Fairness-Accuracy Trade-Offs: A Causal Perspective

Drago Plecko and **Elias Bareinboim**
Department of Computer Science
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 conceptions 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 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 tension, 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. We demonstrate our approach on multiple real-world datasets, and provide a new insight 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. These technologies significantly influence the lives of individuals and are frequently used in high-stakes settings [15, 19, 6]. 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. Past data, often 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 [2], facial recognition [7], online ads [33, 12], and system authentication [31], 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 [4, 5] and racial disparities in legal outcomes [34, 23]. 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 independence (demographic parity [11]), separation (equalized odds [14]), and sufficiency (calibration [9]). These definitions, however, have been shown as mutually incompatible [3, 17]. 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 [18, 16, 21, 37, 36, 35, 8, 29, 26]. 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. Interestingly, such considerations are also the basis of some of the legal frameworks for assessing discrimination. For instance, in the context of employment law, the disparate impact doctrine within the Title VII of the Civil Rights Act of 1964 [1] 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 business purpose" Recover citation. 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 [10]. Even though a canonical argument could be used to argue that an unconstrained predictor always achieves a greater or equal utility than a constrained one, the literature seems to be divided on this issue. For instance, some works argue that fairness and utility trade-offs are negligible in practice [30], while others argue that such trade-offs need not even exist [20, 13]. The key subtlety here is that different works discuss different fairness metrics, and imposing different constraints. Naturally, the implications on the predictor's utility will strongly depend on the exact type of the fairness constraint that is enforced. [1]

Interestingly, the tension between fairness and utility has been largely unexplored in the causal fairness literature (with some exceptions such as [22, 27, 26]). 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:*

$$X \leftarrow \textit{Bernoulli}(0.5) \tag{1}$$
$$W \leftarrow \beta X + \epsilon_w \tag{2}$$
$$Y \leftarrow \alpha X + \gamma W + \epsilon_y, \tag{3}$$

where $\epsilon_w \sim N(0, \sigma_w^2), \epsilon_y \sim N(0, \sigma_y^2)$. Variable X is the protected attribute, and Y is the outcome of interest. A graphical representation of Eqs. 1 -3 is shown in Fig. 1. Attribute X can influence Y along two different pathways: the direct path $X \to Y$, and the indirect path $X \to W \to Y$. We

[1] For instance, the works that discuss how fairness-accuracy trade-offs either do not exist, or are insignificant in practice, often focus on the equality of odds metric [14] (given by the independence $\widehat{Y} \perp\!\!\!\perp X \mid Y$). Notably, this metric always allows for the perfect predictor $\widehat{Y} = Y$, and in settings with good predictive power, the cost of enforcing this constraint may indeed be negligible.

therefore consider fair predictors $\widehat{Y}^S$ of the form:

$$\widehat{Y}^S = \hat{\alpha}_S X + \hat{\gamma}_S W, \tag{4}$$

where the predictor $\widehat{Y}^S$ removes effects in the set S, with S ranging in $\{\emptyset, DE, IE, \{DE, IE\}\}$. For instance, the optimal predictor $\widehat{Y}^\emptyset$ has $\hat{\alpha}_\emptyset = \alpha, \hat{\gamma}_\emptyset = \gamma$, and therefore its mean-squared error (MSE) equals:

$$\mathbb{E}[Y - \widehat{Y}^\emptyset]^2 = \sigma_y^2 \tag{5}$$

The fully fair predictor $\widehat{Y}^{\{DE,\ IE\}}$ has $\hat{\alpha}_{\{DE,\ IE\}} = 0, \hat{\gamma}_{\{DE,\ IE\}} = 0$ and thus has the MSE

$$\mathbb{E}[Y - \widehat{Y}^{\{DE,\ IE\}}]^2 = \frac{(\alpha + \beta\gamma)^2}{2} + \gamma^2\sigma_w^2 + \sigma_y^2. \tag{6}$$

