On the Fairness of Causal Algorithmic Recourse

Julius von Kügelgen[1,2] **Amir-Hossein Karimi**[1,3] **Umang Bhatt**[2]

Isabel Valera[4] **Adrian Weller**[2,5] **Bernhard Schölkopf**[1]

[1] Max Planck Institute for Intelligent Systems Tübingen [2] University of Cambridge
[3] ETH Zürich [4] Saarland University [5] The Alan Turing Institute

Abstract

Algorithmic fairness is typically studied from the perspective of *predictions*. Instead, here we investigate fairness from the perspective of *recourse* actions suggested to individuals to remedy an unfavourable classification. We propose two new fairness criteria at the group and individual level, which—unlike prior work on equalising the average group-wise distance from the decision boundary—explicitly account for causal relationships between features, thereby capturing downstream effects of recourse actions performed in the physical world. We explore how our criteria relate to others, such as counterfactual fairness, and show that fairness of recourse is complementary to fairness of prediction. We study theoretically and empirically how to enforce fair causal recourse by altering the classifier and perform a case study on the Adult dataset. Finally, we discuss whether fairness violations in the data generating process revealed by our criteria may be better addressed by societal interventions as opposed to constraints on the classifier.

1 Introduction

Algorithmic fairness is concerned with uncovering and correcting for potentially discriminatory behavior of automated decision making systems [4, 5, 9, 41]. Given a dataset comprising individuals from multiple legally protected groups (defined, for example, based on age, sex, or ethnicity), and a binary classifier trained to predict a decision (e.g., whether they were approved for a credit card), most approaches to algorithmic fairness seek to quantify the level of unfairness according to some predefined (statistical or causal) criteria, and then aim to correct it by altering the classifier. This notion of *predictive fairness* typically considers the *dataset as fixed*, and thus the *individuals as unalterable*. *Algorithmic recourse*, on the other hand, is concerned with offering recommendations to individuals who were unfavourably treated by a decision-making system in order to overcome their adverse situation [12, 14–16, 33, 35, 36]. For a given classifier and a negatively-classified individual, algorithmic recourse aims to identify which changes the individual could perform to flip the decision. Contrary to predictive fairness, recourse thus considers the *classifier as fixed* but *ascribes agency to the individual*.

Within machine learning (ML), fairness and recourse have mostly been considered in isolation and viewed as separate problems. While recourse has been investigated in the presence of protected attributes—e.g., by comparing recourse actions suggested to otherwise similar male and female individuals [35], or comparing the aggregated cost of recourse across different protected groups [33]—its relation to fairness has only been studied informally, in the sense that differences in recourse have typically been understood as *proxies of predictive unfairness* [13, 33, 35]. However, as we argue in the present work, recourse actually constitutes an interesting fairness criterion *in its own right* as it allows for the notions of agency and effort to be integrated into the study of fairness.

In fact, *discriminatory recourse does not imply predictive unfairness* (and is not implied by it either[1]). To see this, consider the data shown in Fig. 1. Suppose the feature X represents the (centered) income of an individual from one of two sub-groups $A \in \{0, 1\}$, distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$, i.e., only the variances differ. Now consider a binary classifier $h(X) = \text{sign}(X)$ which perfectly predicts whether the individual is approved for a credit card (the true label Y) [2]. While this scenario satisfies

[1] Clearly, the *average cost of recourse* across groups can be the *same*, even if the *proportion* of individuals which are classified as positive or negative is very *different* across groups

Preprint. Under review.

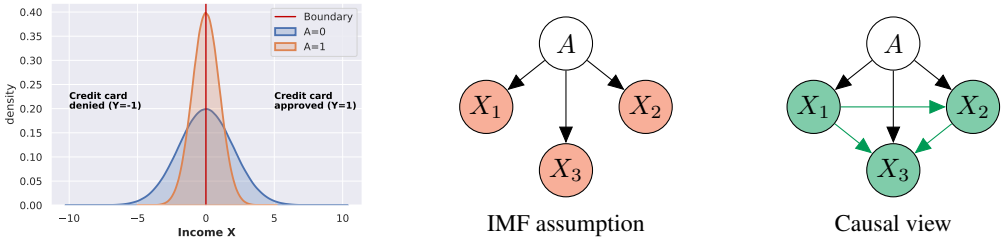


Figure 1: *(Left)* Example demonstrating the difference between fair *prediction* and fair *recourse*: here, only the variance of (centered income) X differs across two protected groups $A \in \{0, 1\}$, while the true and predicted label (whether an individual is approved for a credit card) are determined by $\text{sign}(X)$. This scenario would be considered fair from the perspective of *prediction*, but the cost of *recourse* (here, distance to the decision boundary at $X = 0$) is much larger for individuals in the blue group with $A = 0$. *(Center)* The framework underlying counterfactual explanations and distance-based recourse treats X_i as independently manipulable features (IMF). In a fairness context, this means that the X_i may depend on the protected attribute A (and potentially other unobserved factors) but do not causally influence each other. *(Right)* The present work considers a generalisation the IMF assumption by allowing for causal influences between the X_i, thus modeling the downstream effects of changing some features on others. This causal approach allows us to more accurately quantify recourse unfairness in real-world settings where the IMF assumption is typically violated, and provides a framework for studying alternative routes to achieving fair recourse than changing the classifier.

several *predictive fairness* criteria (e.g., demographic parity, equalised odds, calibration), the required increase in income for negatively-classified individuals to be approved for a credit card (i.e., the effort required to achieve recourse) is much larger for the higher variance group. If individuals from one protected group need to work harder than "similar" ones from another group to achieve the same goal, this violates the concept of equal opportunity, a notion aiming for people to operate on a level playing field [1].[2] However, this type of unfairness is not captured by predictive notions which—in only distinguishing between (unalterable) worthy or unworthy individuals—do not consider the possibility for individuals to deliberately improve their situation by means of changes or interventions.

In this vein, Gupta et al. [8] recently introduced Equalizing Recourse, the first recourse-based and prediction-independent notion of fairness in ML. They propose to measure recourse fairness in terms of the *average group-wise distance to the decision boundary* for those getting a bad outcome, and show that this can be calibrated during classifier training. However, this formulation ignores that *recourse is fundamentally a causal problem* since actions performed by individuals in the real-world to change their situation may have downstream effects [14–16, 23, 24], c.f. also [2, 35, 37]. By not reasoning about causal relations between features, the distance-based approach [8] (i) does not accurately reflect the true (differences in) recourse cost, and (ii) is restricted to the classical prediction-centered approach of changing the classifier to address discriminatory recourse.

In the present work, we address both of these limitations. First, by extending the idea of Equalizing Recourse [8] to the minimal intervention-based framework of recourse [16], we introduce *causal* notions of fair recourse which capture the true differences in recourse cost more faithfully if features are not manipulable independently of each other, as is generally the case. Second, we argue that a causal model of the data generating process opens up a new route to fairness via *societal interventions* in the form of changes to the underlying system. Such societal interventions may reflect common policies like subgroup-specific subsidies or tax breaks. We make the following contributions:

- we introduce a *causal* version (Defn. 3.1) of Equalizing Recourse, as well as a stronger (Prop. 3.3) *individual*-level criterion for fair causal recourse (Defn. 3.2) which we argue is more appropriate;
- we provide the first *formal* study of the relation between fair prediction and fair recourse, and show that they are complementary notions of fairness which do not imply each other (Prop. 3.4);
- we establish sufficient conditions that allow for individually-fair causal recourse (Prop. 3.6);
- in evaluating different fair recourse metrics for several trained classifiers (§ 4), we empirically verify our main results and demonstrate that non-causal metrics misrepresent recourse unfairness;
- we propose societal interventions as an alternative to altering a classifier to address unfairness (§ 5).

2 Preliminaries: explainable AI, recourse, causality, fairness

Notation. Let $\mathbf{X} = (X_1, ..., X_n) \in \mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_n \subseteq \mathbb{R}^n$ denote observed (non-protected) features, $A \in \mathcal{A} = \{1, \dots, K\}$ a protected attribute indicating which group each individual belongs to (based, e.g., on her age, sex, etc), and $h : \mathcal{X} \to \mathcal{Y}$ a given binary classifier with $Y \in \mathcal{Y} = \{\pm 1\}$

[2] This differs from the purely predictive, statistical criterion of equal opportunity commonly-used in ML [9].

denoting the ground truth label (e.g., whether her credit card was approved). We observe a dataset $\mathcal{D} = \{\mathbf{v}^i\}_{i=1}^N$ of i.i.d. observations of the random variables $(\mathbf{X}, A)$ where $\mathbf{v}^i = (\mathbf{x}^i, a^i)$.[3]

Counterfactual explanations. A common framework for explaining decisions made by (black-box) ML models is that of nearest counterfactual explanations (CEs) [37]. A CE is a closest feature vector on the other side of the decision boundary. Given a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, a CE for an individual $\mathbf{x}^\text{F}$ who obtained an unfavourable prediction, $h(\mathbf{x}^\text{F}) = -1$, is defined as a solution to:

$$\min_{\mathbf{x} \in \mathcal{X}} \; d(\mathbf{x}, \mathbf{x}^\text{F}) \qquad \text{subject to} \qquad h(\mathbf{x}) = 1. \tag{1}$$

While CEs are useful to *understand the behaviour of a classifier*, they do not generally lead to *actionable recommendations*: they inform an individual of where she should be to obtain a more favourable prediction, but they may not suggest *feasible* changes she could perform to get there.

