{PSEL}(\{D\} \to \{D, I\}) \tag{21}$$

$$+ \text{PSEL}(\{D, I\} \to \{D, I, S\}). \tag{22}$$

Remark 2 (Non-Uniquness of Decomposition). The decomposition in Thm. 1 is not unique. In particular, the $\text{PSEL}(\emptyset \to \{D, I, S\})$ can be decomposed as

$$\text{PSEL}(\emptyset \to \{S_1\}) + \text{PSEL}(\{S_1\} \to \{S_1, S_2\}) + \text{PSEL}(\{S_1, S_2\} \to \{D, I, S\}) \tag{23}$$

for any choice of $S_1, S_2 \in \{D, I, S\}$ with $S_1 \neq S_2$. Therefore, six different decompositions exist (three choices for S_1, two for S_2).

Fig. 3 provides a graphical overview of all the possible path-speci c excess loses. In the left side, we start with $S = \emptyset$ and the predictor $\hat{Y}^\emptyset$. Then, we can add any of $\{D, I, S\}$ to the S-set, to obtain the predictors $\hat{Y}^D$, $\hat{Y}^I$, or $\hat{Y}^S$, and so on. The graph representing all the possible states and transitions between pairs $(\hat{Y}^S, \hat{Y}^{S'})$ shown in Fig. 3 is labeled G_{PSEL}. There are six paths starting from $\emptyset$ and ending in $\{D, I, S\}$. In Alg. 1, we introduce a procedure that sweeps over all the edges and paths in G_{PSEL} to compute path-speci c excess losses, while also computing the change in the TV measure between groups in order to track the reduction in discrimination. Formally, for any edge (S, S') in G_{PSEL} the value of $\text{PSEL}(S \to S')$ is computed, together with the difference in the TV measure (TVD) from the transition $S \to S'$, de ned by

$$\text{TVD}(S \to S') = \underbrace{E[Y^{S'} \mid x_1] - E[Y^{S'} \mid x_0]}_{\text{TV after removing } S' \cap S} \quad \underbrace{E[Y^S \mid x_1] - E[Y^S \mid x_0]}_{\text{TV before removing } S' \cap S}. \tag{26}$$

[2] Generally, other classi ers that remove only subsets of the causal paths between X and Y may be considered fair, depending on considerations of business necessity [33]. Still, the rationale developed in this paper can be easily adapted to such settings.

Algorithm 1: Path-Specific Excess Loss Attributions

Input: data D, predictors $\hat{Y}_S$ for S-sets $\{\emptyset, I, S, S\cup I\}$

foreach edge $(S, S') \in G_{PSEL}$ do
  compute the path-specific excess loss of S' in S, given by $PSEL(S \to S')$
  compute the TV measure difference of S' in S, written $TVD(S \to S')$ given by
  $TV_{x_0,x_1}(\hat{Y}_{S'}) - TV_{x_0,x_1}(\hat{Y}_S)$

foreach causal path $S_i \in \{\emptyset, I, S, S\cup I\}$ do
  compute the average path-specific excess loss and TV difference across all paths $\emptyset \to \cdots \to \{\emptyset, I, S, S\cup I\}$,

$$APSEL(S_i) = \frac{1}{3!} \sum_{\substack{\text{paths } \in G_{PSEL}: \\ \emptyset \to \cdots \to \{\emptyset, I, S, S\cup I\}}} PSEL(\text{}^{<S_i} \to \text{}^{<S_i} \cup S_i) \tag{24}$$

$$ATVD(S_i) = \frac{1}{3!} \sum_{\substack{\text{paths } \in G_{PSEL}: \\ \emptyset \to \cdots \to \{\emptyset, I, S, S\cup I\}}} TVD(\text{}^{<S_i} \to \text{}^{<S_i} \cup S_i); \tag{25}$$

  where ${}^{<S_i}$ denotes the set of causal paths that precede S_i in the ordering.

return set of $PSEL(S \to S')$, $TVD(S \to S')$, attributions $APSEL(S_i)$, $ATVD(S_i)$

The quantities $PSEL(S \to S')$ and $TVD(S \to S')$ are naturally associated with the effect that was removed, i.e., $S' \cap S$. As there are multiple ways of reaching the set $\{\emptyset, I, S, S\cup I\}$ from $\emptyset$ in G_{PSEL}, each of the corresponding effects (direct, indirect, spurious) will be associated with a number of different PSELs and TVDs (generally, note that the complexity is exponential in the number of causal paths included). In Eqs. 24-25 inside the algorithm, we compute the average PSEL and TVD across all the edges that are associated with a specific effect S_i. This simple intuition, corresponding to taking an average across all of the possible decompositions of the total excess loss (Eq. 23), turns out to be equivalent to a Shapley value [37] of a suitably chosen value function:

Proposition 3 (PSEL Attribution as Shapley Values)**.** Let the functions $f_1(S), f_2(S)$ be defined as:

$$f_1(S) = PSEL(\emptyset \to S), f_2(S) = TVD(\emptyset \to S). \tag{27}$$

The Shapley value for the effect $S_i \in \{\emptyset, I, S, S\cup I\}$ and function f_k, is computed as

$$\phi_k(S_i) = \sum_{S \subseteq \{\emptyset, I, S\} \setminus S_i} \frac{1}{n \binom{n-1}{|S|}} \left(f_k(S \cup S_i) - f_k(S)\right). \tag{28}$$

where $n = 3$ for the choice $\{\emptyset, I, S\}$. The averaged path-specific excess loss of S_i and the averaged TV difference of S_i are equal to the Shapley values of S_i associated with functions f_1, f_2, respectively,

$$\phi_1(S_i) = APSEL(S_i), \quad \phi_2(S_i) = ATVD(S_i), \tag{29}$$

with $APSEL(S_i), ATVD(S_i)$ defined in Eqs. 24, 25, respectively.

The above proposition illustrates how averaging the influence of removing a causal effect over all possible ways of reaching $\{\emptyset, I, S\}$ from $\emptyset$ is equivalent to computing the Shapley values of an appropriate value function. We next introduce the notion of a causal fairness/utility ratios:

Definition 4 (Causal Fairness/Utility Ratio (CFUR))**.** The causal fairness/utility ratio (CFUR) for a causal path S_i is defined as

$$CFUR(S_i) = \frac{ATVD(S_i)}{APSEL(S_i)}. \tag{30}$$

The CFUR quantity may be particularly useful for comparing different causal effects, and the connection of the CFUR with local TVD and PSEL values is described in Appendix D. The intuition behind the quantity is simple – for removing a causal effect S_i from our predictor $\hat{Y}$, we want to compute how much of a reduction in the disparity that results in (measured in terms of the ATVD measure) per unit change in the incurred excess loss. This quantity attempts to assign a single number to a causal path that succinctly summarizes how much fairness is gained vs. how much predictive power is lost by imposing such a causal constraint.

Algorithm 2: Causally-Fair Constrained Learning (CFCL)

Input: training data $\mathcal{D}_t$, evaluation data $\mathcal{D}_e$, precision ε

Set $\varepsilon_{low} = 0$, $\varepsilon_{high} = $ large
while $|\varepsilon_{high} - \varepsilon_{low}| > \varepsilon$ do
  set $\varepsilon_{mid} = \frac{1}{2}(\varepsilon_{low} + \varepsilon_{high})$
  fit a neural network to solves the optimization problem in Eqs. 32-34 with $\varepsilon = \varepsilon_{mid}$ on $\mathcal{D}_t$ to obtain the predictor $\hat{y}^S(\varepsilon_{mid})$
  compute the causal measures of fairness NDE, NIE, NSE of $\hat{y}^S(\varepsilon_{mid})$ on evaluation data $\mathcal{D}_e$
  test the hypothesis

$$H_0^{CE} : \text{NCE}(\hat{y}^S(\varepsilon_{mid})) = \text{NCE}(y) \quad 1(\text{CE} \geq \varepsilon) \tag{31}$$

  where NCE ranges in $\text{NDE}_{x_0,x_1}, \text{NIE}_{x_1,x_0}, \text{NSE}_{x_0}, \text{NSE}_{x_1}$
  if any of $H_0^{DE}, H_0^{IE}, H_0^{SE_0}, H_0^{SE_1}$ rejected then $\varepsilon_{low} = \varepsilon_{mid}$ else $\varepsilon_{high} = \varepsilon_{mid}$
return predictor $\hat{y}^S(\varepsilon_{mid})$

3 Causally-Fair Constrained Learning

In the preceding section, we developed an approach for quantifying the tension between fairness and accuracy from a causal viewpoint. The results were contingent on finding the optimal causally-fair predictors $\hat{y}^S$ following Def. 2. However, computing the predictors $\hat{y}^S$ is quite challenging in practice, due to several complex causal constraints in the optimization problem. In this section we develop a practical approach for solving this problem. We begin by introducing a Lagrangian form of the optimal causally-fair predictor:

Definition 5 (Causal Lagrange Predictor). The causally S-fair ε-optimal predictor $\hat{y}^S(\varepsilon)$ with respect to pathways in S and the loss function L is the solution to the following optimization problem:

$$\arg\min_f E[L(Y, f(X, Z, W))] + \big|\text{NDE}_{x_0,x_1}(f) - \text{NDE}_{x_0,x_1}(y) - 1(\text{DE} \geq \varepsilon)\big|^2 + \tag{32}$$

$$\big|\text{NIE}_{x_1,x_0}(f) - \text{NIE}_{x_1,x_0}(y) - 1(\text{IE} \geq \varepsilon)\big|^2 + \tag{33}$$

$$\big|\text{NSE}_{x_0}(f) - \text{NSE}_{x_0}(y) - 1(\text{SE} \geq \varepsilon)\big|^2 + \big|\text{NSE}_{x_1}(f) - \text{NSE}_{x_1}(y) - 1(\text{SE} \geq \varepsilon)\big|^2 \tag{34}$$