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

$$\underbrace{\mathbb{E}[Y - \widehat{Y}^{\{DE,\ IE\}}]^2}_{\text{fully-fair predictor's loss}} - \underbrace{\mathbb{E}[Y - \widehat{Y}^\emptyset]^2}_{\text{unconstrained loss}} = \underbrace{\mathbb{E}[Y - \widehat{Y}^{\{DE,\ IE\}}]^2 - \mathbb{E}[Y - \widehat{Y}^{DE}]^2}_{\text{excess IE loss}} \tag{7}$$

$$+ \underbrace{\mathbb{E}[Y - \widehat{Y}^{DE}]^2 - \mathbb{E}[Y - \widehat{Y}^\emptyset]^2}_{\text{excess DE loss}}. \tag{8}$$

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

$$\underbrace{\mathbb{E}[\widehat{Y}^{S\cup S'} \mid x_1] - \mathbb{E}[\widehat{Y}^{S\cup S'} \mid x_0]}_{\text{disparity after removing } S'} + \underbrace{\mathbb{E}[Y^S \mid x_0] - \mathbb{E}[Y^S \mid x_0]}_{\text{disparity before removing } S'}. \tag{9}$$

□

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 $\mathbb{E}[\widehat{Y} \mid x_1] - \mathbb{E}[\widehat{Y} \mid x_0]$ (also known as the *parity gap*), relating to previous results on causal decompositions found in the literature [37, 26]. With the above example in mind, we can list the key contributions of this manuscript:

(i) We define the notion of a path-specific excess loss associated with imposing a fairness constraint along a causal path (Def. 4), and we prove how the total excess loss 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 excess loss is affected by different fairness constraints. In this context, we show the equivalence of Alg. 1 with a Shapley value [32] approach (Prop 3),

(iii) For purposes of applying Alg. 1, a key requirement is the construction of causally-fair predictors $\widehat{Y}^S$ which remove effects along pathways in S. We introduce a novel Lagrangian formulation of the optimization problem for such $\widehat{Y}^S$ (Def. 6), and a training procedure for learning the predictor (Alg. 2),

(iv) We introduce the causal fairness/utility ratio (CFUR, Def. 5) that summarizes how much the group disparity can be reduced per 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) as our basic semantical framework [24]. 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

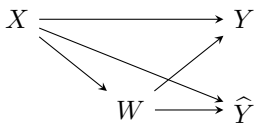


Figure 1: Graphical representation of Ex. 1.

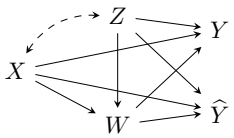


Figure 2: Standard Fairness Model.

each $V_i \in V$, where $V_i \leftarrow f_{V_i}(\text{pa}(V_i), U_{V_i})$ for some $\text{pa}(V_i) \subseteq V$ 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 $\mathcal{G}$ [24] 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 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 [26])**.** *Given an SCM $\mathcal{M}$, a contrast $\mathcal{C}$ is any quantity of the form*

$$\mathcal{C}(C_0, C_1, E_0, E_1) = \mathbb{E}[y_{C_1} \mid E_1] - \mathbb{E}[y_{C_0} \mid E_0], \tag{10}$$

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) [26] over endogenous variables $\{X, Z, W, Y, \widehat{Y}\}$ 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 $\widehat{Y}$ (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[2]. Based on the SFM, we will use the following causal fairness measures for different causal pathways:

Definition 2 (Population-level Causal Fairness Measures [25, 26])**.** *The natural direct, indirect, and spurious effects are defined as:*

$$\textit{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0}) \tag{11}$$

$$\textit{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1}). \tag{12}$$

$$\textit{NSE}_{x}(y) = P(y \mid x) - P(y_x). \tag{13}$$

Importantly, [26] also shows that the TV measure can be decomposed into the NDE, NIE, and NSE contributions:

$$\text{TV}_{x_0,x_1} = \text{NDE}_{x_0,x_1}(y) - \text{NIE}_{x_1,x_0}(y) + \text{NSE}_{x_1}(y) - \text{NSE}_{x_0}(y). \tag{14}$$

Based on the above causal measures of fairness, we can define the causally-fair predictor with respect to set of pathways S:

[2]Partial identification techniques for bounding effects can be used for relaxing these assumptions [38].