Recourse with independently-manipulable features. Ustun et al. [35] refer to a person's ability to change the decision of a model by altering actionable variables as *recourse* and propose to solve

$$\min_{\boldsymbol{\delta} \in \mathcal{F}(\mathbf{x}^\text{F})} \; c(\boldsymbol{\delta}; \mathbf{x}^\text{F}) \quad \text{subject to} \quad h(\mathbf{x}^\text{F} + \boldsymbol{\delta}) = 1 \tag{2}$$

where $\mathcal{F}(\mathbf{x}^\text{F})$ is a set of feasible change vectors and $c(\cdot; \mathbf{x}^\text{F})$ is a cost function defined over these actions, both of which may depend on the individual. As pointed out by Karimi et al. [16], the framework in (2) treats features as manipulable independently of each other (see Fig. 1, *Center*) and does not consider causal relations that may exist between them (see Fig. 1, *Right*): it considers feasibility constraints on actions, but assumes that variables which are not acted-upon ($\delta_i = 0$) remain unchanged. We refer to this assumption as the *independently-manipulable features* (IMF) assumption. While this IMF-view may be appropriate if we only analyse the behaviour of the classifier itself, it falls short of capturing effects of interventions performed in the real world, as is the case in actionable recourse.[4] As a consequence, the IMF-recourse approach of (2) only guarantees recourse if the acted-upon variables have no causal effect on the remaining variables [16].

Structural causal models. A structural causal model (SCM) [27, 29] over a set of observed variables $\mathbf{V} = \{V_i\}_{i=1}^n$ is a pair $\mathcal{M} = (\mathbf{S}, \mathbb{P}_\mathbf{U})$, where the structural equations $\mathbf{S}$ are a set of assignments $\mathbf{S} = \{V_i := f_i(\text{PA}_i, U_i)\}_{i=1}^n$, which compute each V_i as a deterministic function f_i of its direct causes (causal parents) $\text{PA}_i \subseteq \mathbf{V} \setminus V_i$ and an unobserved variable U_i. The distribution $\mathbb{P}_\mathbf{U}$ factorises over the latent $\mathbf{U} = \{U_i\}_{i=1}^n$, incorporating the assumption that there is no unobserved confounding. If the causal graph $\mathcal{G}$ (obtained by drawing a directed edge from each variable in PA_i to V_i) is acyclic, $\mathcal{M}$ induces a unique "observational" distribution over $\mathbf{V}$, defined as the push forward of $\mathbb{P}_\mathbf{U}$ via $\mathbf{S}$. Moreover, SCMs can be used to model the effect of *interventions*: external manipulations to the system that change the generative process (i.e., the structural assignments) of a subset of variables $\mathbf{V}_\mathcal{I} \subseteq \mathbf{V}$, e.g., by fixing their value to a constant $\boldsymbol{\theta}_\mathcal{I}$. Such interventions are denoted using Pearl's do-operator by $do(\mathbf{V}_\mathcal{I} := \boldsymbol{\theta}_\mathcal{I})$, or $do(\boldsymbol{\theta}_\mathcal{I})$ for short. Interventional distributions are obtained from $\mathcal{M}$ by replacing the structural equations for $\mathbf{V}_\mathcal{I}$ by their new assignments to obtain $\mathbf{S}^{do(\boldsymbol{\theta}_\mathcal{I})}$ and then computing the distribution entailed by $\mathcal{M}^{do(\boldsymbol{\theta}_\mathcal{I})} = (\mathbf{S}^{do(\boldsymbol{\theta}_\mathcal{I})}, \mathbb{P}_\mathbf{U})$. Similarly, SCMs allow reasoning about (structural) *counterfactuals*: statements about interventions performed in a hypothetical world where all unobserved noise terms $\mathbf{U}$ are unchanged. The counterfactual distribution for a hypothetical intervention $do(\boldsymbol{\theta}_\mathcal{I})$ given a factual observation $\mathbf{v}^\text{F}$, denoted $\mathbf{v}_{\boldsymbol{\theta}_\mathcal{I}}(\mathbf{u}^\text{F})$, is obtained from $\mathcal{M}$ by first inferring the posterior distribution over the unobserved variables $\mathbb{P}_{\mathbf{U}|\mathbf{v}^\text{F}}$ (*abduction*) and then proceeding as in the interventional case, i.e., it is induced by $\mathcal{M}^{do(\boldsymbol{\theta}_\mathcal{I})|\mathbf{v}^\text{F}} = (\mathbf{S}^{do(\boldsymbol{\theta}_\mathcal{I})}, \mathbb{P}_{\mathbf{U}|\mathbf{v}^\text{F}})$.

Causal recourse. To capture causal relations between features, Karimi et al. [16] propose to approach the actionable recourse task within the framework of SCMs and to shift the focus from nearest CEs to minimal interventions, leading to the optimisation problem

$$\min_{\boldsymbol{\theta}_\mathcal{I} \in \mathcal{F}(\mathbf{x}^\text{F})} \; c(\boldsymbol{\theta}_\mathcal{I}; \mathbf{x}^\text{F}) \quad \text{subj. to} \quad h(\mathbf{x}_{\boldsymbol{\theta}_\mathcal{I}}(\mathbf{u}^\text{F})) = 1, \tag{3}$$

where $\mathbf{x}_{\boldsymbol{\theta}_\mathcal{I}}(\mathbf{u}^\text{F})$ denotes the "counterfactual twin" of $\mathbf{x}^\text{F}$ had $\mathbf{X}_\mathcal{I}$ been $\boldsymbol{\theta}_\mathcal{I}$. In practice, the SCM is unknown and needs to be inferred from data based on additional (domain-specific) assumptions, leading to probabilistic versions of (3) which aim to find actions that achieve recourse with high probability [15]. If the IMF assumptions holds, i.e., if the set of descendants of all actionable variables is empty, (3) reduces to the distance-based IMF approach to recourse [35] from (2) as a special case.

[3] We use $\mathbf{v}$ when there is an explicit distinction between the protected attribute and other features (in the context of fairness) and $\mathbf{x}$ otherwise (in the context of explainability).

[4] E.g., an increase in income will likely also positively affect the individual's savings balance.

Algorithmic and counterfactual fairness. As ML is increasingly used in consequential decision making, many recent works study the problem of algorithmic fairness, i.e., whether model predictions lead to potential discrimination against protected groups. While there are many different statistical notions of fairness [5, 9, 39–41], these are sometimes mutually incompatible [4] and it has been argued that discrimination, at its heart, corresponds to a (direct or indirect) causal influence of a protected attribute on the prediction, thus making fairness a fundamentally causal problem [3, 17, 18, 22, 25, 26, 31, 38, 42, 43]. Of particular interest to our work is the notion of *counterfactual fairness* introduced by Kusner et al. [20] which calls a (probabilistic) classifier h over $\mathbf{V} = \mathbf{X} \cup A$ counterfactually fair if it satisfies $h(\mathbf{v}^{\mathrm{F}}) = h(\mathbf{v}_a(\mathbf{u}^{\mathrm{F}})), \forall a \in \mathcal{A}, \mathbf{v}^{\mathrm{F}} = (\mathbf{x}^{\mathrm{F}}, a^{\mathrm{F}}) \in \mathcal{X} \times \mathcal{A}$, where $\mathbf{v}_a(\mathbf{u}^{\mathrm{F}})$ denotes the "counterfactual twin" of $\mathbf{v}^{\mathrm{F}}$ had the attribute been a instead of a^{F}.

