Overcoming Fairness Trade-offs via Pre-processing: A Causal Perspective

CHARLOTTE LEININGER, LMU Munich, Germany
SIMON RITTEL, LMU Munich, Germany and Munich Center for Machine Learning (MCML), Germany
LUDWIG BOTHMANN, LMU Munich, Germany and Munich Center for Machine Learning (MCML), Germany

Training machine learning models for fair decisions faces two key challenges: The *fairness-accuracy trade-off* results from enforcing fairness which weakens its predictive performance in contrast to an unconstrained model. The incompatibility of different fairness metrics poses another trade-off – also known as the *impossibility theorem*. Recent work identifies the bias within the observed data as a possible root cause and shows that fairness and predictive performance are in accord when predictive performance is measured on unbiased data. We offer a causal explanation for these findings using the framework of the FiND (fictitious and normatively desired) world, a "fair" world, where protected attributes have no causal effects on the target variable. Our contribution is twofold: First, we unify insights from previously separate lines of research and establish a new theoretical link that demonstrates how both the fairness-accuracy and the trade-off between conflicting fairness metrics are naturally resolved in this FiND world. Second, we propose *appFiND*, a new method for evaluating the quality of the FiND world approximation via pre-processing in real-world scenarios where the true FiND world is not observable. In simulations and empirical studies, we demonstrate that these pre-processing methods are successful in approximating the FiND world and resolving both trade-offs. Our results provide actionable solutions for practitioners to achieve fairness and high predictive performance simultaneously.

Keywords: Fairness-accuracy trade-off, impossibility theorem, causal fairness, bias mitigation, pre-processing

Reference Format:
Charlotte Leininger, Simon Rittel, and Ludwig Bothmann. 2025. Overcoming Fairness Trade-offs via Pre-processing: A Causal Perspective. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 24 pages.

1 Introduction

The use of automated decision-making (ADM) systems has become popular in a variety of fields that were previously solely controlled by humans, including sensitive areas such as loan applications [38], hiring choices [26], and the criminal justice system [2]. Such systems have been shown to suffer from bias with respect to certain protected attributes (PAs) – such as gender or race – which is in conflict with a variety of anti-discrimination laws, such as the US Civil Rights Act of 1964 or the Charter of Fundamental Rights of the European Union. In response,

Authors' Contact Information: Charlotte Leininger, C.Leininger@campus.lmu.de, LMU Munich, Munich, Germany; Simon Rittel, simon.rittel@lmu.de, LMU Munich, Munich, Germany and Munich Center for Machine Learning (MCML), Munich, Germany; Ludwig Bothmann, ludwig.bothmann@lmu.de, LMU Munich, Munich, Germany and Munich Center for Machine Learning (MCML), Munich, Germany.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

this concern has led to a growing body of literature on fairness-aware machine learning (fairML), proposing various methods to measure and achieve fairness for ADM systems based on machine learning (ML) models.

This was followed by the observation that fulfilling a fairness notion comes with a decrease in accuracy – the *fairness-accuracy trade-off* (see, e.g., [16, 18, 32]). Concurrently, with the multitude of suggested fairness metrics came the realization that some of them cannot be met at the same time. There appears to be a trade-off between several fairness metrics, that cannot be satisfied simultaneously unless in some special cases – also referred to as the *impossibility theorem* [15, 36]. Newer studies have questioned these seemingly inevitable trade-offs, particularly in relation to the potential bias present in the data [7, 24, 60]. When fairness and accuracy are measured on such biased data, the results must be assumed to be similarly exposed to bias. This raises the need to eliminate bias from the data and evaluate these metrics on unbiased data instead. In doing so, it suggests that fairness and accuracy can enhance each other rather than conflict. Similarly, by obtaining unbiased data that represents a "fair world", measuring fairness with respect to different fairness metrics should yield consistent rather than conflicting results.

1.1 Our Contributions

In this paper, we adopt a causal perspective to form a new theoretical link between previous findings in fairML to show how both the fairness-accuracy trade-off and the trade-off between conflicting fairness metrics can be overcome. Therefore, we investigate the trade-offs in a causal framework of a "fair world" and utilize the FiND (fictitious and normatively desired) world proposed by Bothmann et al. [9]. They provide a philosophically sound definition of a fair world, where the PAs have no causal effect on the target, neither directly nor indirectly.

First, we examine the theoretical implications of this FiND world and demonstrate the **theoretical resolution of the trade-offs:** We show that and explain why the FiND world ensures fairness at both the individual and the group level, overcoming the trade-off between incompatible fairness metrics. At the same time, this allows us to overcome the fairness-accuracy trade-off by explaining how enforcing a fairness metric on a model trained in the real world leads to improved predictive performance in the FiND world.

Second, we examine the practical implications of this newly formed connection between the trade-offs. Since we do not have access to the FiND world in real life, we need to approximate it. For this, we compare two causal pre-processing approaches: *fairadapt* [51] and *residual-based warping* [8]. We demonstrate the effectiveness of the pre-processing methods in a simulation study by evaluating these approximations against the true FiND world, hence **resolving the trade-offs practically via pre-processing.** Additionally, we propose a **new evaluation method "appFiND"** to assess whether the pre-processing approximates the FiND world in real-world use cases (where we do not know the FiND world). Therefore, we utilize an in-processing method that evaluates a model's performance for increasing strengths of a fairness constraint. We observe that for FiND world data as well as for pre-processed data fairness is satisfied without sacrificing predictive performance, confirming that the pre-processing effectively eliminated bias. Concurrently, we show how all fairness metrics that are subject to the trade-off between competing fairness notions are simultaneously satisfied when using pre-processing methods that approximate the FiND world. We therefore present actionable methods for how both the fairness-accuracy trade-off and the trade-off between various fairness metrics can be overcome in practice.

Lastly, we validate our findings from the simulated setup on a real-world dataset, using data from the Home Mortgage Disclosure Act (HMDA) dataset. Our results confirm that the causal pre-processing methods are equally capable of successfully aligning fairness and performance on real datasets while satisfying multiple fairness metrics.

1.2 Related Work

While the widespread assumption that there is a trade-off between fairness and accuracy has already been explored from various angles [16, 18, 32, 49, 66], a growing number of fairML approaches are challenging this trade-off [5, 21, 44, 52, 55]. In particular, recent works suggest bias as a possible cause of this apparent conflict [7, 24, 27, 28, 31, 46, 54]. For instance, enforcing fairness constraints has been shown to enhance accuracy when evaluated on unbiased data [7, 54]. Besides, Menon and Williamson [46] frame the trade-off in terms of the dependence between the PA and the target label, showing that one can only achieve maximum accuracy and fairness simultaneously in the case of perfect independence. As we determine in the causal framework in Section 2.2.1, the assumption of (conditional) independence between PA and outcome aligns with the assumptions of the FiND world.

Closely related to our approach is the work by Wick et al. [60] who tackle the fairness-accuracy trade-off from a non-causal viewpoint. They conclude that the trade-off itself is a false notion if accuracy is measured on test data that is equally biased as the training data. Therefore, one must assume that the corresponding accuracy measurements are also biased and it is crucial to evaluate accuracy on the unbiased – fair – labels. By simulating fair labels on the test set, they show that fairness and accuracy are positively related. We build on these findings by integrating the generation of fair labels into the causal FiND world framework. This allows us to provide a *theoretical explanation* for the trade-off from a causal perspective and, additionally, to offer a *practical solution* for overcoming the trade-off by pre-processing that approximates the FiND world. Therefore, we are able to address the trade-off in both simulated and real-world settings.

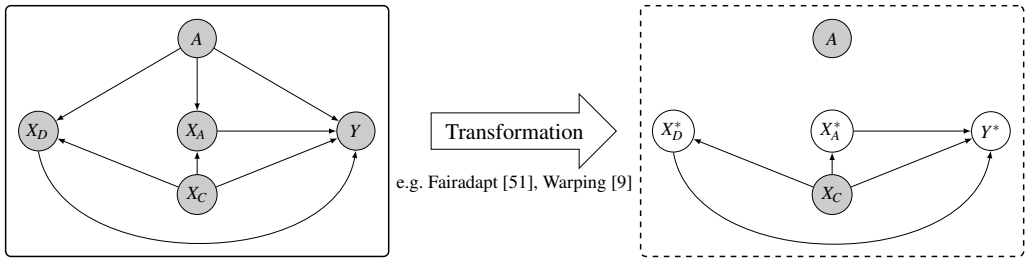
Regarding the impossibility theorem, Bell et al. [4] examine the two exceptions of the impossibility theorem where fairness metrics can be satisfied simultaneously: the case of equal group base rates and the case of perfect prediction. They conceive "fairness regions" in which they quantify the possible options of aligning the fairness metrics subject to the trade-off when one of these special cases is approximately satisfied. We also investigate the special case of equal group base rates and derive that it is, by design, fulfilled in the FiND world.