The above definition reformulates the problem of finding $\hat{y}^S$ to a Lagrangian form. This makes the problem amenable to standard gradient descent methods, and we propose a procedure for finding a suitable predictor $\hat{y}^S$ in Alg. 2 (CFCL). CFCL performs a binary search to find an appropriate value of the ε parameter. For an interval $[\varepsilon_{low}, \varepsilon_{high}]$ it takes the midpoint ε_{mid}. For this parameter value, it computes the optimal predictor $\hat{y}^S(\varepsilon_{mid})$ for the optimization problem in Eqs. 32-34 by fitting a feed-forward neural network with h_h hidden layers and n_v nodes in each layer. After this, the causal measures of fairness are computed for the predictor $\hat{y}^S(\varepsilon_{mid})$ on an evaluation set, and these measures are compared to the causal measures for the true outcome Y (CFCL performs the appropriate hypothesis tests as in Eq. 31). If none of the hypotheses are rejected, it means that ε_{mid} is large enough to enforce the causal fairness constraints, and the algorithm moves to the interval $[\varepsilon_{low}, \varepsilon_{mid}]$. If a hypothesis is rejected, it indicates that ε_{mid} is not large enough, and the algorithm moves to the interval $[\varepsilon_{mid}, \varepsilon_{high}]$. Thus, CFCL leads to a systematic search for the parameter ε.

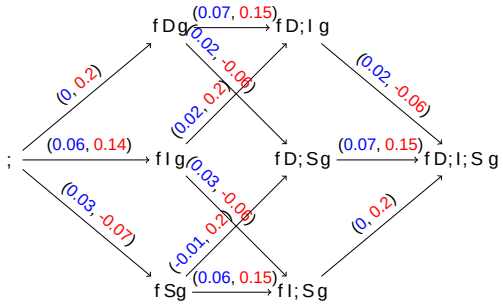
4 Experiments

In this section, we perform the causal fairness-accuracy analysis described in Sec. 2. We analyze the Census 2018 (Ex. 2), COMPAS (Ex. 3), and UCI Credit (Ex. 4 in Appendix E) datasets.

Example 2 (Salary Increase of Government Employees [31]). The US government is building a tool for automated allocation of salaries for new employees. For developing the tool, they use the data collected by the United States Census Bureau in 2018, 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,

(a) (A)PSEL and (A)TVD values.

(b) CFUR values.



(c) Fairness/Utility of $\hat{Y}^S$ predictors.

(d) G_{PSEL} with PSEL (blue) and TVD values (red).

Figure 4: Application of Alg. 1 on the Census 2018 dataset.

education L, and work-related information R. The government wants to predict the outcome Y, the yearly salary of the employees (transformed to a log-scale). The standard fairness model (Fig. 2) is constructed as $\{X = X; Z = \{Z_1; Z_2; Z_3\}; W = \{M; L; R\}; Y = Y\}$.

The team developing the ML predictor is also concerned with the fairness of the allocated salaries. In particular, they wish to understand how the different causal effects from the protected attribute X to the predictor $\hat{Y}$ affect the prediction, and how much the salary predictions would have to deviate from the optimal prediction to remove an effect along a specific pathway (in particular, they focus on the root mean squared error (RMSE) loss). For analyzing this, they utilize the tools from Alg. 1, and build causally fair predictors $\hat{Y}^S$ (for different choices of S-sets) using Alg. 2.

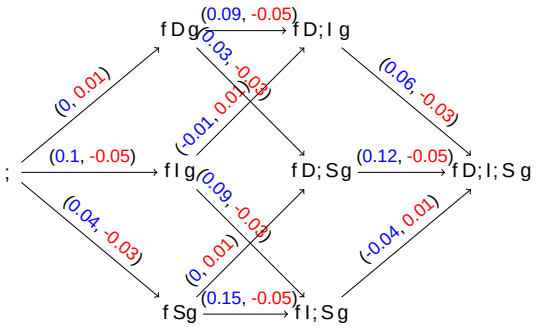
The analysis results are shown in Fig. 4, with uncertainty bars indicating standard deviations over 10 bootstrap repetitions. In the analysis of PSEL and TVD values (Fig. 4a), the team notices that imposing fairness constraints does not reduce RMSE, with the largest excess loss for the indirect effect, and smaller excess losses for direct and spurious effects. When looking at TVD values, they find that removing direct and indirect effects reduces the group differences substantially. In terms of causal fairness-utility ratios (Fig. 4b), the team finds that removing the direct effect has the best value in terms of reducing the disparity between groups vs. increasing the loss. The TV measure vs. excess loss dependence for different predictors $\hat{Y}^S$ is shown in Fig. 4c (binary labels (D, I, S) in the figure indicate which effects were removed), and the graph G_{PSEL} with the values of PSEL and TVD for each transition is shown in Fig. 4d. Based on the analysis, the team decides to use the predictor with the direct effect removed $\hat{Y}^D$.