Equalizing recourse across groups. The main focus of this paper is the *fairness of recourse actions* which, to the best of our knowledge, was studied for the first time by Gupta et al. [8]. They advocate for equalizing the average cost of recourse across protected groups and to incorporate this as a constraint when training a classifier. Taking a distance-based approach in line with the view of CEs, they define the cost of recourse for individual $\mathbf{x}^{\mathrm{F}}$ with $h(\mathbf{x}^{\mathrm{F}}) = -1$ as the minimum achieved in (1):

$$r^{\mathrm{IMF}}(\mathbf{x}^{\mathrm{F}}) = \min_{\mathbf{x} \in \mathcal{X}} d(\mathbf{x}^{\mathrm{F}}, \mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 1, \tag{4}$$

which is equivalent to IMF-recourse (2) [35] if $c(\boldsymbol{\delta}; \mathbf{x}^{\mathrm{F}}) = d(\mathbf{x}^{\mathrm{F}} + \boldsymbol{\delta}, \mathbf{x}^{\mathrm{F}})$ is chosen as cost function. Defining the protected subgroups, $G_a = \{\mathbf{v}^i \in \mathcal{D} : a^i = a\}$, and $G_a^- = \{\mathbf{v} \in G_a : h(\mathbf{v}) = -1\}$, the group-level cost of recourse (here, the average distance to the decision boundary) is then given by,

$$r^{\mathrm{IMF}}(G_a^-) = \tfrac{1}{|G_a^-|} \textstyle\sum_{\mathbf{v}^i \in G_a^-} r^{\mathrm{IMF}}(\mathbf{x}^i). \tag{5}$$

The idea of *Equalizing Recourse* across groups [8] can then be summarised as follows.

Definition 2.1 (Group-level fair IMF-recourse, from [8])**.** The group-level unfairness of *recourse with independently-manipulable features* (IMF) for a dataset $\mathcal{D}$, classifier h, and distance metric d is:

$$\Delta_{\mathrm{dist}}(\mathcal{D}, h, d) := \max_{a,a' \in \mathcal{A}} \left| r^{\mathrm{IMF}}(G_a^-) - r^{\mathrm{IMF}}(G_{a'}^-) \right|.$$

We say recourse for $(\mathcal{D}, h, d)$ is "group IMF-fair" if $\Delta_{\mathrm{dist}}(\mathcal{D}, h, d) = 0$.

3 Fair causal recourse

Since Defn. 2.1 rests on the IMF assumption, it ignores causal relationships between variables, fails to account for downstream effects of actions on other relevant features, and thus generally incorrectly estimates the true cost of recourse [15, 16]. We argue that recourse-based fairness considerations should rest on a causal model that captures the effect of interventions performed in the physical world where features are often causally related to each other. We therefore consider an SCM $\mathcal{M}$ over $\mathbf{V} = (\mathbf{X}, A)$ to model causal relationships between the protected attribute and the remaining features.

3.1 Group-level fair causal recourse

Defn. 2.1 can be adapted to the causal (CAU) recourse framework (3) by replacing the minimum distance in (4) with the cost of recourse within a causal model, i.e., the minimum achieved in (3):

$$r^{\mathrm{CAU}}(\mathbf{v}^{\mathrm{F}}) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{v}^{\mathrm{F}})} c(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\mathrm{F}}) \quad \text{subj. to} \quad h(\mathbf{v}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\mathrm{F}})) = 1,$$

where we recall that the constraint $h(\mathbf{v}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\mathrm{F}})) = 1$ ensures that the counterfactual twin of $\mathbf{v}^{\mathrm{F}}$ in $\mathcal{M}$ falls on the favourable side of the classifier. Let $r^{\mathrm{CAU}}(G_a^-)$ be the average of $r^{\mathrm{CAU}}(\mathbf{v}^{\mathrm{F}})$ across G_a^-, analogously to (5). We can then define group-level fair causal recourse as follows.

Definition 3.1 (Group-level fair causal recourse)**.** The group-level unfairness of *causal* (CAU) recourse for a dataset $\mathcal{D}$, classifier h, and cost function c *w.r.t. an SCM* $\mathcal{M}$ is given by:

$$\Delta_{\mathrm{cost}}(\mathcal{D}, h, c, \mathcal{M}) := \max_{a,a' \in \mathcal{A}} \left| r^{\mathrm{CAU}}(G_a^-) - r^{\mathrm{CAU}}(G_{a'}^-) \right|.$$

We say that recourse for $(\mathcal{D}, h, c, \mathcal{M})$ is "group CAU-fair" if $\Delta_{\mathrm{cost}}(\mathcal{D}, h, c, \mathcal{M}) = 0$.

While Defn. 2.1 is agnostic to the (causal) generative process of the data (note the absence of a reference SCM $\mathcal{M}$ from Defn. 2.1), Defn. 3.1 takes causal relationships between features into account when calculating the cost of recourse. It thus captures the effect of actions and the necessary cost of recourse more faithfully when the IMF-assumption is violated, as is realistic for most applications.

A shortcoming of both Defns. 2.1 and 3.1 is that they are group-level definitions, i.e., they only consider the *average* cost of recourse across all individuals sharing the same protected attribute.

However, it has been argued from causal [3, 20, 42, 43] and non-causal [5] perspectives that fairness is fundamentally an individual-level concept:[5] group-level fairness still allows for unfairness at the level of the individual, provided that positive and negative discrimination cancel out across the group. This is one motivation behind counterfactual fairness [20]: a decision is considered fair at the individual level if it would not have changed, had the individual belonged to a different protected group.

3.2 Individually fair causal recourse

Inspired by counterfactual fairness [20], we propose that (causal) recourse may be considered fair at the level of the individual if the cost of recourse would have been the same had the individual belonged to a different protected group, i.e., under a counterfactual change to A.

Definition 3.2 (Individually fair causal recourse)**.** The individual-level unfairness of *causal* recourse for a dataset $\mathcal{D}$, classifier h, and cost function c *w.r.t. an SCM* $\mathcal{M}$ is

$$\Delta_{\text{ind}}(\mathcal{D}, h, c, \mathcal{M}) := \max_{a \in \mathcal{A}; \mathbf{v}^{\text{F}} \in \mathcal{D}} |r^{\text{CAU}}(\mathbf{v}^{\text{F}}) - r^{\text{CAU}}(\mathbf{v}_a(\mathbf{u}^{\text{F}}))|$$

We say that recourse for $(\mathcal{D}, h, c, \mathcal{M})$ is "individually CAU-fair" if $\Delta_{\text{ind}}(\mathcal{D}, h, c, \mathcal{M}) = 0$.

This is a stronger notion, in the sense that it is possible to satisfy both group IMF-fair (Defn. 2.1) and group CAU-fair recourse (Defn. 3.1), without satisfying individually CAU-fair recourse.

Proposition 3.3. *Neither of the group-level notions of fair recourse (Defn. 2.1 and Defn. 3.1) are sufficient conditions for individually CAU-fair recourse (Defn. 3.2), i.e.,*

Group IMF-fair $\not\Longrightarrow$ *Individually CAU-fair,* $\quad \wedge \quad$ *Group CAU-fair* $\not\Longrightarrow$ *Individually CAU-fair.*

Proof. A counterexample is given by the following combination of SCM and classifier

$$A := U_A, \quad X := AU_X + (1 - A)(1 - U_X), \quad U_A, U_X \sim \text{Bern}(0.5), \quad h(X) = \text{sign}(X - 0.5).$$

with $Y := h(X)$. We have $\mathbb{P}_{X|A=0} = \mathbb{P}_{X|A=1} = \text{Bern}(0.5)$, so the distance to the boundary at $X = 0.5$ is the same across groups and Defn. 2.1 is satisfied. Since there is only one actionable feature X without descendants, so is Defn. 3.1. However, for all $\mathbf{v}^{\text{F}} = (x^{\text{F}}, a^{\text{F}})$ and any $a \neq a^{\text{F}}$, we have $h(x^{\text{F}}) \neq h(x_a(u_X^{\text{F}})) = 1 - h(x^{\text{F}})$, so it is maximally unfair at the individual level: the cost of recourse would have been zero had the protected attribute been different, as the prediction would have flipped.

3.3 Relation to counterfactual fairness

Note that h in the proof of Prop. 3.3 is *not* counterfactually fair. This suggests to investigate their relation more closely: *does a counterfactually fair classifier imply fair (causal) recourse?*

Proposition 3.4. *Counterfactual fairness is insufficient for any of the three notions of fair recourse:*

h *counterfactually fair* $\not\Longrightarrow Q, \quad \forall Q \in \{$*Group IMF-fair, Group CAU-fair, Individually CAU-fair*$\}$

Proof. A counterexample is given by the following combination of SCM and classifier:

$$A := U_A, \qquad X := (2 - A)U_X, \qquad U_A \sim \text{Bernoulli}(0.5), \qquad U_X \sim \mathcal{N}(0, 1), \tag{6}$$

with $Y := h(X) = \text{sign}(X)$, which we used to generate Fig. 1 *(left)*. As $\text{sign}(X) = \text{sign}(U_X)$, and U_X is assumed fixed when reasoning about a counterfactual change of A, h is counterfactually fair. However, $\mathbb{P}_{X|A=0} = \mathcal{N}(0, 4)$ and $\mathbb{P}_{X|A=1} = \mathcal{N}(0, 1)$, so the distance to the boundary differs at the group level. Moreover, X either doubles or halves when counterfactually changing A.