2 Theoretical Framework

2.1 Conception of FiND World

Fairness in ML has often been approached by employing quantitative group fairness metrics. However, the use of such "classical" metrics has been criticized, as their blind application – without considering the underlying structure of the data or potential biases – can be blatantly unfair [17, 25, 47]. Bothmann et al. [9] therefore derive a philosophically grounded definition of what a normatively fair treatment is and support this by taking into account causal considerations.

In fact, even from a philosophical perspective, there appears to be no single answer to "what fairness is". Instead, they show that the conception of fairness always depends on multilayered normative evaluations that depend on the context or task at hand: Building on Aristotle, a treatment is considered fair if equals are treated equally and unequals unequally. They define the concept of "task-specific equality" and thereby require normative stipulations to define when individuals are considered to be equal in a certain task. When considering causal relationships in the real world, a PA can have a causal effect on the target. This may be due to a variety of reasons, including biases introduced during the data collection process or historical biases, stemming from historic discrimination against the protected group. They argue that taking such causal relationships into account in a prediction process thus can



(a) Factual real world *(b)* Counterfactual FiND world

Fig. 1. Assumed causal DAGs of the *(a)* real world and *(b)* FiND world for the credit application example introduced in Section 2.1.1, with shaded nodes being observed, i.e., accessible for model training. In the FiND world, the PA A has no children, its descendants according to the causal graph from the real world lack a causal bias from A. To distinguish these counterfactual features from their real equivalents, we denote them with an asterisk. For further details, we refer to Section 4.1.

represent a normatively unfair treatment. The reasoning behind this is that societal norms as well as several legal requirements demand not to differentiate between individuals based on their PAs. This implies that individuals are to be considered equal if they only differ in their association with a protected group. In order for a treatment to be normatively fair, one must therefore move the decision-making process to a world without causal effects from the PA on the target.

2.1.1 The FiND World

In the "fictitious, normatively desired" (FiND) world introduced by Bothmann et al. [9], the PAs have no causal effect on the target, neither directly nor indirectly.[1] They distinguish this counterfactual FiND world from the real world, in which direct and indirect causal effects from the PAs on the target may exist.

Assuming acyclic causal relations and causal sufficiency, the model of the FiND world and real world can be represented using a Bayesian network where nodes represent random variables. A directed edge denotes a causal effect, while a directed path with more than two variables implies an indirect one. Throughout the paper we will use the example of a credit application with binary PA A, target Y (risk of credit non-repayment), features X_A (credit amount), X_D (debt) and a confounder X_C (age). The corresponding directed acyclic graphs (DAGs) for the real and FiND world are illustrated in Figure 1. Any PA is assumed to be a root node and is shared between both worlds, i.e., remains unchanged. By contrast, the missing effect of PAs on features yields a counterfactual FiND representation which we denote by $\mathbf{X}^*$ to distinguish them from $\mathbf{X}$ that are observed in the real world.

For any a, a' of the set of possible group memberships $\mathscr{A}$, the setting of the FiND world ensures that there is no difference in distributions of outcomes Y^* regarding the protected group a and unprotected group a'. Further, individuals of a and a' receive equal outcome Y^* if they are equal with respect to their counterfactual features $\mathbf{X}^*$ and factual feature(s) X_C. In this manner, the FiND world represents a conceptual framework that ensures fairness both on the group level and on the individual level (further explanation follows in Section 2.2).

[1] They also provide an adjusted version for the case that certain path-specific effects are normatively deemed admissible, c.f. their Def. 3.7. For the ease of presentation, we work with the stricter definition of no causal effects in the remainder – while extensions are straightforward and only depend on the availability of suitable pre-processing methods.

2.1.2 FiND World vs. "We're All Equal" Worldview

While differing in their philosophical derivation of a fair world, the "We're All Equal" (WAE) worldview by Friedler et al. [29, 30] shares theoretical implications with the FiND world, but from a non-causal approach. The WAE worldview reflects the assumption that in a construct space – an idealized, unobservable space holding unbiased information of individuals – all groups are essentially the same. However, biases arise when mapping from the construct space to the observed space, resulting in discriminatory representations that get transported into the decision space. In order to be fair, the decision-making process must therefore be shifted to the construct space. Therefore, WAE demands the target Y to be independent of the PA A:

$$Y \perp\!\!\!\perp A. \tag{1}$$

The FiND world can be transferred to the WAE worldview to the extent that if there exists no (direct nor indirect) causal effect from A on Y, Y is also statistically independent of A. Therefore, (1) is fulfilled in the FiND world.

2.1.3 FiND World vs. Counterfactual Fairness

Counterfactual fairness [39] is a causal fairness notion that compares factual and counterfactual predictions. In the counterfactual world, an individual with a PA $A = a$ belongs counterfactually to the group $A = a'$. A predictor $\hat{Y}$, defined with latent exogenous variables $\mathbf{U}$ and observable features $\mathbf{X}$, is considered counterfactually fair iff, for all $\mathbf{x} \in \mathbb{R}^p$ and $a, a' \in \mathscr{A}$:

$$P(\hat{Y}_{A \leftarrow a}(\mathbf{U}) = y \mid \mathbf{X} = \mathbf{x}, A = a) = P(\hat{Y}_{A \leftarrow a'}(\mathbf{U}) = y \mid \mathbf{X} = \mathbf{x}, A = a). \tag{2}$$

This ensures that the predicted outcome $\hat{Y}$ remains unchanged when the individual's PA A is counterfactually altered. This condition is fulfilled in the FiND world.[2] In addition, counterfactual fairness does not necessarily fulfill fairness on the group level [56], which is a key difference to the FiND world (see Section 2.2.1).

2.1.4 FiND World vs. Individual Fairness

Individual fairness, as proposed by Dwork et al. [25], postulates that similar individuals should be treated similarly. This is formalized using task-specific similarity metrics that ensure that the outcomes for two individuals of different protected groups are similar if their features are similar. However, they acknowledge the difficulty of specifying these metrics and propose to base their choice on the context of the task and feature space. The FiND world ties in with this by normatively redefining the"task-specific equality" on the basis of social and legal requirements. This involves a transformation procedure from the observed real world – where dissimilarities based on PAs may exist but are deemed unjust – to a FiND world – where these dissimilarities are removed. This means individuals are treated equally if they are equal in the FiND world (and if the predictor is individually well-calibrated, see [9], Def. 3.8) and thereby, individual level fairness is fulfilled. Note that this is generally different from the approach of Dwork et al. [25], where similarity is evaluated in the real world.

2.2 Relation to Fairness Trade-offs

We can derive relations between the FiND world and the resolution of several fairness trade-offs. A key factor responsible for this lies in the fact that the FiND world also inherently fulfills several group fairness notions.

[2] See [9], Section 3.4, for a more detailed comparison of the two concepts.

2.2.1 Relation to Group Fairness

The assumption of $Y \perp\!\!\!\perp A$ in the FiND world has important implications for the fulfillment of several group metrics, as also highlighted by Loftus et al. [43]. They make a similar connection between group fairness and counterfactual fairness for cases where there are no causal paths between Y and A. This exactly is characterized in the FiND world.

Demographic Parity. Perhaps the most common group metric is *demographic parity*, defined as follows:

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a'), \tag{3}$$

for all $y \in \{0,1\}$, $a, a' \in \mathscr{A}$. An alternative way of framing this condition is by the assumption of *Independence* [3],

$$\hat{Y} \perp\!\!\!\perp A. \tag{4}$$

Since A is assumed to be a root node and no back door paths can exist, any statistical dependence between A and $\hat{Y}$ is either due to a causal effect from A on Y or introduced by a common child X. From $A \perp\!\!\!\perp \mathbf{X}^*, Y^*$ in the FiND world immediately follows that a predictor $\hat{Y}^*$ trained on its representation, i.e., $\hat{Y}^* = \hat{f}(\mathbf{X}^*)$ satisfies demographic parity.