Example 3 (Recidivism Prediction on COMPAS [23]). Courts in Broward County, Florida use machine learning algorithms, developed by a private company, 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. The grouping $\{X = X; Z = \{Z_1; Z_2\}; W = \{J; P; D\}; Y = Y\}$ constructs the standard fairness model.

The team from ProPublica wishes to understand the tension between fairness and accuracy, since the latter is an important concern of the district court. For the analysis, they use the complement

(a) (A)PSEL and (A)TVD values.

(b) CFUR values.



(c) Fairness/Utility of $\hat{Y}^{S}$ predictors.

(d) $\mathcal{G}_{\text{PSEL}}$ with PSEL (blue) and TVD values (red).

Figure 5: Application of Alg. 1 on the COMPAS dataset.

of the area under the receiver operator characteristic curve (AUROC). They apply Alg. 1, and find that the APSEL value is the largest for the indirect effect (around 12% of AUROC), while it is smaller for the spurious effect (6%), and negligible for the direct effect, implying that the direct effect does not play a significant role in the prediction (see Fig. 5a). Regarding TVD values, the team finds that the removal of the spurious effect results in a 3% decrease in the TV measure, while the removal of the indirect effect decreases the disparity by 5%. Based on this, they find that the spurious effect is the best for reducing the disparity between groups (see Fig. 5b for CFUR values). Finally, the team also visualizes the TV and excess loss dependence (Fig. 5c), and the graph $\mathcal{G}_{\text{PSEL}}$ with PSEL and TVD values associated with each transition and effect removal (Fig. 5a). In the upcoming court hearing, the team proposes the usage of the causally-constrained predictor $\hat{Y}_{f_{D,S}}$ that removes direct and spurious effects, and they use the above analysis to demonstrate the impact of this choice on the predictor's accuracy.

5 Conclusion

The tension between fairness and accuracy is a fundamental concern in the modern applications of machine learning. The importance of this tension is also recognized in the legal frameworks of anti-discrimination, such as the disparate impact doctrine, which may allow for the usage of covariates correlated with the protected attribute if they are sufficiently important for the decision-maker's utility (this concept is known as business necessity). In this work, we developed tools for analyzing the fairness-accuracy trade-off from a causal standpoint. Our approach allows the system designer to quantify how much excess loss is incurred when removing a path-specific causal effect (Def. 3). We also showed how the total excess loss, defined as the difference between the loss of the predictor fair along all causal pathways vs. an unconstrained predictor, can be decomposed into a sum of path-specific excess losses (Thm. 1). Based on this, we developed an algorithm for attributing excess loss to different causal pathways (Alg. 1), and introduced the notion of a causal fairness-utility ratio that captures the $\frac{\text{fairness gain}}{\text{excess loss}}$ ratio and in this way summarizes the trade-off for each causal path. Since our approach is dependent on causally-fair predictors (Def. 2), we introduced a new neural approach for constructing such predictors (Def. 5, Alg. 2). Finally, we demonstrated empirically that from causal perspective fairness and accuracy are almost always in tension (Exs. 2-4).

References

[1] Elston v. Talladega County Bd. of Educ. 997 F.2d 1394 (11th Cir. 1993), 1993. United States Court of Appeals for the Eleventh Circuit.

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

[3] 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.

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

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

[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1):3–44, 2021.

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

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

[9] 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.

[10] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classi cation. 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.

[11] S. Chiappa. Path-speci c counterfactual fairness. In Proceedings of the AAAI Conference on Arti cial Intelligence, volume 33, pages 7801–7808, 2019.

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

[13] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pages 797–806, 2017.

[14] R. B. Darlington. Another look at “cultural fairness”. Journal of educational measurement, 8 (2):71–82, 1971.

[15] 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.

[16] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In International conference on machine learning, pages 2803–2813. PMLR, 2020.

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

[18] 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.

[19] 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.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

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

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

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

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

[25] S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun. There is no trade-off: enforcing fairness can improve accuracy, 2020.

[26] R. Nabi and I. Shpitser. Fair inference on outcomes. In Proceedings of the AAAI Conference on Arti cial Intelligence, volume 32, 2018.

[27] 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.

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

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

[30] J. Pearl. Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Arti cial Intelligence, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

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

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

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

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

[35] K. T. Rodolfa, H. Lamba, and R. Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nature Machine Intelligence, 3(10):896–904, 2021.

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

[37] L. S. Shapley et al. A value for n-person games. Princeton University Press Princeton, 1953.

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

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

[40] M. Taddeo and L. Floridi. How ai can be a force for good. Science, 361(6404):751–752, 2018.

[41] E. J. Topol. High-performance medicine: the convergence of human and arti cial intelligence. Nature medicine, 25(1):44–56, 2019.