Remark 3.5. An important characteristic of the above counterexample is that h is *deterministic*, which makes it possible that h is counterfactually fair, even though it depends on a descendant of A. This would generally not be the case if h were *probabilistic* (e.g., a logistic regression), $h : \mathcal{X} \to [0, 1]$, so that the probability of a positive classification decreases with the distance from the decision boundary.

3.4 Achieving fair causal recourse algorithmically at the classifier level

Constrained optimisation. A first approach is to explicitly take constraints on the (group or individual level) fairness of causal recourse into account when training a classifier, as implemented for non-causal recourse under the IMF assumption by Gupta et al. [8]. Herein we can control the potential trade-off between accuracy and fairness with a hyperparameter. However, the optimisation problem in (3) involves optimising over the combinatorial space of intervention targets $\mathcal{I} \subseteq \{1, ..., n\}$, so it is unclear whether fairness of causal recourse may easily be included as a differentiable constraint.

[5] after all, it is not much consolation for an individual who was unfairly given an unfavourable prediction to find out that other members of the same group were treated more favourably

Restricting the classifier inputs. An approach that only requires *qualitative* knowledge in form of the causal graph (but not a fully-specified SCM), is to restrict the set of input features to the classifier to only contain non-descendants of the protected attribute. In this case, and subject to some additional assumptions stated in more detail below, individually fair causal recourse can be guaranteed.[6]

Proposition 3.6. *Assume h only depends on a subset $\tilde{\mathbf{X}}$ which are non-descendants of A in $\mathcal{M}$, $\tilde{\mathbf{X}} \subseteq \mathbf{V} \setminus (A \cup \mathrm{desc}(A))$; and that the set of feasible actions and their cost remain the same under a counterfactual change of A, $\mathcal{F}(\mathbf{v}^F) = \mathcal{F}(\mathbf{v}_a(\mathbf{u}^F))$ and $c(\cdot\,; \mathbf{v}^F) = c(\cdot\,; \mathbf{v}_a(\mathbf{u}^F))\ \forall a \in \mathcal{A}, \mathbf{v}^F \in \mathcal{D}$. Then $(\mathcal{D}, h, c, \mathcal{M})$ is "individually CAU-fair".*

The proof is provided in Appendix A. The assumption of Prop. 3.6 that both the set of feasible actions $\mathcal{F}(\mathbf{v}^F)$ and the cost function $c(\cdot\,; \mathbf{v}^F)$ remain the same under a counterfactual change to the protected attribute may not always hold. For example, if a protected group were precluded (by law) or discouraged from performing certain recourse actions such as taking on a particular job or applying for a certification, that would constitute such a violation due to a separate source of discrimination. Moreover, since protected attributes usually represent socio-demographic features (e.g., age, gender, ethnicity, etc), they often appear as root nodes in the causal graph and have downstream effects on numerous other features. Forcing the classifier to only consider non-descendants of A as inputs, as in Prop. 3.6, can therefore lead to a drop in accuracy which can be a restriction in practice.

Abduction / representation learning. We have shown that considering only non-descendants of A is a way to achieve individually CAU-fair recourse. In particular, this also applies to the unobserved variables $\mathbf{U}$ which are, by definition, not descendants of any observed variables. This suggests to use U_i in place of any descendants X_i of A when training the classifier—in a way, U_i can be seen as a "fair representation" of X_i since it is an exogenous component that is not due to A. However, as $\mathbf{U}$ is unobserved, it needs to be inferred from the observed $\mathbf{v}^F$, corresponding to the abduction step of counterfactual reasoning. Great care needs to be taken in learning such a representation in terms of the (fair) background variables as (untestable) counterfactual assumptions are required [20, § 4.1].

4 Experiments

We perform two sets of experiments. First, we verify our main claims in numerical simulations (§ 4.1). Second, we use our causal measures of fair recourse to conduct a preliminary case study on the Adult dataset (§ 4.2). We refer to Appendix B for further experimental details and C for additional results.

4.1 Numerical simulations

Data. Since computing recourse actions in the general case requires knowledge (or estimation) of the true SCM, we first consider a controlled setting with two kinds of synthetically generated data:

- Independently-manipulable features (IMF): the setting underlying IMF recourse [35]; features do not causally influence each other, but may depend on the protected attribute A, c.f. Fig. 1 *(center)*.
- Causally-dependent features (CAU): features causally depend on each other and on A, c.f. Fig. 1 *(right)*. We use $\{X_i := f_i(A, \mathrm{PA}_i) + U_i\}_{i=1}^n$ with linear (CAU-LIN) and nonlinear (CAU-ANM) f_i.

We use $n = 3$ non-protected features X_i and a binary protected attribute $A \in \{0, 1\}$ in all our experiments and generate labelled datasets of $N = 500$ observations using the SCMs in B.1. The ground truth (GT) labels y^i used to train different classifiers are sampled as $Y^i \sim \mathrm{Bernoulli}(h(\mathbf{x}^i))$ where $h(\mathbf{x}^i)$ is a linear or nonlinear logistic regression, independently of A, as detailed in B.2.

Classifiers. On each data set, we train several ("fair") classifiers. We consider linear and nonlinear logistic regression (LR), and different support vector machines (SVMs) [32] (for ease of comparison with Gupta et al. [8]), trained on varying input sets:

- LR/SVM$(\mathbf{X}, A)$: trained on all features (*naïve baseline*);
- LR/SVM$(\mathbf{X})$: trained only on non-protected features $\mathbf{X}$ (*unaware baseline*);
- FairSVM$(\mathbf{X}, A)$: the method of Gupta et al. [8], designed to equalise the average distance to the decision boundary across different protected groups;
- LR/SVM$(\mathbf{X}_{\mathrm{nd}})$: trained only on features $\mathbf{X}_{\mathrm{nd}(A)}$ which are non-descendants of A, see § 3.4;
- LR/SVM$(\mathbf{X}_{\mathrm{nd}}, \mathbf{U}_{\mathrm{d}})$: trained on the non-descendants $\mathbf{X}_{\mathrm{nd}(A)}$ of A and on the unobserved variables $\mathbf{U}_{\mathrm{d}(A)}$ corresponding to features $\mathbf{X}_{\mathrm{d}(A)}$ which are descendants of A, see § 3.4.

To make distances comparable across classifiers, we use either a linear or polynomial kernel for all SVMs (depending on the GT labels) and select all remaining hyperparameters (including the trade-off parameter λ for FairSVM) using 5-fold cross validation. Results for kernel selection by

[6] Only using non-descendants of A is also sufficient for counterfactual fairness, see Lemma 1 of [20].

Table 1: Results for § 4.1; best performing method highlighted in **bold**. Consistent with § 3, there is no clear relation between $\Delta_{\textbf{dist}}$, $\Delta_{\textbf{cost}}$, and $\Delta_{\textbf{ind}}$, and blindness to the protected attribute alone does not improve any of the recourse fairness metrics (c.f. *naïve* and *unaware* baselines). FairSVM generally performs well on $\Delta_{\textbf{dist}}$ (which it has been trained for) at a cost to accuracy. Our causally-motivated setups, LR/SVM($\mathbf{X}_{\text{nd}}$) and LR/SVM($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$), achieve $\Delta_{\textbf{ind}} = 0$ throughout (c.f. Prop. 3.6), and they are the only methods to do so.