Equalized Odds. Similar connections apply to *equalized odds*, which requires that both groups have equal error rates. Therefore, it includes both the conditions of *false positive error rate balance* and *false negative error rate balance*. Formally, equalized odds is defined as:

$$P(\hat{Y} = y \mid A = a, Y = y) = P(\hat{Y} = y \mid A = a', Y = y), \tag{5}$$

for all $y \in \{0,1\}$, $a, a' \in \mathscr{A}$. This condition ensures that the predictor $\hat{Y}$ is conditionally independent of A given the true outcome Y, which also falls under the term *Separation* [3] and can be written as:

$$\hat{Y} \perp\!\!\!\perp A \mid Y. \tag{6}$$

The independence of the target Y^* (1) and the predictor $\hat{Y}^*$ (4) w.r.t. the PA A in the FiND world directly imply that the predictor satisfies (6) by the compositional and weak union axioms of the compositional graphoid induced by the causal DAG [41].

Predictive Parity. In the same way, relations can be drawn to *Predictive Parity*[3], which ensures that for a predicted outcome $\hat{Y}$ individuals in both protected and unprotected group have equal probability to truly belong to class Y. Formally, this requires:

$$P(Y = y \mid \hat{Y} = y, A = a) = P(Y = y \mid \hat{Y} = y, A = a'), \tag{7}$$

for all $y \in \{0,1\}$, $a, a' \in \mathscr{A}$. This condition is also known as *Sufficiency* [3] and implies that the true outcome Y is independent of A when conditioned on the predicted outcome $\hat{Y}$, expressed as:

$$Y \perp\!\!\!\perp A \mid \hat{Y}. \tag{8}$$

Following the same reasoning as for equalized odds, (8) is a direct consequence of $Y^* \perp\!\!\!\perp A$, leading to predictive parity being inherently satisfied in the FiND world.

[3]For hard-label predictors, this condition is referred to as *predictive parity*, as it ensures equal positive predictive values (PPV) for a and a'. For predicted scores S, this condition is commonly termed *calibration*, demanding that for each score s the probability of belonging to the positive class is equal for both a and a'

2.2.2 *Resolving the Trade-off Between Group Fairness Metrics*

Parts of the fairML literature surround the problem that several group fairness metrics are mathematically incompatible with each other. This is known as the *impossibility theorem* [15, 36], which states that false positive rate balance, false negative rate balance, and predictive parity cannot be satisfied simultaneously unless for two special cases: one must either have equal group base rates (also falls under the term *prevalence*) for the protected and unprotected group or perfect prediction.

However, as we have just derived, these metrics are all inherently fulfilled in the FiND word. In fact, the FiND world represents one of the special cases of the impossibility theorem, namely equal base rates among groups [22]. This is due to the independence assumption of $Y \perp\!\!\!\perp A$ of the FiND world, which implies that individuals from protected and unprotected groups have the same probability of belonging to the positive class:

$$P(Y = 1|A = a) = P(Y = 1|A = a') = P(Y = 1). \tag{9}$$

This assumption can be normatively justified if we presume that different group base rates in the real world are entirely attributable to historical discrimination and biases stemming from the data collection process. Therefore, the FiND world is able to overcome the trade-off between competing fairness metrics naturally by design.

2.2.3 *Resolving the Conflict Between Group Fairness and Individual Fairness*

Additionally, as the FiND world approaches fairness on the individual level, it also overcomes a conflict between group fairness and individual fairness notions. A similar observation is made by Binns [6], who argue that the conflict between individual and group fairness depends on the worldview and the underlying normative principles a decision-maker considers. By, e.g., adopting a WAE worldview, both individual and group approaches are driven by the same moral assumptions that groups should be treated equally based on their PAs. The same considerations apply to the FiND world, as observed group differences are assumed to be due to biases in the real world and should therefore not be taken into account.

Coming back to the financial lending example, the existence of discrimination based on race and gender has repeatedly been observed in various US mortgage markets [42]. From a policy perspective, to base the approval of a loan on such historical disparities would constitute an unjust practice and is legally restricted under, e.g., the Fair Housing Act and the Home Mortgage Disclosure Act (HMDA). This raises the need for decision-makers (in this scenario, banks) to proactively implement *substantive equality* as described by Wachter et al. [59]. They argue that in order to achieve fair treatment, one has to account for historical inequalities by actively re-leveling discriminated individuals via *bias-transforming* actions (which can also be understood as some type of *affirmative action*). This can be connected to the transformation from the real world to FiND world.

2.2.4 *Resolving the Fairness-Accuracy Trade-off*

As we move the prediction process into the FiND world, we are also able to overcome the fairness-accuracy trade-off. Since all the investigated fairness notions are inherently embedded within the FiND world, enforcing a fairness metric no longer leads to a decrease in predictive performance; instead, fairness and predictive performance become aligned. Furthermore, this alignment enables us to reframe the quest for fair models: We can focus on high predictive performance using the data representation of the FiND world for which fairness naturally holds.

The FiND world thus resolves both fairness trade-offs by providing a unified causal framework. Moreover, our findings have significant practical implications: if we can approximate the FiND world through suitable pre-processing methods, it becomes possible to overcome these trade-offs in real-world applications. Successfully doing so would eliminate the need to explicitly enforce specific fairness metrics, allowing high predictive performance to naturally serve as the primary notion of fairness.

3 Approximating the FiND World

3.1 Pre-processing

Since we do not have access to data from the FiND world in practice, we need to project the real-world data into the FiND world as a preliminary step. This can be understood as some type of *fair representation learning* [65]. Various methods have been proposed to learn fair representations of the data (e.g., [11, 40, 53, 61, 65]). However, most of them present non-causal techniques and do not apply to the specific assumptions of the FiND world. Against this background, we consider two causal pre-processing techniques to approximate the FiND world.

With *fairadapt*, Plecko et al. [50] present a pre-processing method using a causal approach. Their method is based on quantile preservation and uses quantile regression forests [45]. They aim to approximate a counterfactual world by constructing "fair twins". Their method constructs these fair twins by transforming the observational distribution for each individual to a fair-projection distribution. In doing so, they demand all individuals to have the same PA after the fair twin projection, thereby setting the protected group to its baseline value. Note that this differs slightly from the warping method (see below), which does not intervene on the PA but only on its descendants. However, this distinction is not a problem here, as the protected attribute from the real world can replace the baseline value in the joint distribution after the projection of the features $\mathbf{X}$ to meet the definition of the FiND world. We use their R package `fairadapt`[4] to adapt the train and the test dataset, and in the following refer to this as the "adapted world" data.

Bothmann et al. [8] propose another causal approach by introducing a *residual-based warping* method to approximate the FiND world. Their method consists of intervening on the paths from the PA to its descendants by transforming the observations for the protected group to the corresponding quantile of the unprotected group's distribution. They reduce the problem of estimating full distributions to estimating models for the location parameters of their distributions. Therefore, they derive individual probability ranks by using a residual-based approach. The method thereby derives "rank-preserving interventional distributions" (RPID), i.e., individuals of the protected group maintain the group-specific pre-intervention ranks for each warped variable. By symmetry, reversing the warping direction from the unprotected to the protected group is also feasible, as it produces the same result of eliminating the effect of the PA. We use R functions from their GitHubrepository[5] for warping training and test data, which will be referred to as the "warped world" data.