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

[43] I.-C. Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

[44] J. Zhang and E. Bareinboim. Equality of opportunity in classi cation: 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, pages 3671–3681, Montreal, Canada, 2018. Curran Associates, Inc.

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

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

Technical Appendices for Fairness-Accuracy Trade-Offs: A Causal Perspective

The source code for reproducing all the experiments can be found in the anonymized code repository. The code is also included with the supplementary materials, in the folder source-code. All experiments were performed on a MacBook Pro, with the M3 Pro chip and 36 GB RAM on macOS 14.1 (Sonoma). Each experiment can be run with less than 1 hour of compute on the above-described machine or equivalent. Within each experiment, calls to Alg. 2 are included, which includes optimizing neural networks (using pytorch). Here, we use feed-forward networks with $n_{\text{hidden}} = 2$ hidden layers, and $d_{\text{hidden}} = 16$ nodes in each layer. Training is carried out using the Adam optimizer [20], with the learning rate $= 0.001$ fixed, for 500 epochs with batch size 512, and early stopping regularization (20 epochs of patience), and 5 random initial restarts.

A Linear Path-Specific Excess Losses

Consider the SCM in Eqs. 1-3. The unconstrained, DE-fair, IE-fair, and fully-fair predictors are given by

$$\hat{f}_{\varnothing} = X + W \tag{35}$$

$$\hat{f}_{\text{DE}} = W \tag{36}$$

$$\hat{f}_{\text{IE}} = X \tag{37}$$

$$\hat{f}_{\{\text{DE, IE}\}} = 0. \tag{38}$$

Note that we can compute:

$$\begin{aligned}
E[Y - \hat{f}_{\{\text{DE, IE}\}}]^2 &= EY^2 && (39)\\
&= E[X + W + \varepsilon_y]^2 && (40)\\
&= E[\varepsilon_y^2] + \alpha^2 EX^2 + \beta^2 EW^2 + 2E\varepsilon_w E X + E\varepsilon_w E W && (41)\\
&\quad + 2E X E W && (42)\\
&= \sigma_y^2 + \frac{\beta^2}{2} + \alpha^2 E[X + \varepsilon_w]^2 + 2\frac{\beta}{2} E[X + \varepsilon_w] && (43)\\
&= \sigma_y^2 + \frac{\beta^2}{2} + \alpha^2 \sigma_w^2 + \frac{\alpha^2 \beta^2}{2} + \frac{\beta^2}{2} && (44)\\
&= \sigma_y^2 + \frac{\beta^2 + \alpha^2 \beta^2 + \beta^2}{2} + \alpha^2 \sigma_w^2. && (45)
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
E[Y - \hat{f}_{\text{DE}}]^2 &= E[X + W + \varepsilon_y - W]^2 && (46)\\
&= \sigma_y^2 + \alpha^2 EX^2 && (47)\\
&= \sigma_y^2 + \frac{\beta^2}{2} && (48)\\
E[Y - \hat{f}_{\text{IE}}]^2 &= E[X + W + \varepsilon_y - X]^2 && (49)\\
&= \sigma_y^2 + \beta^2 EW^2 && (50)\\
&= \sigma_y^2 + \frac{\alpha^2 \beta^2}{2} + \beta^2 \sigma_w^2 && (51)\\
E[Y - \hat{f}_{\varnothing}]^2 &= E[X + W + \varepsilon_y - X - W]^2 && (52)\\
&= E\varepsilon_y^2 = \sigma_y^2. && (53)
\end{aligned}$$

Therefore, we have that

$$E[Y - \hat{\beta}_{f\{D,E\},\{E\}}]^2 - E[Y - \hat{\beta}_{;}]^2 = \frac{\sigma^2_{\epsilon} + \sigma^2_{\epsilon}\sigma^2_{\epsilon} + \sigma^2_{\epsilon}}{2} + \sigma^2_{\epsilon}\sigma^2_{w} \tag{54}$$

$$= \frac{\sigma^2_{\epsilon}}{2} + \frac{\sigma^2_{\epsilon}\sigma^2_{\epsilon} + \sigma^2_{\epsilon}}{2} + \sigma^2_{\epsilon}\sigma^2_{w} \tag{55}$$

$$= E[Y - \hat{\beta}_{DE}]^2 - E[Y - \hat{\beta}_{;}]^2 \tag{56}$$

$$+ E[Y - \hat{\beta}_{f\{D,E\},\{E\}}]^2 - E[Y - \hat{\beta}_{DE}]^2. \tag{57}$$

B Theorem Proofs

Thm. 1 Proof: The total excess loss (TEL) is defined as:

$$PSEL(; \to f\{D;I;S\}) = E[L(\hat{\beta}_{f\{D;I;S\}}; Y)] - E[L(\hat{\beta}_{;}; Y)]. \tag{58}$$