Classifier	**GT labels from *linear* log. reg. → using *linear* kernel / *linear* log. reg.**												**GT labels from *nonlinear* log. reg. → using *polynomial* kernel / *nonlinear* log. reg.**											
	IMF				**CAU-LIN**				**CAU-ANM**				**IMF**				**CAU-LIN**				**CAU-ANM**			
	Acc	Δ_{dist}	Δ_{cost}	Δ_{ind}	**Acc**	Δ_{dist}	Δ_{cost}	Δ_{ind}	**Acc**	Δ_{dist}	Δ_{cost}	Δ_{ind}	**Acc**	Δ_{dist}	Δ_{cost}	Δ_{ind}	**Acc**	Δ_{dist}	Δ_{cost}	Δ_{ind}	**Acc**	Δ_{dist}	Δ_{cost}	Δ_{ind}
SVM($\mathbf{X}, A$)	86.5	0.96	0.40	1.63	89.5	1.18	0.44	2.11	**88.2**	0.65	0.27	2.32	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.03	1.06	90.6	0.04	0.03	1.40
LR($\mathbf{X}, A$)	**86.7**	0.48	0.50	1.91	89.5	0.63	0.53	2.11	87.7	0.40	0.34	2.32	90.5	0.08	0.03	1.06	90.6	0.09	**0.01**	1.00	90.6	0.19	0.22	1.28
SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	89.4	1.61	0.61	2.11	88.0	0.56	0.29	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.08	1.09	**91.0**	**0.02**	0.03	1.64
LR($\mathbf{X}$)	86.6	0.47	0.53	1.80	89.5	0.64	0.57	2.11	87.7	0.41	0.43	2.79	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	90.9	0.08	0.06	1.66
FairSVM($\mathbf{X}, A$)	68.1	**0.04**	0.28	1.36	66.8	0.26	**0.12**	0.78	66.3	0.25	0.21	1.50	90.1	**0.02**	**0.00**	1.15	90.7	0.06	0.04	1.16	90.3	0.37	**0.02**	1.64
SVM($\mathbf{X}_{\text{nd}}$)	65.5	0.05	0.06	**0.00**	67.4	**0.15**	0.17	**0.00**	65.9	0.31	0.37	**0.00**	66.7	0.10	0.06	**0.00**	58.4	0.05	0.06	**0.00**	62.0	0.13	0.11	**0.00**
LR($\mathbf{X}_{\text{nd}}$)	65.3	0.05	**0.05**	**0.00**	67.3	0.18	0.18	**0.00**	65.6	0.31	0.31	**0.00**	64.7	**0.02**	0.04	**0.00**	58.4	**0.02**	0.02	**0.00**	61.1	**0.02**	0.03	**0.00**
SVM($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$)	86.5	0.96	0.58	**0.00**	**89.6**	1.07	0.70	**0.00**	88.0	**0.21**	**0.14**	**0.00**	90.7	**0.02**	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	90.1	0.15	0.12	**0.00**
LR($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$)	**86.7**	0.43	0.90	**0.00**	89.5	0.35	0.77	**0.00**	87.8	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**

cross-validation are also provided in Tab. 4 in C.3. Linear (nonlinear, resp.) LR is used when the GT labels are generated using linear (nonlinear, resp.) logistic regression, as detailed in B.2.

Solving the Causal Recourse Optimisation Problem. We treat A and all U_i as non-actionable and all X_i as actionable. For each negatively predicted individual, we discretise the space of feasible actions, compute the efficacy of each action using a learned approximate SCM ($\mathcal{M}_{\text{KR}}$) (following [15]; see C.2 for details), and select the least costly valid action resulting in a favorable outcome. Results using the true oracle SCM ($\mathcal{M}^{\star}$) and a linear estimate thereof ($\mathcal{M}_{\text{LIN}}$) are included in Tabs. 3 and 4 in Appendix C.2; the trends are largely the same as for $\mathcal{M}_{\text{KR}}$.

Metrics. We report (a) accuracy (**Acc**) on a held out test set of size 3000; and (b) fairness of recourse as measured by average distance to the boundary ($\Delta_{\textbf{dist}}$, Defn. 2.1) [8], and our causal group-level ($\Delta_{\textbf{cost}}$, Defn. 3.1) and individual level ($\Delta_{\textbf{ind}}$, Defn. 3.2) criteria. For (b), we select 50 negatively classified individuals from each protected group and report the difference in group-wise means ($\Delta_{\textbf{dist}}$ and $\Delta_{\textbf{cost}}$) or the maximum difference over all 100 individuals ($\Delta_{\textbf{ind}}$). To facilitate a comparison between the different SVMs, $\Delta_{\textbf{dist}}$ is reported in terms of absolute distance to the decision boundary in units of margins. As a cost function in the causal recourse optimisation problem, we use the L2 distance between the intervention value $\boldsymbol{\theta}_{\mathcal{I}}$ and the factual value of the intervention targets $\mathbf{x}^{\text{F}}_{\mathcal{I}}$.

Results. Results are shown in Tab. 1. We find that the *naïve* and *unaware* baselines LR/SVM($\mathbf{X}, A$) and LR/SVM($\mathbf{X}$) generally exhibit high accuracy and rather poor performance in terms of fairness metrics, though they achieve surprisingly low $\Delta_{\textbf{cost}}$ on some data sets. We observe no clear preference of one baseline over the other across datasets which is consistent with prior work showing that blindness to the protected attribute is not necessarily beneficial for fairness [5, 17]. The FairSVM generally performs well in terms of $\Delta_{\textbf{dist}}$ (which is what it is trained for), especially on the two IMF data sets, and sometimes (though not consistently) outperforms the baselines in terms of the causal fairness metrics. However, this comes at decreased accuracy, particularly on the linearly-separable data. Both of our causally-motivated setups, LR/SVM($\mathbf{X}_{\text{nd}(A)}$) and LR/SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$), achieve $\Delta_{\textbf{ind}} = 0$ throughout *as expected per* Prop. 3.6, and they are the only methods to do so. Whereas the former comes at a substantial drop in accuracy, as discussed in § 3.4, the latter maintains high accuracy by additionally relying on (the true) $\mathbf{U}_{\text{d}(A)}$ for prediction. Since these are not observed practice, its accuracy should therefore be understood as an upper bound on what is possible while preserving "individually CAU-fair" recourse if abduction is done correctly, c.f. § 3.4. Generally, we observe no clear relationship between the different fairness metrics. For example, low $\Delta_{\textbf{dist}}$ does not imply low $\Delta_{\textbf{cost}}$ (nor vice versa) justifying the need for taking causal relations between features into account (if present) to enforce fair recourse at the group-level. Likewise, *neither small* $\Delta_{\textbf{dist}}$ *nor small* $\Delta_{\textbf{cost}}$ *imply small* $\Delta_{\textbf{ind}}$*, consistent with* Prop. 3.3, and, empirically, the converse does not hold either.

4.2 Case study on the Adult dataset

Data. We use the Adult dataset [21], which consists of 45K+ samples without missing data. We process the dataset similarly to Chiappa [3] and Nabi and Shpitser [25] and adopt the causal graph assumed therein (c.f. Fig. 3c in Appendix B.1). A mix of eight heterogeneous variables are considered, including three binary protected attributes: sex, age (binarised as $\mathbb{I}\{\text{age} \geq 38\}$), and nationality (Nat, US vs non-US); and five non-protected features: marital status (MS, categorical), education level (Edu, integer), working class (WC, categorical), occupation (Occ, categorical), and hours per week (Hrs, integer). We consider marital status as non-intervenable when searching for recourse actions. All data loading and preprocessing scripts can be found in the amended implementation.

Experimental setup. We extend the probabilistic framework of Karimi et al. [15] to consider causal recourse in the presence of heterogeneous features, see C.2 for details. We use a nonlinear LR($\mathbf{X}$) as a classifier (78.4% **Acc**) and solve the recourse optimisation problem (3) as in § 4.1. We compute the

Table 2: Individual-level recourse discrimination on the Adult dataset (§ 4.2). Factual (F) observation highlighted in cyan, counterfactual (CF) twin with largest individual-level recourse difference in magenta. Consistent with the group-level trends, we observe quantitative discrimination across each protected attribute (favoring older age, male gender, and US nationalism), and qualitative differences in the suggested recourse actions across groups (e.g., favorable predictions based on higher education for men and more working hours for non-US nationals).