3.2 In-processing

We propose *appFiND*, a new evaluation method consisting of in-processing to assess whether the pre-processing successfully approximates the FiND world. In-processing methods generally work by enforcing a desired fairness

[4] https://cran.r-project.org/web/packages/fairadapt/index.html

[5] https://github.com/slds-lmu/paper_2024_rpid

constraint during model training, see, e.g., [12, 19, 20, 23, 63, 64]. In particular, we are interested in whether for a model trained on i.i.d data from the real world, $\mathscr{D}_{\text{train}} \coloneqq \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, and evaluated on pre-processed data, the performance improves when enforcing a fairness constraint. For this purpose, we consider the following constraint optimization problem, which we implement by adding a fairness regularization term to the empirical risk:

Regularized Empirical Risk. For individuals $i \in \{1, 2, \ldots, N\}$ in $\mathscr{D}_{\text{test}}$, their corresponding targets $y_i \in \{0, 1\}$ and features $\mathbf{x}_i \in \mathbb{R}^d$, we obtain the regularized empirical risk

$$R_{\text{reg}}(f, \mathscr{D}_{\text{train}}) = \sum_{i=1}^{N} L\big(y_i, f(\mathbf{x}_i)\big) + \lambda \cdot C(\hat{\boldsymbol{\pi}}_{\text{train}}), \tag{10}$$

where $L\big(y_i, f(\mathbf{x}_i)\big)$ denotes the Bernoulli loss with predicted log-odds $f(\mathbf{x}_i)$, and $\hat{\boldsymbol{\pi}}_{\text{train}}$ the vector of predicted probabilities for all data samples $\mathbf{x}_i \in \mathscr{X}_{\text{train}} = \{\mathbf{x}_i\}_{i=1}^N$ with

$$\hat{\pi}_i \coloneqq \sigma\big(f(\mathbf{x}_i)\big) = \frac{1}{1 + e^{-f(\mathbf{x}_i)}}. \tag{11}$$

The regularization term $C(\hat{\boldsymbol{\pi}})$ represents the fairness constraint controlled by the penalty parameter λ. We refer to $C(\hat{\boldsymbol{\pi}})$ as a regularization term, yet its purpose is not to reduce model complexity or improve generalization, but to favor models that lead to fairer predictions.

Fairness constraint C. We penalize the difference in the average predicted probabilities between the protected group memberships a and a' with

$$C(\hat{\boldsymbol{\pi}}) \coloneqq \left| \frac{1}{N_a} \sum_{i:A_i=a} \hat{\pi}_i - \frac{1}{N_{a'}} \sum_{j:A_j=a'} \hat{\pi}_j \right|, \tag{12}$$

where N_a and $N_{a'}$ denote the total number of individuals per group. This can be seen as a form of demographic parity constraint [1, 37] that is a necessary condition of the FiND world as we proved in Section 2.2.1.

Gradient Boosted Trees. To learn our model, we use gradient boosted trees in the R implementation of `xgboost` [13]. We split our data randomly into 80% training data $\mathscr{D}_{\text{train}}$ and 20% test data $\mathscr{D}_{\text{test}} \coloneqq \{(\mathbf{x}_i, y_i)\}_{i=N+1}^M$ and determine optimal hyperparameters d (depth of the trees) and η (learning rate) via random search with 3-fold cross-validation on $\mathscr{D}_{\text{train}}$.

3.3 Fairness-Accuracy Trade-off Evaluation

To assess whether the pre-processing is able to resolve the fairness-accuracy trade-off, *appFiND* investigates whether the performance of a model evaluated on pre-processed data improves as the fairness constraint C, used for training on real-world data, increases. As a preliminary step, we find the smallest penalty parameter λ^* for which $C\big(f(x)\big) < \varepsilon$ is fulfilled via grid search (trained on $\mathscr{D}_{\text{train}}$ with the above optimal hyperparameters and evaluated on $\mathscr{D}_{\text{test}}$) . Since $\varepsilon = 0$ is unfeasible for probabilistic classifiers (for non-trivial cases such as $\hat{\boldsymbol{\pi}} = 1.0$ or $\hat{\boldsymbol{\pi}} = 0.0$), we set $\varepsilon = 0.01$, thereby allowing fairness differences of up to 1%[6].

[6]The choice of $\varepsilon = 0.01$ is in some sense arbitrary and can be adapted based on the application context. While it allows fairness differences of up to 1%, this threshold may lead to more substantial disparities when base rates are low. Alternatively, one could consider making ε relative to the base rate to better account for context-specific disparities.

Algorithm 1 Fairness-Accuracy Trade-off Evaluation

Input Training dataset $\mathscr{D}_{\text{train}}$, test dataset $\mathscr{D}_{\text{test}}$, learning rate η, tree depth d, optimal λ^*, interpolation steps S
Output AUC values $\mathbf{p}$, empirical group disparities $\hat{\delta}$

```
for n ← 0, 1, ..., S+1 do
    w ← n/(S+1)
    f̂ ← argmin_{f∈ℋ} Σ_{x_i∈𝒟_train} L(y_i, f(x_i)) + w·λ*·C(π̂_train)   ▷ Train model f̂ with parameters η, d with C from (12)
    for x_i ∈ 𝒳_test do
        π̂_i ← σ(f̂(x_i))                                   ▷ Predict probabilities for test features 𝒳_test := 𝒟_test \ 𝒴_test
    end for
    δ̂_n ← C(π̂_test)                                       ▷ Evaluate fairness via group disparity defined in (12)
    p_n ← auc(π̂_test, 𝒴_test)                              ▷ Evaluate performance via AUC
end for
```

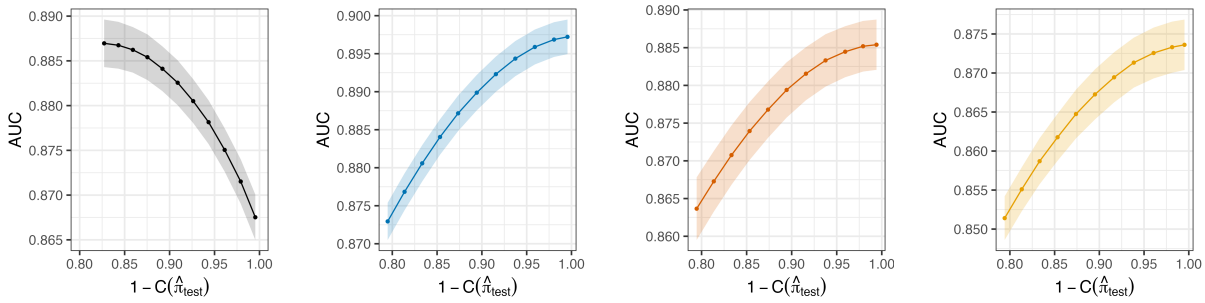
We then apply Algorithm 1 that trains and evaluates models for different penalty parameters $\lambda = w\lambda^*$, where w ranges from 0.0 to 1.0 in increments of $\frac{1}{S+1}$, i.e., interpolates between the unconstrained model and the one with an $(1-\varepsilon)$-fulfillment of the fairness constraint C. In our experiments, we set the number of interpolation steps to $S = 9$ leading to 11 models in total. To evaluate the relation between fairness and accuracy, we record for each model its fulfillment of fairness by measuring the difference in predicted group probabilities and evaluate performance using the area under the ROC curve (AUC). We choose AUC over accuracy as the performance measure, since it allows for a more nuanced assessment when classes are imbalanced.

4 Resolving the Trade-offs via Pre-Processing

4.1 Testing Approximation to the FiND World

The approximation of the FiND world via pre-processing can be evaluated using *appFiND* (Algorithm 1), based on the following rationale: We already derived in Section 2.2.1 that in the FiND world demographic parity (DP) holds. Thus, enforcing the fairness constraint C during model training using real world training data $\mathscr{D}_{\text{train}}$ should increase performance on FiND world test data $\mathscr{D}^*_{\text{test}}$, as the FiND world accurately reflects the enforced fairness. If for pre-processed test data we observe this same relationship – namely the same increase in performance – we can conclude that the pre-processed data was able to learn the representation of the FiND world. In the following, we investigate whether the fairadapt and warping pre-processing methods are able to approximate the FiND world and can therefore overcome the fairness-accuracy trade-off and the trade-off between several fairness metrics. To further validate the successful approximation of the FiND world by the pre-processing methods, we additionally check whether the differences in distributions of transformed variables between groups were resolved (see A.2 and B.2).