Definition 3 (Causally Fair Predictor [28]). *The causally S-fair predictor $\widehat{Y}^S$ with respect to a loss function L and pathways in S is the solution to the following optimization problem:*

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

$$\text{subject to} \quad \textit{NDE}_{x_0,x_1}(f) = \textit{NDE}_{x_0,x_1}(y) \cdot \mathbb{1}(\textit{DE} \notin S) \tag{16}$$

$$\textit{NIE}_{x_1,x_0}(f) = \textit{NIE}_{x_1,x_0}(y) \cdot \mathbb{1}(\textit{IE} \notin S) \tag{17}$$

$$\textit{NSE}_{x_0}(f) = \textit{NSE}_{x_0}(y) \cdot \mathbb{1}(\textit{SE} \notin S) \tag{18}$$

$$\textit{NSE}_{x_1}(f) = \textit{NSE}_{x_1}(y) \cdot \mathbb{1}(\textit{SE} \notin S). \tag{19}$$

The definition of $\widehat{Y}^S$ has a straightforward intuition. For any pathway in the set S, the corresponding causal effect should equal to 0, as proposed in the path-specific causal fairness literature [21, 8]. However, importantly, pathways that are not in S also need to be constrained – the effect along these paths should not change compared to the true outcome Y [28], e.g., if the direct path is not in S, then we expect to have $\text{NDE}_{x_0,x_1}(\widehat{y}) = \text{NDE}_{x_0,x_1}(y)$ (and similarly for other effects).

2 Path-Specific Excess Loss

In this section, we introduce the concept of a path-specific excess loss, and then demonstrate how the total excess loss can be decomposed into path-specific excess losses. We start with the following definition:

Definition 4 (Path-Specific Excess Loss). *Let $L(\widehat{Y}, Y)$ be a loss function. Let $\widehat{Y}^S$ be the optimal S-fair predictor with respect to L. Define the path-specific excess loss with respect to a pair S, S' as:*

$$\textit{PSEL}(S \to S') = \mathbb{E}[L(\widehat{Y}^{S'}, Y)] - \mathbb{E}[L(\widehat{Y}^{S}, Y)]. \tag{20}$$

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

The total excess loss computes the increase in the loss for the totally constrained predictor $\widehat{Y}^{\{D,I,S\}}$ with direct, indirect, and spurious effects removed compared to the unconstrained predictor $\widehat{Y}^{\emptyset}$. As it turns out, the total excess loss can be decomposed as a sum of path-specific excess losses:

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

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

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

The proof of the theorem is given in Appendix B. Importantly, we can make the following remark:

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

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

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-specific excess loses. In the left side, we start with $S = \emptyset$ and the predictor $\widehat{Y}^{\emptyset}$. Then, we can add any of $\{D, I, S\}$ to the S-set, to obtain the predictors $\widehat{Y}^{D}, \widehat{Y}^{I}$, or $\widehat{Y}^{S}$, and so on. 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 $\mathcal{G}_{\text{PSEL}}$ to compute path-specific 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 $\mathcal{G}_{\text{PSEL}}$ we can compute $\text{PSEL}(S \to S')$, and we also compute the difference in the TV measure from the transition $S \to S'$, defined by

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

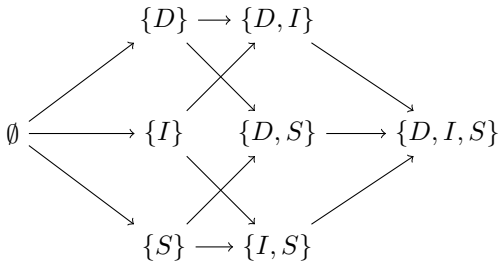


Figure 3: Graphical representation $\mathcal{G}_{\text{PSEL}}$ of possible transitions for causally-fair predictors.