Type	Sex	Age	Nat	MS	Edu	WC	Occ	Hrs	Recourse action	Cost
CF	male	young	US	Married (civ)	Some Collg.	Private	Sales	32.3	$do(\{\text{Edu} : \text{Prof-school}, \text{WC} : \text{Private}\})$	6.2
CF	male	young	non-US	Married (civ)	HiSch. Grad	Private	Sales	27.8	$do(\{\text{WC} : \text{Self-empl. } (\neg \text{ inc.}), \text{Hrs} : 92.0\})$	64.2
CF	male	old	US	Married (civ)	Some Collg. / Bachelors	Private	Cleaner	36.2	$do(\{\text{Edu} : \text{Prof-school}, \text{WC} : \text{Private}\})$	5.5
CF	male	old	non-US	Married (civ)	HiSch. Grad	Private	Sales	30.3	$do(\{\text{WC} : \text{Self-empl. } (\neg \text{ inc.}), \text{Hrs} : 92.0\})$	61.7
CF	female	young	US	Married (civ)	Some Collg.	Self-empl. (¬inc.)	Sales	27.3	$do(\{\text{Hrs} : 92.0\})$	64.7
CF	female	young	non-US	Married (civ)	HiSch. Grad	Self-empl. (¬inc.)	Sales	24.0	$do(\{\text{Edu} : \text{Some Collg.}, \text{WC} : \text{Self-empl. } (\neg \text{ inc.}), \text{Hrs} : 92.0\})$	68.0
CF	female	old	US	Married (civ)	HiSch. / Some Collg.	Private	Sales	28.8	$do(\{\text{Edu} : \text{Prof-school}, \text{WC} : \text{Private}\})$	6.4
F	female	old	non-US	Married (civ)	HiSch. Grad	W/o pay	Sales	25	$do(\{\text{Hrs} : 92.0\})$	67.0

optimal recourse actions for 10 (uniformly sampled) negatively predicted individuals from each of the eight different protected groups (all combinations of the three protected attributes), as well as for each of their seven counterfactual twins, and use the same metrics as in § 4.1 for evaluation.

Results. We obtain $\Delta_{\textbf{dist}} = 0.89$ and $\Delta_{\textbf{cost}} = 33.32$, indicating group-level recourse discrimination. Moreover, the maximum difference in *distance* is between *old US males* and *old non-US females* (latter is furthest from the boundary), while that in *cost* is between *old US females* and *old non-US females* (latter is most costly). This quantitative and qualitative difference between $\Delta_{\textbf{dist}}$ and $\Delta_{\textbf{cost}}$ emphasises the general need to account for causal-relations in fair recourse, as present in the Adult dataset [3, 25]. At the individual-level, we find an average difference in recourse cost to the counterfactual twins of 24.32 and a maximum ($\Delta_{\textbf{ind}}$) of 61.53 (see a summary for the latter in Tab. 2).

5 On societal interventions

Our notions of fair causal recourse (Defns. 3.1 and 3.2) depend on multiple components $(\mathcal{D}, h, c, \mathcal{M})$. As discussed in § 1, in fair ML, the typical procedure is to *alter the classifier* h. This is the approach proposed for Equalizing Recourse by Gupta et al. [8], which we have discussed in the context of fair *causal* recourse (§ 3.4) and explored experimentally (§ 4). However, requiring the learnt classifier h to satisfy some constraint implicitly places the cost of an intervention on the deployer. For example, a bank might need to modify their classifier so as to offer credit cards to some individuals who would not otherwise receive them. Another possibility is to *alter the data-generating process* (as captured by the SCM $\mathcal{M}$ and manifested in the form of the observed data $\mathcal{D}$) via a *societal intervention* in order to achieve fair causal recourse with a *fixed* classifier h. By considering changes to the underlying SCM or to some of its mechanisms, we may facilitate outcomes which are more societally fair overall, and perhaps end up with a dataset that is more amenable to fair causal recourse (either at the group or individual level). Unlike the setup of Gupta et al. [8], our causal approach here is perhaps particularly well suited to exploring this perspective, as we are already explicitly modelling the causal generative process, i.e., how changes to parts of the system will affect the other variables.

To make things concrete, we demonstrate our ideas for the toy example (6) with different variances across groups from Fig. 1 *(left)*. Here, the difference in recourse cost across groups cannot easily be resolved by changing the classifier $h(X)$ (e.g., per the techniques in § 3.4): to achieve perfectly fair recourse, we would have to use a constant classifier, i.e., either approve all credit cards, or none, irrespective of income. Essentially, changing h does not address the root of the problem, namely the discrepancy in the two populations. Instead, we investigate how to reduce the larger cost of recourse within the higher-variance group by altering the data generating process via societal interventions.

Let i_k denote a societal intervention that modifies the original data generating process (6) by changing the original SCM $\mathcal{M}$ to $\mathcal{M}'_k = i_k(\mathcal{M})$. For example, i_k may introduce additional variables or modify a subset of the original structural equations. Specifically, we consider subsidies to particular eligible individuals. We introduce a new treatment variable T which randomly selects a proportion $0 \leq p \leq 1$ of individuals from $A = 0$ who are awarded a subsidy s if their latent variable U_X is below a threshold t.[7] This is captured by the modified structural equations

$$T := (1 - A)\mathbb{I}\{U_T < p\}, \qquad U_T \sim \text{Uniform}[0, 1],$$
$$X := (2 - A)U_X + sT\mathbb{I}\{U_X < t\}, \qquad U_X \sim \mathcal{N}(0, 1).$$

Here, each societal intervention i_k thus corresponds to a particular way of setting the triple (p, t, s). To avoid changing the predictions $\text{sgn}(X)$, we only consider $t \leq 0$ and $s \leq -2t$. The modified distribution resulting from $i_k = (1, -0.75, 1.5)$ is shown in Fig. 2 *(left)*, see the caption for details. To evaluate the effectiveness of different societal interventions i_k in reducing recourse unfairness, we

[7] s could also vary depending on income, c.f. [7] for further discussion on the role of randomness in fairness.

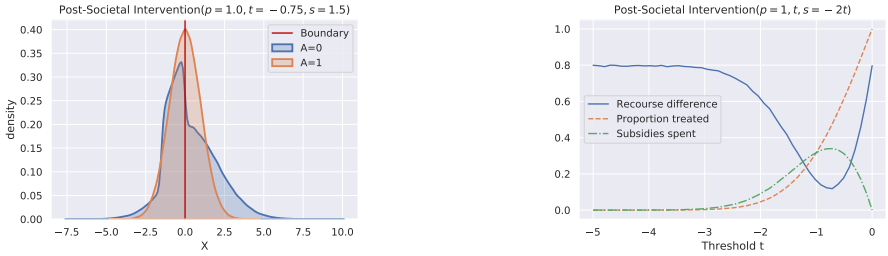


Figure 2: *(Left)* Distribution after applying a societal intervention to the credit-card example from Fig. 1 *(left)*. We randomly select a *proportion* $p = 1$ of individuals from the disadvantaged group (blue, $A = 0$) to receive a *subsidy* $s = 1.5$ if U_X is below the *threshold* $t = -0.75$. As a result, the distribution of negatively-classified individuals ($X < 0$) shifts towards the boundary which makes it more similar to those in $A = 1$, thus resulting in fairer recourse. At the same time, the distribution of positively-classified individuals ($X > 0$) remains unchanged. *(Right)* Comparison of different societal interventions $i_k = (1, t, -2t)$ with respect to their benefit (reduction in recourse difference) and cost (paid-out subsidies). The threshold $t \approx -0.75$ (corresponding to the distribution shown on the left) leads to the largest reduction in recourse difference, but also incurs the highest cost. Smaller reductions can be achieved using two different thresholds: one corresponding to giving a larger subsidy to fewer individuals, and the other to giving a smaller subsidy to more individuals.

compare their associated societal costs c_k and benefits b_k. Here, the cost c_k of implementing i_k can reasonably be chosen as the total amount of paid-out subsidies, and the benefit b_k, as the reduction in the difference of average recourse cost across groups. We then reason about different societal interventions i_k by simulating the proposed change via sampling data from $\mathcal{M}'_k$ and computing b_k and c_k based on the simulated data. To decide which intervention to implement, we compare the societal benefit b_k and cost c_k of i_k for different k and choose the one with the most favourable trade-off. We show the societal benefit and cost tradeoff for $i_k = (1, t, -2t)$ with varying t in Fig. 2 *(Right)* and refer to the caption for further details. Plots similar to Fig. 2 for different choices of (p, t, s) are shown in Fig. 4 in C.1. Effectively, our societal intervention does not change the outcome of credit card approval but ensures that the effort required (additional income needed) for rejected individuals from two groups is the same. Instead of using a threshold to select eligible individuals as in the toy example above, for more complex settings, our individual-level unfairness metric (Definition 3.2) may provide a useful way to inform whom to target with societal interventions as it can be used to identify individuals for whom the counterfactual difference in recourse cost is particularly high.

6 Discussion

With data-driven decision systems pervading our societies, establishing appropriate fairness metrics and paths to recourse are gaining major significance. There is still much work to do in identifying and conceptually understanding the best path forward. Here we make progress towards this goal by applying tools of graphical causality. We are hopeful that this approach will continue to be fruitful as we search together with stakeholders and broader society for the right concepts and definitions, as well as for assaying interventions on societal mechanisms.