Let $S_1; S_2 \in f\{D;I;S\}$ such that $S_1 \neq S_2$. The quantity can be expanded as follows:

$$PSEL(; \to f\{D;I;S\}) = E[L(\hat{\beta}_{f\{D;I;S\}}; Y)] - E[L(\hat{\beta}_{;}; Y)] \tag{59}$$

$$= E[L(\hat{\beta}_{f\{D;I;S\}}; Y)] - E[L(\hat{\beta}_{S_1}; Y)] \tag{60}$$

$$+ E[L(\hat{\beta}_{S_1}; Y)] - E[L(\hat{\beta}_{;}; Y)] \tag{61}$$

$$= \underbrace{E[L(\hat{\beta}_{f\{D;I;S\}}; Y)] - E[L(\hat{\beta}_{f\{S_1;S_2\}}; Y)]}_{PSEL(f\{S_1;S_2\} \to f\{D;I;S\})} \tag{62}$$

$$+ \underbrace{E[L(\hat{\beta}_{f\{S_1;S_2\}}; Y)] - E[L(\hat{\beta}_{S_1}; Y)]}_{PSEL(S_1 \to f\{S_1;S_2\})} \tag{63}$$

$$+ \underbrace{E[L(\hat{\beta}_{S_1}; Y)] - E[L(\hat{\beta}_{;}; Y)]}_{PSEL(; \to S_1)} \tag{64}$$

$$= PSEL(; \to S_1) + PSEL(S_1 \to f\{S_1;S_2\}) \tag{65}$$

$$+ PSEL(f\{S_1;S_2\} \to f\{D;I;S\}); \tag{66}$$

completing the theorem's proof for the choice $S_1 = \{D;S_2 = I\}$. The proof also implies other decompositions of the total excess loss $PSEL(; \to f\{D;I;S\})$ as mentioned in Rem. 2. □

Prop. 3 Proof: Let $S = \{S_1;\ldots;S_m\}$ be a set of causal pathways between the attribute X and the outcome Y (such as direct, indirect, and spurious effects in the context of the Standard Fairness Model). Consider the value function

$$f(S) = PSEL(; \to S). \tag{67}$$

The Shapley value of pathway S_i associated with f is given by:

$$(S_i) = \sum_{S \subseteq f\{S_i\}} \frac{1}{m \binom{m-1}{|S|}} (f(S \cup \{S_i\}) - f(S)) : \tag{68}$$

Consider now the generalized G_{PSEL} graph form pathways shown in Fig. 6. The graph is split into columns, and in column ℓ all of the sets S of size ℓ appear. The quantity $APSEL(S_i)$ considers all pathways that start in $;$ and end in $f\{S_1;\ldots;S_m\}$, and takes the average of excess losses associated with S_i along each path, written $PSEL(\ldots \to S_i \to \ldots)$. The APSEL quantity is written as:

$$\frac{1}{n!} \sum_{\substack{\text{paths } \in G_{PSEL}: \\ ; \text{ to } f\{D;I;S\}}} PSEL(\ldots \to S_i \to \ldots): \tag{69}$$

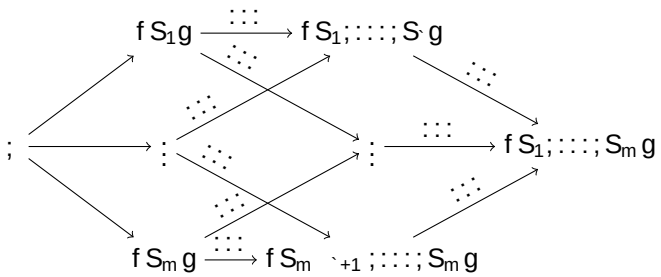


Figure 6: Generalized G_{PSEL} graph form causal pathways.

We now split the expression for the Shapley value by conditioning on the size of the set S

$$\phi(S_i) = \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \frac{1}{m\binom{m-1}{\ell}} \left(f(S \cup \{S_i\}) - f(S)\right) \tag{70}$$

$$= \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \frac{1}{m\binom{m-1}{\ell}} \left(\text{PSEL}(\ldots) - \text{PSEL}(\ldots)\right) \tag{71}$$

$$= \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \frac{1}{m\binom{m-1}{\ell}} \text{PSEL}(\ldots). \tag{72}$$

Now, for a set S of size $|S| = \ell$. There are exactly $\ell!$ paths leading from the root to S. This is followed by a fixed transition $S \to S \cup \{S_i\}$ is chosen. The remaining steps from $S \cup \{S_i\}$ to $\{S_1,\ldots,S_m\}$ along the path can be chosen in $(m-\ell-1)!$ ways. Therefore, across all paths, the contribution $\text{PSEL}(S \to S \cup \{S_i\})$ appears $\ell!(m-\ell-1)!$ times. Therefore, the APSEL quantity can be re-written as:

$$\text{APSEL}(S_i) = \frac{1}{m!} \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \ell!(m-\ell-1)!\,\text{PSEL}(S \to S \cup \{S_i\}) \tag{73}$$

$$= \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \frac{\ell!(m-\ell-1)!}{m!} \text{PSEL}(S \to S \cup \{S_i\}) \tag{74}$$

$$= \sum_{\ell=0}^{m-1} \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}:|S|=\ell} \frac{1}{m\binom{m-1}{\ell}} \text{PSEL}(S \to S \cup \{S_i\}) \tag{75}$$

$$= \sum_{S \subseteq \{S_1,\ldots,S_m\}\setminus\{S_i\}} \frac{1}{m\binom{m-1}{|S|}} \text{PSEL}(S \to S \cup \{S_i\}) \tag{76}$$