Algorithm 1: Path-Specific Excess Loss Attributions

Input: data $\mathcal{D}$, predictors $\widehat{Y}^S$ for S-sets $\subseteq \{D, I, S\}$

foreach *edge* $(S, S') \in \mathcal{G}_{PSEL}$ **do**

  compute the path-specific excess loss of $S' \setminus S$, given by $\text{PSEL}(S \to S')$

  compute the TV measure difference of $S' \setminus S$, written $\text{TVD}(S \to S')$ given by $\text{TV}_{x_0,x_1}(\widehat{Y}^{S'}) - \text{TV}_{x_0,x_1}(\widehat{Y}^{S})$

foreach *causal path* $S_i \in \{D, I, S\}$ **do**

  compute the average path-specific excess loss and TV difference across all $\emptyset$

$$\text{APSEL}(S_i) = \frac{1}{n(S_i)} \sum_{\substack{\text{edges } (S,S') \in \mathcal{G}_{\text{PSEL}}: \\ S' \setminus S = S_i}} \text{PSEL}(S \to S') \tag{24}$$

$$\text{ATVD}(S_i) = \frac{1}{n(S_i)} \sum_{\substack{\text{edges } (S,S') \in \mathcal{G}_{\text{PSEL}}: \\ S' \setminus S = S_i}} \text{TVD}(S \to S'), \tag{25}$$

  where $n(S_i)$ is the number of edges associated with the causal path S_i.

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

The quantities $\text{PSEL}(S \to S')$ and $\text{TVD}(S \to S')$ are naturally associated with the effect that was removed, i.e., $S' \setminus S$. As there are multiple ways of reaching the set $\{D, I, S\}$ from $\emptyset$ in $\mathcal{G}_{\text{PSEL}}$, each of the causal effects (direct, indirect, spurious) will be associated with a number of different PSELs and TVDs. In Eqs. 24-25, 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 [32] based approach:

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

$$f_1(S) = \textit{PSEL}(\emptyset \to S). \tag{27}$$

$$f_2(S) = \textit{TVD}(\emptyset \to S). \tag{28}$$

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

$$\phi^k(S_i) = \sum_{S \subseteq \{D,I,S\} \setminus \{S_i\}} \frac{1}{n\binom{n}{|S|}} \left(f_k(S \cup \{S_i\}) - f(S) \right). \tag{29}$$

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:

$$\phi^1(S_i) = \textit{APSEL}(S_i) \tag{30}$$

$$\phi^2(S_i) = \textit{ATVD}(S_i). \tag{31}$$

The above proposition illustrates how averaging the influence of removing a causal effect over all possible ways of reaching $\{D, I, S\}$ from $\emptyset$ is equivalent to computing the Shapley values with

respect to an appropriate value function f. We now introduce a notion of a causal fairness/utility ratio for each pathway.

Definition 5 (Causal Fairness/Utility Ratio (CFUR))**.** *The causal fairness/utility ratio (CFUR) for a causal path S_i is defined as*

$$\mathit{CFUR}(S_i) = \frac{\mathit{ATVD}(S_i)}{\mathit{APSEL}(S_i)}. \tag{32}$$

The CFUR quantity may be particularly useful for comparing different causal effects. The intuition behind the quantity is simple – for removing a causal effect S_i from our predictor $\widehat{Y}$, 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 succintly summarizes how much fairness can be gained vs. how much the predictive power is reduced. In Sec. 4, we compute the CFUR values on several real-world datasets.

3 Causally 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 $\widehat{Y}^S$ following Def. 3. However, computing the predictors $\widehat{Y}^S$ in practice is a challenge, and 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 6 (Lagrange Form of $\widehat{Y}^S$)**.** *The causally S-fair λ-optimal predictor $\widehat{Y}^S(\lambda)$ with respect to pathways in S and the loss function L is the solution to the following optimization problem:*

$$\widehat{Y}^S(\lambda) = \arg\min_f \mathbb{E}\, L(Y, f(X, Z, W)) + \tag{33}$$

$$\lambda\big(\mathit{NDE}_{x_0,x_1}(f) - \mathit{NDE}_{x_0,x_1}(y) \cdot \mathbb{1}(\mathit{DE} \notin S)\big)^2 + \tag{34}$$

$$\lambda\big(\mathit{NIE}_{x_1,x_0}(f) - \mathit{NIE}_{x_1,x_0}(y) \cdot \mathbb{1}(\mathit{IE} \notin S)\big)^2 + \tag{35}$$

$$\lambda\big(\mathit{NSE}_{x_0}(f) - \mathit{NSE}_{x_0}(y) \cdot \mathbb{1}(\mathit{SE} \notin S)\big)^2 + \tag{36}$$

$$\lambda\big(\mathit{NSE}_{x_1}(f) - \mathit{NSE}_{x_1}(y) \cdot \mathbb{1}(\mathit{SE} \notin S)\big)^2 \tag{37}$$