We specifically considered the fairness of recourse as opposed to the fairness of predictions. Following earlier work, we take a causal perspective and argue that current non-causal notions of fair recourse are limited in that they do not account for the downstream effects of recourse actions on other (causally related) features. To address this limitation, we introduced causal notions of fair recourse at the group- and individual level and showed that they are complementary to fairness of prediction. While our fairness criteria may help assess the fairness of recourse, it is still unclear how best to achieve fair causal recourse algorithmically. Here, we argue that fairness considerations may benefit from considering the larger system at play—instead of focusing solely on the classifier—and that a causal model of the underlying data generating process provides a principled framework for addressing issues such as multiple sources of unfairness, different costs and benefits to the individual, to institutions, and to society, and changes to the system in the form of societal interventions.

Societal interventions to overcome (algorithmic) discrimination constitute a complex topic which not only applies to fair recourse but also to other notions of fairness. It deserves further study well beyond the scope of the present work. We may also question whether it is appropriate to perform a societal intervention on all individuals in a subgroup. For example, when considering who is approved for a credit card, an individual might not be able to pay their statements on time and this could imply costs to them, to the bank, or to society. This idea relates to the economics literature which studies the effect of policy interventions on society, institutions, and individuals [10, 11]. Thus, future work could focus on formalising the effect of these interventions to the SCM, as such a framework would help trade off the costs and benefits for individuals, companies, and society.

Acknowledgements

We are very grateful to Chris Russell for insightful feedback on connections to existing fairness notions and philosophy. We also thank Matthäus Kleindessner, Adrián Javaloy Bornás, and the anonymous reviewers for helpful comments and suggestions.

AHK is appreciative of NSERC and CLS for generous funding support. UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI), and from the Mozilla Foundation. AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via CFI. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] Richard Arneson. Equality of Opportunity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2015 edition, 2015.

[2] Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

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

[4] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller. On fairness, diversity, and randomness in algorithmic decision making. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[8] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[10] James J Heckman. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature*, 48(2):356–98, 2010.

[11] James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

[12] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[13] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905, 2020.

[14] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[15] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 265–277, 2020.

[16] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[18] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

[19] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

[20] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[21] Moshe Lichman et al. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/adult, 2013.

[22] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[23] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[24] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[25] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 1931, 2018.

[26] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[29] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. MIT Press, 2017.

[30] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[31] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[32] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, USA, 2002.

[33] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

[35] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[36] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 284–293, 2020.

[37] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2017.

[38] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414, 2019.

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[41] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[42] Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681, 2018.

[43] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

APPENDIX

Overview:

- Appendix A provides the proof of Prop. 3.6.
- Appendix B contains additional experimental details.
- Appendix C contains additional results and analysis.

A Proof of Prop. 3.6

Proof. According to Defn. 3.2, it suffices to show that

$$r^{\text{CAU}}(\mathbf{v}^{\text{F}}) = r^{\text{CAU}}(\mathbf{v}_a(\mathbf{u}^{\text{F}})), \qquad \forall a \in \mathcal{A}, \mathbf{v}^{\text{F}} \in \mathcal{D}. \tag{7}$$

Substituting our assumptions in the definition of r^{CAU} from § 3.1, we obtain:

$$r^{\text{CAU}}(\mathbf{v}^{\text{F}}) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \mathcal{F}(\mathbf{v}^{\text{F}})} c(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\text{F}}) \quad \text{subject to} \quad h(\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})) = 1,$$
$$r^{\text{CAU}}(\mathbf{v}_a(\mathbf{u}^{\text{F}})) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \mathcal{F}(\mathbf{v}^{\text{F}})} c(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\text{F}}) \quad \text{subject to} \quad h(\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}},a}(\mathbf{u}^{\text{F}})) = 1.$$

It remains to show that

$$\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}},a}(\mathbf{u}^{\text{F}}) = \tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}}), \qquad \forall \boldsymbol{\theta}_{\mathcal{I}} \in \mathcal{F}(\mathbf{v}^{\text{F}}), a \in \mathcal{A}$$

which follows from applying do-calculus [27] since $\tilde{\mathbf{X}}$ does not contain any descendants of A by assumption, and is thus not influenced by counterfactual changes to A. □

B Experimental Details

In this Appendix, we provide additional details on our experiment setup.

B.1 SCM Specification

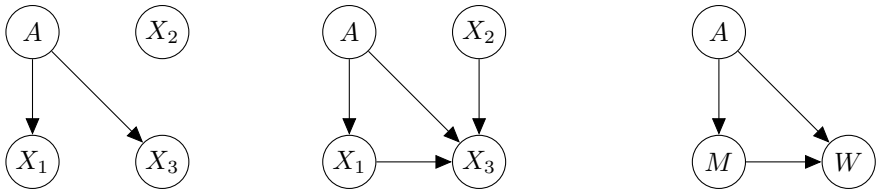
First, we give the exact form of SCMs used to generate our three synthetic data sets IMF, CAU-LIN, and CAU-ANM. Besides the desired characteristics of independently-manipulable (IMF) or causally dependent (CAU) features and linear (LIN) or nonlinear (ANM) relationships with additive noise, we choose the particular form of structural equations for each setting such that all features are roughly standardised, i.e., such that they all approximately have a mean of zero and a variance one.

We use the causal structures shown in Fig. 3. Apart from the desire two make the causal graphs similar to facilitate a better comparison and avoid introducing further nuisance factors while respecting the different structural constraints of the IMF and CAU settings, this particular choice is motivated by having at least one feature which is not a descendant of the protected attribute A. This is so that LR/SVM($\mathbf{X}_{\text{nd}(A)}$) and LR/SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$) always have access to at least one actionable variable (X_2) which can be manipulated to achieve recourse.

B.1.1 IMF

For the IMF data sets, we sample the protected attribute A and the features X_i according to the following SCM:

$$\begin{aligned} A &:= 2U_A - 1, & U_A &\sim \text{Bernoulli}(0.5) \\ X_1 &:= 0.5A + U_1, & U_1 &\sim \mathcal{N}(0,1) \\ X_2 &:= U_2, & U_2 &\sim \mathcal{N}(0,1) \\ X_3 &:= 0.5A + U_3, & U_3 &\sim \mathcal{N}(0,1) \end{aligned}$$



(a) IMF (b) CAU (c) Adult (from [3, 25])

Figure 3: (a) & (b) Causal graphs used to generate synthetic data for our experiments. (c) The (assumed) causal graph used for the Adult dataset [21]; here A denotes the three protected attributes {sex, age, nationality}; M denotes {marital status, education level}; and W corresponds to {working class, occupation, hrs per week}. Here, we show the coarse-grained causal graph for simplicity. In practice, we model each node separately. For example, the single arrow from A to M actually corresponds to six directed edges, one from each feature in A to each feature in M.

B.1.2 CAU-LIN

For the CAU-LIN data sets, we sample A and X_i according to the following SCM:

$$\begin{aligned} A &:= 2U_A - 1, & U_A &\sim \text{Bernoulli}(0.5) \\ X_1 &:= 0.5A + U_1, & U_1 &\sim \mathcal{N}(0,1) \\ X_2 &:= U_2, & U_2 &\sim \mathcal{N}(0,1) \\ X_3 &:= 0.5(A + X_1 - X_2) + U_3, & U_3 &\sim \mathcal{N}(0,1) \end{aligned}$$

B.1.3 CAU-ANM

For the CAU-ANM data sets, we sample A and X_i according to the following SCM:

$$\begin{aligned} A &:= 2U_A - 1, & U_A &\sim \text{Bernoulli}(0.5) \\ X_1 &:= 0.5A + U_1, & U_1 &\sim \mathcal{N}(0,1) \\ X_2 &:= U_2, & U_2 &\sim \mathcal{N}(0,1) \\ X_3 &:= 0.5A + 0.1(X_1^3 - X_2^3) + U_3, & U_3 &\sim \mathcal{N}(0,1) \end{aligned}$$

B.2 Label generation

To generate ground truth labels on which the different classifiers are trained, we consider both a linear and a nonlinear logistic regression. Specifically, we generate ground truth labels according to

$$Y := \mathbb{I}\{U_Y < h(X_1, X_2, X_3)\}, \qquad U_Y \sim \text{Uniform}[0,1].$$

In the linear case, $h(X_1, X_2, X_3)$ is given by

$$h(X_1, X_2, X_3) = \left(1 + e^{-2(X_1 - X_2 + X_3)}\right)^{-1}.$$

In the nonlinear case, $h(X_1, X_2, X_3)$ is given by

$$h(X_1, X_2, X_3) = \left(1 + e^{4 - (X_1 + 2X_2 + X_3)^2}\right)^{-1}.$$

B.3 Fair model architectures and training hyper-parameters

We use the implementation of Gupta et al. [8] for the FairSVM and the `sklearn SVC` class [28] for all other SVM variants. We consider the following values of hyperparameters (which are the same as those reported in [8] for ease of comparison) and choose the best by 5-fold cross validation (unless stated otherwise): kernel type $\in$ {linear, poly, rbf}, regularisation strength $C \in \{1, 10, 100\}$, RBF kernel bandwidth $\gamma_{\text{RBF}} \in \{0.001, 0.01, 0.1, 1\}$, polynomial kernel degree $\in \{2, 3, 5\}$; following [8], we also pick the fairness trade-off parameter $\lambda = \{0.2, 0.5, 1, 2, 10, 50, 100\}$ by cross-validation.