Simulation Setup. We return to the credit application example introduced in Section 2.1.1. According to the corresponding DAG depicted in Figure 1, we generate synthetic data for both worlds. We consider the PA to be binary and equally distributed, i.e., $\pi_a = 50\%$ (this can, e.g., mimic gender). The simulated real world model contains direct causal effects of the PA A on the features $\mathbf{X}$ and target Y, while the FiND world model has no causal path from A to neither $\mathbf{X}$ nor Y. For more details on the setup, see Appendix A.1. In total, we conduct 25



(a) Real world $\mathscr{D}_{\text{test}}$ (b) FiND world $\mathscr{D}^{*}_{\text{test}}$ (c) Adapted world $\mathscr{D}^{\text{a}}_{\text{test}}$ (d) Warped world $\mathscr{D}^{\text{w}}_{\text{test}}$

Fig. 2. Fairness-performance curves of predictors trained on real-world data $\mathscr{D}_{\text{train}}$ for the simulation study in Section 4.2, with 95% confidence intervals. *(a)* Real world test data does not approximate *(b)* the FiND world due to inherent bias. By contrast, *(c)* the adapted world and *(d)* warped world data do approximate the FiND world, as increasing fairness leads to increased AUC (same monotonic relation as for *(b)* FiND world data).

simulation runs that each produces a dataset for the real and FiND world. We then apply the fairadapt and the warping pre-processing to all simulated real world datasets. (Code for all experiments will be published upon acceptance.)

4.2 Resolving the Trade-off between Fairness and Performance

Figure 2 presents the fairness-accuracy trade-offs in the different worlds. Using Algorithm 1, we train models for increasing weights of fairness on real world $\mathscr{D}_{\text{train}}$, but evaluate them regarding performance (AUC) and fairness on real world, FiND world, adapted world, and warped world test data, respectively. We use $1 - C(\hat{\pi}_{\text{test}})$ as a fairness measure for better interpretability (high value is better). We additionally record the 95% confidence intervals of the AUC.

For models trained and tested on real world data, we observe the typical trade-off between fairness and performance (Figure 2a). In Figure 2b, we evaluate the same set of classifiers trained on real world data, but this time measure performance on FiND world $\mathscr{D}^{*}_{\text{test}}$. We see the exact opposite pattern: Classifiers with low fairness violation $C(\hat{\pi}_{\text{test}})$ are also more accurate. This also aligns with our theory: When evaluating models that enforce fair predictions (in terms of DP) on data that corresponds to the FiND world where fairness is naturally fulfilled, a model that is "fairer" will also be more accurate.

We observe the same relationship between fairness and performance also for adapted world (Figure 2c) and warped world (Figure 2d) test data. On both pre-processed test sets, the performance of models increases with lower fairness violations. We conclude that the pre-processed data are equally capable of reflecting a world that incorporates fairness in the sense that there are no differences between groups in their outcomes and thus, $Y \perp\!\!\!\perp A$. Consequently, there can be no direct or indirect causal effects from PA A on Y. Therefore, we conclude that both pre-processing methods are able to overcome the trade-off between fairness and performance in the simulated setting and were successful in approximating the FiND world.

Table 1. Comparison of fairness and performance metrics in real, adapted, and warped world for *(a)* the simulation study to evaluate the pre-processing methods (Section 4.1) and *(b)* the HMDA experiment (Section 5). For *(a)* the simulation study, we additionally record the standard deviation (sd) over the iterations in the subscript. For confidence intervals, see Table 2 in A.2.

(a) Simulation Study

World	Fairness				Performance
	DP_{sd}	FPR_{sd}	FNR_{sd}	PPV_{sd}	AUC_{sd}
Real	$0.825_{0.020}$	$0.782_{0.024}$	$0.954_{0.017}$	$0.986_{0.014}$	$0.887_{0.007}$
FiND	$0.987_{0.011}$	$0.989_{0.011}$	$0.991_{0.007}$	$0.988_{0.012}$	$0.897_{0.006}$
Adapted	$0.982_{0.012}$	$0.972_{0.016}$	$0.984_{0.013}$	$0.975_{0.015}$	$0.886_{0.008}$
Warped	$0.982_{0.011}$	$0.964_{0.015}$	$0.974_{0.016}$	$0.971_{0.018}$	$0.893_{0.013}$

(b) HMDA Experiment

World	Fairness				Performance
	DP	FPR	FNR	PPV	AUC
Real	0.809	0.840	0.823	0.891	0.716
FiND	—	— unkown —		—	—
Adapted	0.989	0.991	0.989	0.990	0.712
Warped	0.994	0.988	0.990	0.975	0.713

4.3 Resolving the Trade-off between Fairness Metrics

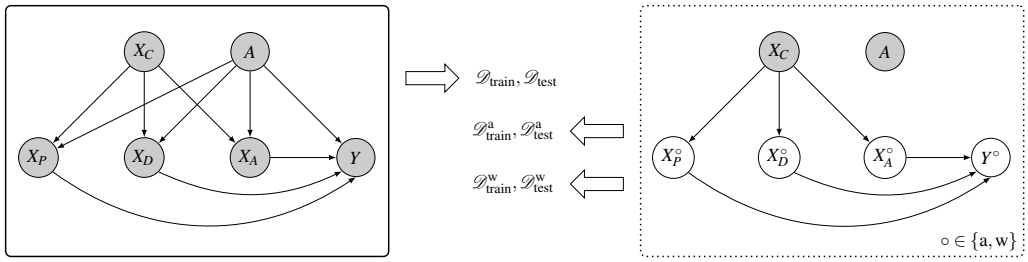
The results from Section 4.2 offer another important insight. When the pre-processing methods are successful in approximating the FiND world, we do not have to rely on using the in-processing method anymore to achieve fairness by enforcing a fairness constraint. Instead, we can directly train models on pre-processed data that reflects the FiND world in order to achieve fair predictions.

In the following, we train unconstrained models on the adapted world and warped world datasets, evaluating them in terms of the compatibility of various fairness metrics. We record the fulfillment of fairness regarding demographic parity (DP), false positive rate balance (FPR), false negative rate balance (FNR) and predictive parity (PPV) between the protected a and unprotected group a' (averaged over 25 simulations). Regarding performance, we report the average AUC. To allow for comparative analyses, we perform the same evaluations on the real world and FiND world datasets. The results are listed in Table 1a.

In the real world, we observe large fairness differences for DP and FPR up to 22%. In the FiND world, we obtain much higher fairness values, the empirical differences are less than 1.3%. This matches our theoretical results that in the FiND world the concerned fairness metrics are inherently satisfied. The same pattern can be validated for the adapted and warped world, although we observe slightly higher differences than in the FiND world of up to 3.6%. However, this is still considerably lower than in the real world. We therefore consider both pre-processed worlds to approximately fulfill all fairness metrics. From this, we conclude that in our simulated setting the fairadapt and warping pre-processing methods are both able to successfully overcome the trade-off between several fairness metrics. We also conclude that models of real, FiND, adapt and warped world are all equally well performing (where FiND, adapted and warped have even slightly better AUCs than the real world model). However, note that the performance of the real world model is not directly interpretable as it still inherits bias.

5 Real Data Experiments

To further validate our results, we demonstrate that our findings from the simulated setup can be transferred to real data experiments for which the FiND world model is inaccessible and needs to be approximated. For this purpose,



(a) Observed real world *(b)* Approximated FiND worlds

Fig. 3. Assumed causal DAGs for the HMDA dataset of *(a)* the real and *(b)* the approximated FiND world alongside their corresponding datasets. The non-shaded nodes in the approximated worlds are transformed since they are descendants of the PA A in the real world. They are denoted with either a or w to indicate the applied transformation, FairAdapt [51] or Warping [9].

we use the 2022 Home Mortgage Disclosure Act (HMDA) dataset for the state Wisconsin.[7] This dataset is similar to our previously considered credit assessment example and contains information on housing loan applications, including details about the applicants and the corresponding loan decisions.

We consider race as the binary PA A. We filter the data regarding race by only selecting Black (protected group a) and non-Hispanic White borrowers (unprotected group a'), as previous mortgage studies have highlighted the most significant disparities in approval rates and treatment between these two groups [42]. This results in a total of 83,808 observations. The original dataset encompasses over 100 variables, for simplicity, we focus on the subset of the following six: We include loan amount (X_A), purpose of the loan (X_P), and debt ratio (X_D) as features as well as age and gender ($\mathbf{X}_C$) jointly as confounders. The target Y indicates whether an applicant was granted a loan (1) or not (0)[8]. For more details on the data setup, see Appendix B. Figure 3 shows the assumed DAGs. The final dataset consists of 94.8% White vs. 5.2% Black applicants and has an overall loan approval base rate of 66.8%. The group-specific base rate for White applicants is 67.9%, and for Black applicants 48.6%.