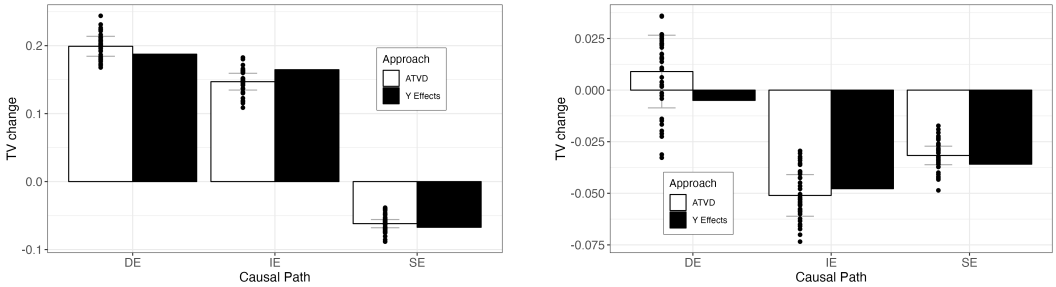
$$= \psi(S_i). \tag{77}$$

The proof for the ATVD quantity follows the same steps. □

C Connection to Causal Decompositions

In this appendix, we explore the connection of the fairness/accuracy trade-off described in the main text with the literature on causal decompositions [45, 31]. As described in the main text and Alg. 1, along with tracking the reduction in predictive power from imposing a causal fairness constraint, we were also interested in the change in the $\text{TV}_{x_0,x_1}(y|S)$, captured by the TV difference (TVD) quantity defined as:

$$\text{TVD}(S \to S') = \underbrace{\left[\mathbb{E}[Y^{S'} \mid x_1] - \mathbb{E}[Y^{S'} \mid x_0]\right]}_{\text{TV after removing } S'\setminus S} - \underbrace{\left[\mathbb{E}[Y^{S} \mid x_1] - \mathbb{E}[Y^{S} \mid x_0]\right]}_{\text{TV before removing } S'\setminus S}. \tag{78}$$



(a) TVD and TVR comparison on Census 2018 data. (b) TVD and TVR comparison on COMPAS data.

Figure 7: Comparison of alternative causal disparity quantifications TVD, TVR.

In words, the quantity TVD($S \; ! \; S^0$) measures how much the TV measure changes once the effect in $S^0 \; n \; S$ is constrained to 0. Put differently, the TVD($S \; ! \; S^0$) is a quantification of the amount of disparity that is transmitted along the path $S^0 \; n \; S$. Other notions for quantifying the size of the disparity transmitted along a specific causal pathway have also been explored in the literature on causal decompositions [45, 31]. An important result from [31] we draw a connection to is the decomposition of the TV measure into the NDE, NIE, and NSE contributions:

Proposition 4 (TV Decomposition [31])**.** *The TV measure can be decomposed into the natural direct, indirect, and spurious contributions:*

$$TV_{x_0;x_1}(y) = NDE_{x_0;x_1}(y) \quad NIE_{x_1;x_0}(y) + NSE_{x_1}(y) \quad NSE_{x_0}(y): \tag{79}$$

Based on the decomposition in Eq. 79, we can try to quantify how much the disparity would be reduced by if we were to remove an effect. For instance, if the direct effect was removed, the reduction in the disparity would be

$$\text{TV}_{x_0;x_1}(y^{\text{DE}}) \quad \text{TV}_{x_0;x_1}(y) = \quad \text{NDE}_{x_0;x_1}(y) \tag{80}$$

where y^{DE} is the outcome with the direct effect removed, which we assume satisfies:

$$\text{NDE}_{x_0;x_1}(y^{\text{DE}}) = 0 \tag{81}$$
$$\text{NIE}_{x_1;x_0}(y^{\text{DE}}) = \text{NIE}_{x_1;x_0}(y) \tag{82}$$
$$\text{NSE}_{x_0}(y^{\text{DE}}) = \text{NSE}_{x_0}(y) \tag{83}$$
$$\text{NSE}_{x_1}(y^{\text{DE}}) = \text{NSE}_{x_1}(y): \tag{84}$$
$$\tag{85}$$

A similar reasoning could be applied to understand the reduction in the TV resulting from removing indirect, and spurious effects. Therefore, in this way we can obtain a different quantification of the TVD measure, which we call TV reduction (TVR), satisfying:

$$\text{TVR}(S \; ! \; S \; [\; \text{DE}) = \quad \text{NDE}_{x_0;x_1}(y) \tag{86}$$
$$\text{TVR}(S \; ! \; S \; [\; \text{IE}) = \text{NIE}_{x_1;x_0}(y) \tag{87}$$
$$\text{TVR}(S \; ! \; S \; [\; \text{SE}) = \quad \text{NSE}_{x_1}(y) + \text{NSE}_{x_0}(y); \tag{88}$$

where in each line we assume the set S does not contain DE, IE, and SE, respectively. To establish a connection between our approach with the literature on causal decompositions, we empirically compare the TVR and TVD measures. We again use the same datasets as in the main text, namely the Census 2018 dataset from Ex. 2 and the COMPAS dataset from Ex. 3.