The above definition reformulates the problem of finding $\widehat{Y}^S$ to a Lagrangian form. This makes the problem amenable to standard gradient descent methods, and in Alg. 2 we propose a procedure for finding a suitable predictor $\widehat{Y}^S$. The procedure can be described as follows. We are performing a binary search with the intent of finding an appropriate value of the λ parameter. For an interval $[\lambda_{\text{low}}, \lambda_{\text{high}}]$ we take the midpoint λ_{mid}. For this parameter value we compute the optimal predictor $\widehat{Y}^S(\lambda_{\text{mid}})$ for the optimization problem in Eqs. 33-37 by fitting a feed-forward neural network with two hidden layers. After this, the causal measures of fairness are computed for the predictor $\widehat{Y}^S(\lambda_{\text{mid}})$ on an evaluation set, and these measures are compared to the causal measures for the true outcome Y (we perform the appropriate hypothesis tests as in Eqs. 38-41). If none of the hypotheses are rejected, it means that λ_{mid} is large enough to enforce the causal fairness constraints, and we move to the interval $[\lambda_{\text{low}}, \lambda_{\text{mid}}]$. If a hypothesis is rejected, it indicates that λ_{mid} is not large enough, and we move to the interval $[\lambda_{\text{mid}}, \lambda_{\text{high}}]$. In this way, the algorithm allows us to choose the tuning parameter systematically.

4 Experiments

In this section, we perform the causal fairness-accuracy analysis described in Sec. 2. We begin with an example of allocating salaries performed by the US government:

Example 2 (Salary Increase of Government Employees)**.** *An office within 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. The data includes demographic*

Algorithm 2: Causally Constrained Fair-Learning

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

Set $\lambda_{\text{low}} = 0$, $\lambda_{\text{high}} = \text{large}$

while $|\lambda_{high} - \lambda_{low}| > \epsilon$ **do**

  set $\lambda_{\text{mid}} = \frac{1}{2}(\lambda_{\text{low}} + \lambda_{\text{high}})$

  fit a neural network to solves the optimization problem in Eqs. 33-37 with $\lambda = \lambda_{\text{mid}}$ on $\mathcal{D}_t$ to obtain the predictor $\widehat{Y}^S(\lambda_{\text{mid}})$

  compute the causal measures of fairness NDE, NIE, NSE of $\widehat{Y}^S(\lambda_{\text{mid}})$ on evaluation data $\mathcal{D}_e$

  test the hypothesis

$$H_0^{DE} : \text{NDE}_{x_0,x_1}(\widehat{y}^S(\lambda_{\text{mid}})) = \text{NDE}_{x_0,x_1}(y) \cdot \mathbb{1}(\text{DE} \notin S) \tag{38}$$

$$H_0^{IE} : \text{NIE}_{x_1,x_0}(\widehat{y}^S(\lambda_{\text{mid}})) = \text{NIE}_{x_1,x_0}(y) \cdot \mathbb{1}(\text{IE} \notin S) \tag{39}$$

$$H_0^{SE_0} : \text{NSE}_{x_0}(\widehat{y}^S(\lambda_{\text{mid}})) = \text{NSE}_{x_0}(y) \cdot \mathbb{1}(\text{SE} \notin S) \tag{40}$$

$$H_0^{SE_1} : \text{NSE}_{x_1}(\widehat{y}^S(\lambda_{\text{mid}})) = \text{NSE}_{x_1}(y) \cdot \mathbb{1}(\text{SE} \notin S) \tag{41}$$

  if *any of* $H_0^{DE}, H_0^{IE}, H_0^{SE_0}, H_0^{SE_1}$ *rejected* **then**

    $\lambda_{\text{low}} = \lambda_{\text{mid}}$

  else

    $\lambda_{\text{high}} = \lambda_{\text{mid}}$

return *predictor* $\widehat{Y}^S(\lambda_{mid})$

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 government is predicted the outcome Y, the yearly salary of the employees (transformed to a log-scale). The grouping $\{X = X, Z = \{Z_1, Z_2\}, W = \{M, L, R\}, Y = \{Y\}\}$ constructs the standard fairness model for this data.