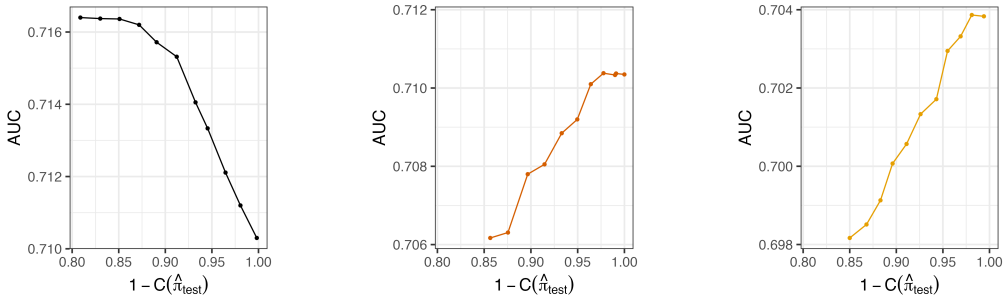
5.1 Resolving the Trade-off between Fairness and Performance

As we do not have access to the FiND world in this real data setting, we rely solely on pre-processing to obtain unbiased data that reflects the FiND world. For this purpose, we apply the fairadapt and the warping pre-processing methods on the real HMDA train and test data. To check if the pre-processing successfully eliminated all causal effects from the PA A on Y and is thereby able to approximate the FiND world, we apply *appFiND* (Algorithm 1). The results are displayed in Figure 4. Additionally, we compare distributions in Appendix B.2.

For the real world HMDA data in Figure 4a, we again observe the common trade-off between fairness and performance, where an increase in fairness from 80% (fairness of the unconstraint model) to 100% is accompanied by a decrease in performance. The absolute decrease in performance is rather small in this case, which is due to the highly imbalanced classes of the PA where a' only makes up 5% of the data. In Figure 4b and Figure 4c, we again see a positive connection between fairness and performance. Models that enforce higher fairness constraints during

[7]The dataset is accessible at https://ffiec.cfpb.gov/data-browser/data/2022?category=states&items=WI.

[8]Note that the target Y in this case indicates the loan approval and does not contain information about whether the individual repaid the loan or not. This varies slightly from the previous credit application example, where Y denoted the risk of credit repayment.



(a) Real $\mathscr{D}_{\text{test}}$ (b) Adapted $\mathscr{D}^{\text{a}}_{\text{test}}$ (c) Warped $\mathscr{D}^{\text{w}}_{\text{test}}$

Fig. 4. Fairness-performance curves of predictors trained on HMDA $\mathscr{D}_{\text{train}}$.

training on real world train data also achieve higher performance on pre-processed test data. On both pre-processed test sets, we see a range in fairness increase from 85% (unconstraint model) up to 100% which is accompanied by an increase in performance. This increase is again rather small in absolute values but nonetheless indicates that the pre-processed data reflect a world that incorporates this fairness. However, we observe differences between fairadapt and warping as for warping we see a larger increase in performance while for fairadapt, the performance is higher from the start (see varying y-axis values). Yet, we have to be cautious in over-interpreting these differences in the AUC as we are more interested in the general trend towards increased performance. Similarly, the AUC on the real world data is consistently higher than on pre-processed data. Note that this higher performance must not be misinterpreted as favoring real-world data for evaluation; on the contrary: while the fully constrained real world classifier is now fair wrt. DP, it still carries bias from the unfair representation of the data it was trained on. Instead, in order to obtain a model that incorporates a fair representation of the data, we need to train and test models on the pre-processed data and only in this way we can meaningfully evaluate performance.

5.2 Resolving the Trade-off between Fairness Metrics

We observe that the pre-processing methods are also able to overcome the trade-off between several fairness metrics for the HMDA data. We proceed in the same manner as in Section 4.3 by directly training and evaluating models on the pre-processed data. The results are displayed in Table 1b. For the real HMDA data, the fairness metrics are not simultaneously satisfied, with fairness differences up to 20%. However, for the adapted and warped data, all fairness metrics are approximately fulfilled. Regarding performance, we observe that models trained and evaluated on real, adapted, and warped data all achieve similarly high AUCs.

6 Summary & Discussion

By establishing a new link between previous findings on fairML trade-offs and the causal framework of the FiND world, we show that both the trade-off between different fairness metrics and the trade-off between fairness and accuracy can be resolved, both theoretically and practically. We show that the FiND world naturally satisfies multiple fairness notions at both the individual and group levels, including those subject to the impossibility theorem – thereby addressing the first trade-off. This theoretical foundation also reveals how the second trade-off

can be overcome: by learning a FiND world representation of the data through pre-processing methods, practitioners can achieve high predictive performance while maintaining fairness. To operationalize these insights, we introduce a novel evaluation method *appFiND* that assesses the quality of FiND world approximations in practice.

However, for the purpose of our study, we considered all dependencies between the PA and the target to be unjust – and thus eliminated all corresponding causal paths. This varies from several other proposals in fairML literature that only eliminate some path-specific effects from the PA on the target as they consider some features dependent on the PA to be resolving variables [14, 34, 51]. Bothmann et al. [9] also consider an alternative definition of the FiND world in which such paths can be deemed fair and need not be removed. Future work could investigate how to extend our findings on such alternative FiND worlds, e.g., using [51], who propose conditional fairness metrics. The question of which variables can be considered (non-)resolving also depends on the specific use case.

The assumed knowledge of the causal graph is a necessary requirement for causal pre-processing methods, otherwise it is limited to purely statistical relations. Learning the true causal graph from data is beyond the scope of this work and forms a separate research branch known as causal discovery or causal structure learning. For a recent overview of the field, we refer the reader to Squires and Uhler [58] and Kitson et al. [35]. Nevertheless, we highlight the limitation of a correctly specified DAG. In the presence of latent confounding or selection bias, the resulting independence model can no longer be expressed by a DAG over the observed variables only. Besides, the misspecification of the DAG can also degrade the effectiveness of the pre-processing methods in eliminating all causal effects from the PA on Y. In the HMDA dataset, Gender can also be considered an additional PA which we did not consider in our study. Modeling multiple PAs opens up the topic of intersectionality (see, e.g., [10, 33, 48, 62]), which imposes another major challenge in fairML.

Since the FiND world represents a causal framework, we have focused on causal pre-processing methods. However, future work could include evaluating non-causal pre-processing methods in their ability to approximate the FiND world. Moreover, our study considered relatively simple DAGs. To further validate our results, applications involving structurally more complex DAGs with a larger number of variables could be explored. While the shown theory is independent from the complexity of the DAGs, it would be interesting to analyze how well the transformation methods scale. Additionally, analyzing further real-world datasets would help to further demonstrate the practical applicability and generalizability of our approach.

7 Conclusion & Outlook

Our results offer practical solutions for fairness-aware machine learning, where one does not have to choose between fairness and high predictive performance anymore, nor between several classical fairness metrics. Instead, one can shift the modeling and evaluation process to an approximated FiND world, using appropriate pre-processing methods such as the fairadapt method or warping. To this end, we provide practitioners with *appFiND*, an evaluation method that indicates whether the applied pre-processing method has successfully approximated the FiND world. By directly training and testing models on data that approximates this FiND world, we overcome the need to explicitly enforce certain fairness metrics, as the approximation of the FiND world already incorporates a holistic assurance of these fairness metrics. Rather, we can now focus on "just" achieving high predictive performance. We have shown in a simulation study that the two covered pre-processing methods are able to approximate the FiND world and showcased an application with real-world mortgage data.