The results of the empirical comparisons of TVR and TVD values are shown in Fig. 7, where the indicate bars are standard deviations obtained from 10 bootstrap samples of the data. The figure illustrates that on both COMPAS and Census data, TVD and TVR values are not statistically different, across direct, indirect, and spurious effects. This finding adds further validity to our analysis in the main text, meaning that the TVD measure corresponds closely with existing notions for quantifying discrimination in the literature.

D CFUR Quantity Representation

The CFUR quantity (Def. 4) of a causal path S_i is given by

$$\frac{\text{ATVD}(S_i)}{\text{APSEL}(S_i)}. \tag{89}$$

The ATVD and APSEL quantities are defined as the average values of TVD, PSEL quantities, defined through Eqs. 24-25. In this appendix, we discuss the connection of the CFUR value with the local fairness-utility ratios, defined via:

$$\text{LCFUR}(S; S_i) = \frac{\text{TVD}(S \;!\; S \,[\, S_i)}{\text{PSEL}(S \;!\; S \,[\, S_i)}. \tag{90}$$

From Eqs. 24-25, we see that

$$\text{CFUR}(S_i) = \frac{\frac{1}{m!} \sum_{S \in I} \text{TVD}(S \;!\; S \,[\, S_i)}{\frac{1}{m!} \sum_{S \in I} \text{PSEL}(S \;!\; S \,[\, S_i)}, \tag{91}$$

where m is the number of causal paths and in the summations the set S ranges within I, the set of all path prefixes of S_i along all paths connecting $\varnothing$ and $\{S_1, \ldots, S_m\}$ in the G_{PSEL} graph. Now, we want to connect the global CFUR with the local LCFUR values. This can be done through the following proposition:

Proposition 5 (LCFUR and CFUR)**.** *The $CFUR(S_i)$ value of a causal path S_i can be written as:*

$$CFUR(S_i) = \sum_{S \in I} \alpha(S_i) \quad LCFUR(S; S_i); \tag{92}$$

where the weights $\alpha(S; S_i)$ are defined as:

$$\alpha(S; S_i) = \frac{PSEL(S \;!\; S \,[\, S_i)}{\sum_{S' \in I} PSEL(S' \;!\; S' \,[\, S_i)}. \tag{93}$$

The above proposition shows that the global CFUR value is a weighted average of all the local LCFUR values. The weight assigned to each local LCFUR is proportional to the size of its path-specific excess loss $\text{PSEL}(S \;!\; S \,[\, S_i)$ among the sum of all path-specific excess losses that are considered $\sum_{S' \in I} \text{PSEL}(S' \;!\; S' \,[\, S_i)$. Therefore, we see that the CFUR value places larger weights on $S \;!\; S \,[\, S_i$ transitions that result in a larger path-specific excess loss.

E UCI Credit Example

Example 4 (UCI Credit [43])**.** *A commercial bank is developing an automated payment default scoring system based on the UCI Credit Dataset, collected in 2018. This dataset includes demographic variables Z (Z_1 for age), the protected attribute gender X (x_0 female, x_1 male), marital status M, education level L, employment status E, and financial history F. The institution aims to predict the binary outcome Y that indicates whether the individual defaulted on a credit payment ($Y = 1$ for defaulting, $Y = 0$ otherwise). The grouping $\{X = X; Z = \{Z_1\}; W = \{L; M; E; F\}; Y = Y\}$ induces the standard fairness model [31] for this analysis.*

The development team is focused on ensuring fairness in the risk of default predictions. They examine the causal effects from the protected attribute X to the predicted default probability $\hat{p}_S$, evaluating how much scores need to be adjusted to mitigate specific causal pathways, using the complement of AUROC ($1 - AUROC$) as the loss metric. They employ Alg.1 for this purpose and develop fair scoring models $\hat{p}_S$ using different S-sets as per Alg.2.

The analysis results are depicted in Fig.8, with uncertainty bars indicating standard deviations over 10 bootstrap repetitions. Fig.8a) shows the TVD and PSEL values. The team notes that for direct and spurious effects the APSEL and ATVD values are not significantly different from 0. In words, they find that imposing a fairness constraint along these causal pathways does not significantly reduce the predictive power (in terms of reducing AUROC) of the predictor $\hat{p}_S$. For the indirect effect, they find