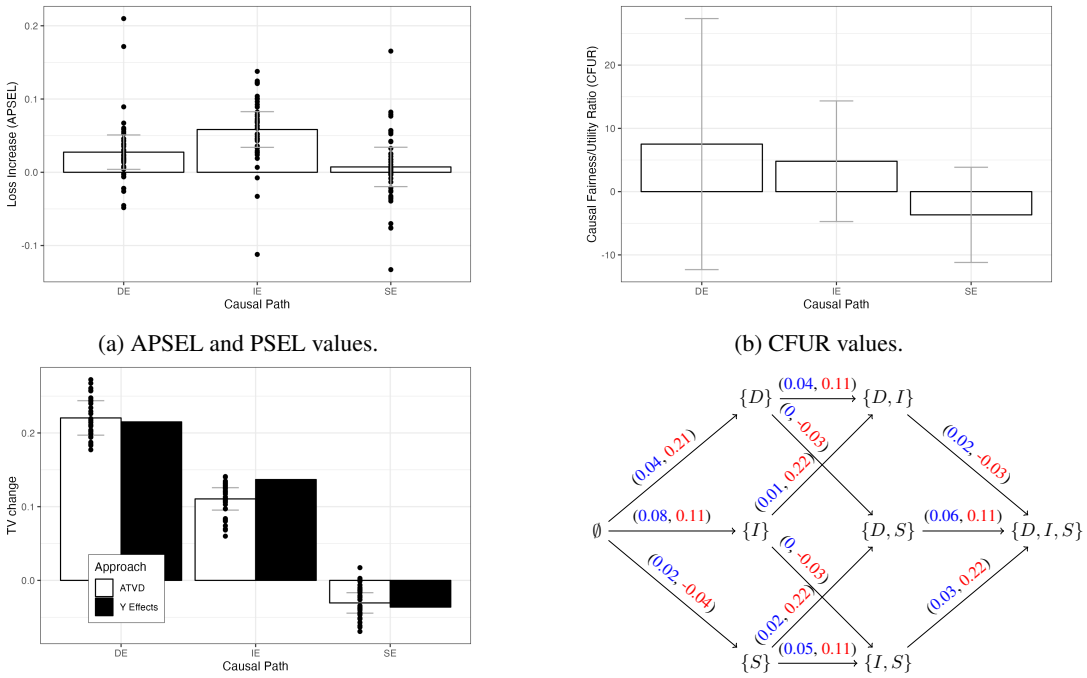
The team developing the ML predictor is also concerned with the fairness of the allocated salaries. In particular, they wish to understand how the different causal effects from the protected attribute X to the predictor $\widehat{Y}$ affect the prediction, and how much the salary predictions would have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

The results of the analysis are shown in Fig. 4. In the analysis of PSEL values (Fig. 4a), the team notices that imposing any fairness constraint incurs an increase in the MSE, with the increase being the largest for the indirect effect. 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. In Fig. 4c, the team visualized the ATVD values, and decided to compare these values to the causal decomposition of the TV measure from [26, Thm. 4.2]. As it turns out, the ATVD values correspond closely with the decomposition terms appearing in the causal decomposition of the outcome Y (as per Eq. 14), adding further validity to the analysis. Finally, the team also visualizes the graph $\mathcal{G}_{PSEL}$ and plots the values of PSEL and TVD for each transition. Based on the analysis, the team decides to use the predictor with the direct effect removed $\widehat{Y}^D$. □

We next look at a well-known example from criminal justice:

Example 3 (Recidivism Prediction on COMPAS)**.** *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.*

In a previous analysis, the team from ProPublica showed that the predictions produced by the private company are discriminatory. They now wish to better understand the tension between fairness and



(a) APSEL and PSEL values.

(b) CFUR values.

(c) ATVD values and Y decomposition.