Future work could investigate if other pre-processing methods could be used for approximating the FiND world and compare different pre-processing methods on diverse simulated and real-world data sets. Worthwhile directions are further the incorporation of multiple PAs, aiming at intersectionality, and an investigation of the sensitivity of the methods with respect to DAGs (partially) learned from data.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica* (May 23 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[4] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. 2023. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 400–422. https://doi.org/10.1145/3593013.3594007

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. http://arxiv.org/abs/1706.02409

[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 514–524. https://doi.org/10.1145/3351095.3372864

[7] Avrim Blum and Kevin Stangl. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference) (LIPIcs, Vol. 156)*, Aaron Roth (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:20. https://doi.org/10.4230/LIPICS.FORC.2020.3

[8] Ludwig Bothmann, Susanne Dandl, and Michael Schomaker. 2023. Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, October 1st, 2023 (CEUR Workshop Proceedings, Vol. 3523)*, Roberta Calegari, Andrea Aler Tubella, Gabriel González-Castañé, Virginia Dignum, and Michela Milano (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-3523/paper1.pdf

[9] Ludwig Bothmann, Kristina Peters, and Bernd Bischl. 2024. What Is Fairness? On the Role of Protected Attributes and Fictitious Worlds. http://arxiv.org/abs/2205.09622

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[11] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html

[12] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 319–328. https://doi.org/10.1145/3287560.3287586

[13] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, and Jiaming Yuan. 2014. xgboost: Extreme Gradient Boosting. https://doi.org/10.32614/CRAN.package.xgboost Institution: Comprehensive R Archive Network Pages: 1.7.8.1.

[14] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 7801–7808. https://doi.org/10.1609/AAAI.V33I01.33017801

[15] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/BIG.2016.0047

[16] A. Feder Cooper, Ellen Abrams, and Na Na. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 46–54. https://doi.org/10.1145/3461702.3462519

[17] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. *J. Mach. Learn. Res.* 24 (2023), 312:1–312:117. http://jmlr.org/papers/v24/22-1511.html

[18] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 797–806. https://doi.org/10.1145/3097983.3098095

[19] Andrew Cotter, Heinrich Jiang, Maya R. Gupta, Serena Lutong Wang, Taman Narayan, Seungil You, and Karthik Sridharan. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *J. Mach. Learn. Res.* 20 (2019), 172:1–172:59. https://jmlr.org/papers/v20/18-616.html

[20] André F. Cruz, Catarina G. Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. 2023. FairGBM: Gradient Boosting with Fairness Constraints. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=x-mXzBgCX3a

[21] André F. Cruz and Moritz Hardt. 2024. Unprocessing Seven Years of Algorithmic Fairness. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=jr03SfWsBS

[22] MaryBeth Defrance and Tijl De Bie. 2025. Maximal Combinations of Fairness Definitions. *J. Artif. Intell. Res.* 82 (2025), 1495–1579. https://doi.org/10.1613/JAIR.1.16776

[23] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 2796–2806. https://proceedings.neurips.cc/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html

[24] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2803–2813. https://proceedings.mlr.press/v119/dutta20a.html

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness Through Awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. https://doi.org/10.1145/2090236.2090255

[26] Evanthia Faliagka, Athanasios K. Tsakalidis, and Giannis Tzimas. 2012. An Integrated E-Recruitment System for Automated Personality Mining and Applicant Ranking. *Internet Res.* 22, 5 (2012), 551–568. https://doi.org/10.1108/10662241211271545

[27] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? A study of label and selection bias. *Mach. Learn.* 112, 12 (2023), 5081–5104. https://doi.org/10.1007/S10994-023-06401-1

[28] Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, Sanjay Chawla Venkatasubramanian and Wagner Meira Jr. (Eds.). SIAM, 144–152. https://doi.org/10.1137/1.9781611974348.17

[29] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. http://arxiv.org/abs/1609.07236

[30] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (April 2021), 136–143. https://doi.org/10.1145/3433949

[31] Sofie Goethals, Toon Calders, and David Martens. 2024. Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics. http://arxiv.org/abs/2401.13391

[32] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[33] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 2569–2577. http://proceedings.mlr.press/v80/kearns18a.html

[34] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 656–666. https://proceedings.neurips.cc/paper/2017/hash/f5f8590cd58a54e94377e6ae2eded4d9-Abstract.html

[35] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. 2023. A survey of Bayesian Network structure learning. *Artif. Intell. Rev.* 56, 8 (2023), 8721–8814. https://doi.org/10.1007/S10462-022-10351-W

[36] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23. https://doi.org/10.4230/LIPICS.ITCS.2017.43

[37] Nikola Konstantinov and Christoph H. Lampert. 2021. Fairness Through Regularization for Learning to Rank. https://arxiv.org/abs/2102.05996

[38] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. 2022. Fairness in credit scoring: Assessment, implementation and profit implications. *Eur. J. Oper. Res.* 297, 3 (2022), 1083–1094. https://doi.org/10.1016/J.EJOR.2021.06.023

[39] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4066–4076. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[40] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[41] Steffen Lauritzen and Kayvan Sadeghi. 2018. Unifying Markov properties for graphical models. *The Annals of Statistics* 46, 5 (2018), 2251 – 2278. https://doi.org/10.1214/17-AOS1618

[42] Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. *Minds Mach.* 31, 1 (2021), 165–191. https://doi.org/10.1007/S11023-020-09529-4

[43] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal Reasoning for Algorithmic Fairness. http://arxiv.org/abs/1805.05859

[44] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. 2021. Does enforcing fairness mitigate biases caused by subpopulation shift?. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 25773–25784. https://proceedings.neurips.cc/paper/2021/hash/d800149d2f947ad4d64f34668f8b20f6-Abstract.html

[45] Nicolai Meinshausen. 2006. Quantile Regression Forests. *J. Mach. Learn. Res.* 7 (2006), 983–999. https://jmlr.org/papers/v7/meinshausen06a.html

[46] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 107–118. https://proceedings.mlr.press/v81/menon18a.html

[47] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, Volume 8, 2021 (March 2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902 Publisher: Annual Reviews.

[48] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2020. Auditing and Achieving Intersectional Fairness in Classification Problems. http://arxiv.org/abs/1911.01468

[49] Drago Plecko and Elias Bareinboim. 2024. Fairness-Accuracy Trade-Offs: A Causal Perspective. https://doi.org/10.48550/ARXIV.2405.15443

[50] Drago Plecko, Nicolas Bennett, and Nicolai Meinshausen. 2024. fairadapt: Causal Reasoning for Fair Data Preprocessing. *J. Stat. Softw.* 110, 4 (2024). https://doi.org/10.18637/JSS.V110.I04

[51] Drago Plečko and Nicolai Meinshausen. 2020. Fair Data Adaptation with Quantile Preservation. *Journal of Machine Learning Research* 21 (2020), 1–44. http://jmlr.org/papers/v21/19-966.html

[52] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.* 3, 10 (2021), 896–904. https://doi.org/10.1038/S42256-021-00396-X

[53] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin T. Vechev. 2020. Learning Certified Individually Fair Representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/55d491cf951b1b920900684d71419282-Abstract.html

[54] Mohit Sharma and Amit Deshpande. 2024. How Far Can Fairness Constraints Help Recover From Biased Data?. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=RfQT6vJt8b

[55] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York NY USA, 358–364. https://doi.org/10.1145/3375627.3375865

[56] Ricardo Silva. 2024. Counterfactual Fairness Is Not Demographic Parity, and Other Observations. https://doi.org/10.48550/arXiv.2402.02663

[57] Oleg Sofrygin, Mark J. van der Laan, and Romain Neugebauer. 2017. simcausal R Package: Conducting Transparent and Reproducible Simulation Studies of Causal Effect Estimation with Complex Longitudinal Data. *Journal of Statistical Software* 81, 2 (2017), 1–47. https://doi.org/10.18637/jss.v081.i02

[58] Chandler Squires and Caroline Uhler. 2022. Causal structure learning: a combinatorial perspective. *Foundations of Computational Mathematics* (2022), 1–35. https://doi.org/10.1007/s10208-022-09581-9

[59] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3792772

[60] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8780–8789. https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html

[61] Shizhou Xu and Thomas Strohmer. 2023. Fair Data Representation for Machine Learning at the Pareto Frontier. *J. Mach. Learn. Res.* 24 (2023), 331:1–331:63. http://jmlr.org/papers/v24/22-0005.html

[62] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. 2021. Causal Intersectionality and Fair Ranking. In *2nd Symposium on Foundations of Responsible Computing (FORC 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 192)*, Katrina Ligett and Swati Gupta (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 7:1–7:20. https://doi.org/10.4230/LIPIcs.FORC.2021.7

[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1171–1180. https://doi.org/10.1145/3038912.3052660

[64] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.* 20 (2019), 75:1–75:42. https://jmlr.org/papers/v20/18-262.html

[65] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013 (JMLR Workshop and Conference Proceedings, Vol. 28)*. JMLR.org, 325–333. http://proceedings.mlr.press/v28/zemel13.html

[66] Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. http://arxiv.org/abs/1505.05723

A Simulation Study

A.1 Simulation Setup

We simulate the causal relationships of the fictitious credit application example depicted in Figure 1, using the R package `simcausal` [57]. In the FiND world, we eliminate the PA's effect by setting amount X_A, debt X_D, and the target Y of $A = a$ to their corresponding values among the $A = a'$ distributions X_A^*, X_D^* and Y^*:

Real world:

$$\begin{aligned} A &\sim \mathrm{B}(\pi_A) \\ X_C &\sim \mathrm{Ga}(\alpha_C, \beta_C) \\ X_A | X_C, A &\sim \mathrm{Ga}(\alpha_A(X_C, A), \beta_A(X_C, A)) \\ X_D | X_C, A &\sim \mathrm{B}(\pi_D(X_C, A)) \\ Y | X_A, X_D, X_C, A &\sim \mathrm{B}(\pi_Y(X_A, X_D, X_C, A)) \end{aligned}$$

FiND world:

$$\begin{aligned} A &\sim \mathrm{B}(\pi_A) \\ X_C &\sim \mathrm{Ga}(\alpha_C, \beta_C) \\ X_A^* | X_C &\sim \mathrm{Ga}(\alpha_{Am}(X_C, a'), \beta_A(X_C, a')) \\ X_D^* | X_C &\sim \mathrm{B}(\pi_D(X_C, a')) \\ Y^* | X_A^*, X_D^*, X_C &\sim \mathrm{B}(\pi_Y(X_A^*, X_D^*, X_C, a')) \end{aligned}$$

In both worlds, the PA A is generated by a Bernoulli distribution with success probability $\pi_A = 0.5$, while the confounder is Gamma distributed with $\alpha_C = 3.26$ and $\beta_C = 10.91$. For α_A and β_A, we take linear combinations of the features in combination with a log link, and for π_D and π_Y a logit link. We simulate datasets of size $N = 10,000$ for each world, where we use the same seed for both worlds to assure comparability and perform 25 iterations.

A.2 Approximating the FiND world

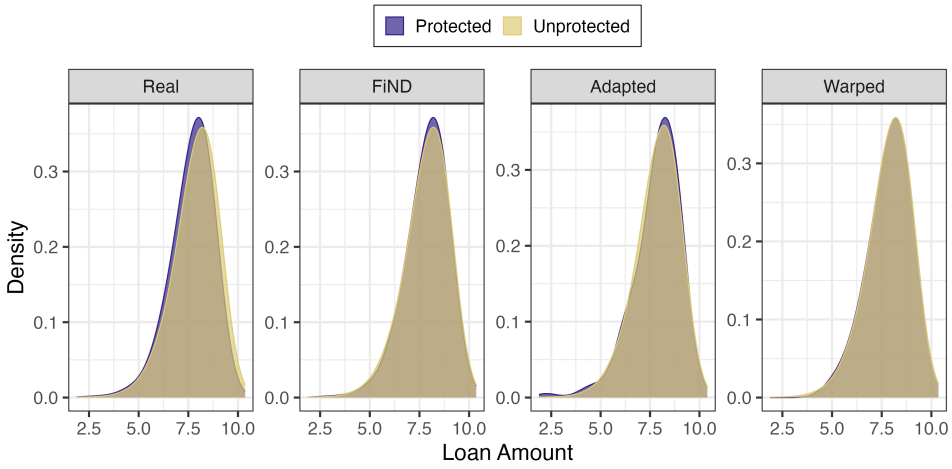


Fig. 5. Distribution of X_A in simulated real, FiND, adapted and warped world per protected a and unprotected group a'

Table 2. Distribution of Y and X_D in simulated real, FiND, adapted and warped world per protected a and unprotected group a'

	Real	**FiND**	**Adapted**	**Warped**
$P(Y=1)$	$a=0.755$ $a'=0.569$	$a=0.569$ $a'=0.569$	$a=0.585$ $a'=0.569$	$a=0.569$ $a'=0.569$
$P(X_D=1)$	$a=0.732$ $a'=0.927$	$a=0.928$ $a'=0.927$	$a=0.922$ $a'=0.927$	$a=0.927$ $a'=0.927$

Table 3. Fairness and performance metrics in real, FiND, adapted, and warped world for the simulation study in Section 4.2 alongside their 95% confidence intervals. All predictors are trained and evaluated using data from the same world, e.g., trained and evaluated on real world, trained and evaluated on FiND world, etc.

World	Fairness				Performance
	DP	FPR	FNR	PPV	AUC
Real	$0.825_{[0.792,0.862]}$	$0.782_{[0.750,0.836]}$	$0.954_{[0.926,0.983]}$	$0.986_{[0.955,0.998]}$	$0.887_{[0.895,0.899]}$
FiND	$0.987_{[0.964,0.999]}$	$0.989_{[0.963,1.000]}$	$0.991_{[0.976,0.999]}$	$0.988_{[0.957,0.999]}$	$0.897_{[0.895,0.899]}$
Adapted	$0.982_{[0.959,0.997]}$	$0.972_{[0.942,0.995]}$	$0.984_{[0.954,0.997]}$	$0.975_{[0.952,0.999]}$	$0.886_{[0.883,0.889]}$
Warped	$0.982_{[0.959,0.996]}$	$0.964_{[0.938,0.992]}$	$0.974_{[0.943,0.998]}$	$0.971_{[0.943,0.996]}$	$0.893_{[0.888,0.899]}$

B HMDA Experiment

B.1 Data Setup

We encode and filter the 2022 Home Mortgage Disclosure Act (HMDA) data of the state Wisconsin in the following way[9]:

- Y: Binary target indicating loan approved (1) or not approved (0). The original variable "action taken" has eight categories and encodes the status of the loan.
- A: Binary PA race with levels a Black applicant or a' non-Hispanic White applicant.
- X_A: Numerical variable of the amount of the covered loan, log-transformed.
- X_P: Binary variable indicating the purpose of the loan, (1) home purchase or not (0). The original variable has four categories.
- X_D: The debt to income ratio, with binary category (1) high ratio or not (0).
- X_C: The joint confounders age and gender. Binary age indicates (1) age above 62 or not (0). Binary gender indicates (1) female or not (0). (Note that gender is assumed to be binary purely for simplicity reasons and does not reflect the authors' personal view.)

B.2 Approximating the FiND world

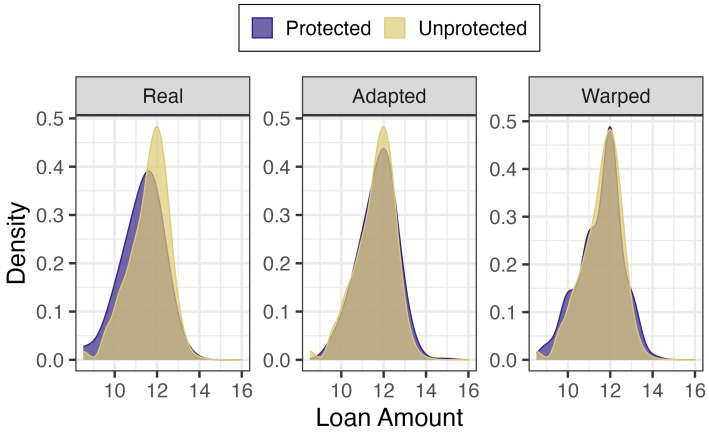


Fig. 6. Distribution of X_A on real-world, adapted and warped HMDA data per protected a and unprotected group a'

[9]A detailed description of all variables is provided here: https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields

Table 4. Distribution of Y, X_P and X_D on real-world, adapted and warped HMDA data per protected a and unprotected group a'

	Real	**Adapted**	**Warped**
$P(Y=1)$	$a=0.486$ $a'=0.679$	$a=0.677$ $a'=0.679$	$a=0.679$ $a'=0.679$
$P(X_P=1)$	$a=0.347$ $a'=0.390$	$a=0.398$ $a'=0.390$	$a=0.384$ $a'=0.390$
$P(X_D=1)$	$a=0.298$ $a'=0.371$	$a=0.335$ $a'=0.371$	$a=0.371$ $a'=0.371$