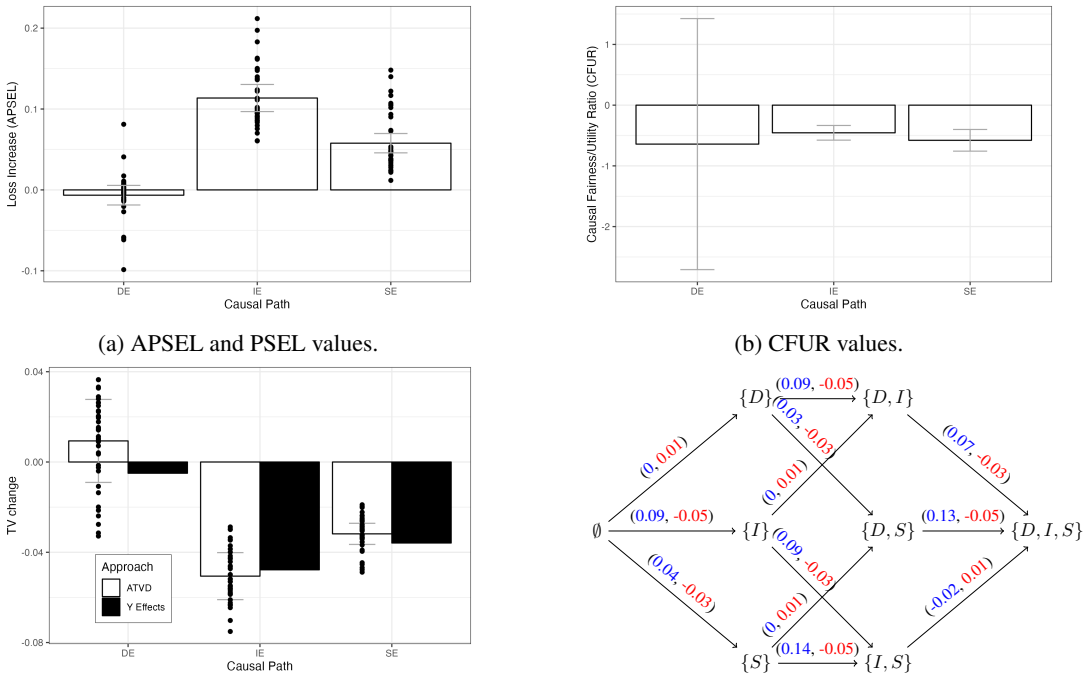
(d) $\mathcal{G}_{\text{PSEL}}$ with PSEL (blue) and TVD values (red).

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

accuracy, since the latter is an important concern of the district court. For the analysis, they use the complement of the area under reciever operator characteristic curve (1−AUC). They apply Alg. 1, and find that the APSEL value is the largest for the indirect effect (around 10% of AUC), while it is smaller for the spurious effect, and negligible for the direct effect (Fig. 5a). The team finds that the ATVD values closely match the causal decomposition of the $TV_{x_0,x_1}(y)$ measure of the true outcome Y into its direct, indirect, and spurious parts (Fig. 5c). They also find that the direct effect does not play a significant role in the prediction. However, they find that the reomoval of the spurious effect results in a 3% decrease in AUC, while it decreases the disparity between Majority and Minority groups by an average of 3% (i.e., the spurious effect is the best in terms of the CFUR metric). Finally, the team also visualizes the graph $\mathcal{G}_{PSEL}$, and the values of PSEL and TVD associated with each transition and effect removal. Based on these findings, in the upcoming court hearing, the team will propose the usage of the causally-constrained predictor $\widehat{Y}^{\{D,S\}}$ that removes direct and spurious effects, and they will use their analysis to demonstrate the impact of this choice on the predictor's accuracy. □

5 Conclusion

The tension between fairness and accuracy is one of the important topics in the literature on fair 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. 4). 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. 3), we introduced a new neural approach



(a) APSEL and PSEL values.

(b) CFUR values.

(c) ATVD values and Y decomposition.

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

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

for constructing such predictors (Def. 6, Alg. 2). Finally, we applied our approach to real-world datasets, and demonstrated conclusively that from causal perspective fairness and accuracy are almost always in tension (Exs. 2-3).

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

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

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

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

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

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

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

[22] 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.

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

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

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

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

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

[28] 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.

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

[30] 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.

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

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.