For the nonlinear logistic regression model, we opted for an instance of the `sklearn` `MLPClassifier` class with two hidden layers (10 neurons each) and ReLU activation functions. This model was then optimised on its inputs using the default optimiser and training hyperparameters.

B.4 Optimisation approach

Since an algorithmic contribution for solving the causal recourse optimisation problem is not the main focus of this work, we choose to discretise the space of possible recourse actions and select the best (i.e., lowest cost) valid action by performing a brute-force search. For an alternative gradient-based approach to solving the causal recourse optimisation problem, we refer to [15].

For each actionable feature X_i, denote by $\max_i$ and $\min_i$ its maximum and minimum attained in the training set, respectively. Given a factual observation x_i^{F} of X_i, we discretise the search space and pick possible intervention values θ_i using 15 equally-spaced bins in the range $[x_i^{\mathsf{F}} - 2(x_i^{\mathsf{F}} - \min_i), x_i^{\mathsf{F}} + 2(\max_i - x_i^{\mathsf{F}})]$. We then consider all possible combinations of intervention values over all subsets $\mathcal{I}$ of the actionable variables. We note that for LR/SVM($\mathbf{X}_{\text{nd}}$) and LR/SVM($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$), only X_2 is actionable, while for the other LR/SVMs all of $\{X_1, X_2, X_3\}$ are actionable.

B.5 Adult dataset case study

The causal graph for the Adult dataset informed by expert knowledge [3, 25] is depicted in Fig. 3c.

Because the true structural equations are not known, we learn an approximate SCM for the Adult dataset by fitting each parent-child relationship in the causal graph. Since most variables in the Adult dataset are categorical, additive noise is not an appropriate assumption for most of them. We therefore opt for modelling each structural equation, $X_i := f_i(\text{PA}_i, U_i)$, using a latent variable model; specifically, we use a conditional variational autoencoder (CVAE) [34], similar to [15].

We use deterministic conditional decoders $D_i(\text{PA}_i, U_i; \psi_i)$, implemented as neural nets parametrised by ψ_i, and use an isotropic Gaussian prior, $U_i \sim \mathcal{N}(0, \mathbf{I})$, for each X_i.

For continuous features, the decoders directly output the value assigned to X_i, i.e., we approximate the structural equations as

$$X_i^{\text{continuous}} := D_i(\text{PA}_i, U_i; \psi_i), \qquad U_i \sim \mathcal{N}(0, \mathbf{I}). \tag{8}$$

For categorical features, the decoders output a vector of class probabilities (by applying a softmax operation after the last layer). The arg max is then assigned as the value of the corresponding categorical feature, i.e.,

$$X_i^{\text{categorical}} := \arg\max D_i(\text{PA}_i, U_i; \psi_i), \qquad U_i \sim \mathcal{N}(0, \mathbf{I}). \tag{9}$$

The decoders $D_i(\text{PA}_i, U_i; \psi_i)$ are trained using the standard variational framework [19, 30], amortised with approximate Gaussian posteriors $q_{\phi_i}(U_i | X_i, \text{PA}_i)$ whose means and variances are computed by encoders in the form of neural nets with parameters ϕ_i. For continuous features, we use the standard reconstruction error between real-valued predictions and targets, i.e., L2/MSE$(X_i, D_i(\text{PA}_i, U_i; \psi_i))$, whereas for categorical features, we instead use the cross entropy loss between the one-hot encoded value of X_i and the predicted vector of class probabilities, i.e., CrossEnt$(X_i, \text{softmax}(D_i(\text{PA}_i, U_i; \psi_i))$.

The CVAEs are trained on $6,000$ training samples using a fixed learning rate of 0.05, and a batch size of 128 for 100 epochs with early stopping on a held-out validation set of 250 samples. For each parent-child relation, we train 10 models with the number of hidden layers and units randomly drawn from the following configurations for the encoder: $\text{enc}_{\text{arch}} = \{(\zeta, 2, 2), (\zeta, 3, 3), (\zeta, 5, 5), (\zeta, 32, 32, 32)\}$, where ζ is the input dimensionality; and similarly for the decoders from: $\text{dec}_{\text{arch}} = \{(2, \eta), (2, 2, \eta), (3, 3, \eta), (5, 5, \eta), (32, 32, 32, \eta)\}$, where η is either one for continuous variables, or alternatively the size of the one-hot embedding for categorical variables (e.g., Work Class, Marital Status, and Occupation have 7, 10, and 14 categories, respectively). Moreover, we also randomly pick a latent dimension from $\{1, 3, 5\}$. We then select the model with the smallest MMD score [6] between true instances and samples from the decoder post-training.

To perform abduction for counterfactual reasoning with such an approximate CVAE-SCM, we sample U_i from the approximate posterior. For further discussion, we refer to [15], Appendix C.

Finally, using this approximate SCM, we solve the recourse optimisation problem similar to the synthetic experiments above. The caveat with this approach (and any real-world dataset absent a true SCM for that matter) is that we are not able to certify that a given recourse action generated under

the assumption of an approximate SCM will guarantee recourse when executed in the true world governed by the real (unknown) SCM.

C Additional Results

In this Appendix, we provide additional experimental results omitted from the main paper due to space constraints.

C.1 Additional Societal Interventions

In § 5, we only showed plots for i_k with $p = 1$ since this has the largest potential to reduce recourse unfairness. However, it may not be feasible to give subsidies to all eligible individuals, and so, for completeness, we also show plots similar to Fig. 2 for different choices of (p, t, s) in Fig. 4.

C.2 Using different SCMs for Recourse

The results presented in § 4.1 of the main paper use an estimate $\hat{\mathcal{M}}_{\text{KR}}$ of the ground truth SCM $\mathcal{M}^\star$ (learnt via kernel ridge regression under an additive noise assumption) to solve the recourse optimisation problem.

In Tab. 3 we show a more complete account of results which also includes the cases where the ground truth SCM $\mathcal{M}^\star$ or a linear ($\hat{\mathcal{M}}_{\text{LIN}}$) estimate thereof is used as the basis for computing recourse actions. When using an SCM estimate for recourse, we only consider *valid* actions to compute $\Delta_{\textbf{cost}}$ and $\Delta_{\textbf{ind}}$, where the validity of an action is determined by whether it results in a changed prediction according the oracle $\mathcal{M}^\star$.

We find that, as expected, using different SCMs does not affect **Acc** or $\Delta_{\textbf{dist}}$ since these metrics are, by definition, agnostic to the underlying causal generative process encoded by the SCM. However, using an estimated SCM in place of the true one may result in different values for $\Delta_{\textbf{cost}}$ and $\Delta_{\textbf{ind}}$ since these metrics take the downstream effects of recourse actions on other features into account and thus depend on the underlying SCM, c.f. Defns. 3.1 and 3.2.

We observe that using an estimated SCM may lead to underestimating or overestimating the true fair causal recourse metric (without any apparent clear trend as to when one or the other occurs). Moreover, the mis-estimation of fair causal recourse metrics is particularly pronounced when using the linear SCM estimate $\hat{\mathcal{M}}_{\text{LIN}}$ in a scenario in which the true SCM is, in fact, nonlinear, i.e., on the CAU-ANM data sets. This behaviour is intuitive and to be expected and should caution against using overly strong assumptions or too simplistic parametric models when estimating an SCM for use in (fair) recourse. We also remark that, in practice, underestimation of the true fairness metric is probably more problematic than overestimation.

Despite some small differences, the overall trends reported in § 4.1 remain very much the same, and thus seem relatively robust to small differences in the SCM which is used to compute recourse actions.

C.3 Kernel selection by cross validation

For completeness, we perform the same set of experiments shown in Tab. 3 where we also choose the kernel function by cross validation, instead of fixing it to either a linear or a polynomial kernel as before. The results are shown in Tab. 4 and the overall trends, again, remain largely the same.

As expected, we observe variations in accuracy compared to Tab. 3 due to the different kernel choice. Perhaps most interestingly, the FairSVM seems to generally perform slightly better in terms of $\Delta_{\textbf{dist}}$ when given the "free" choice of kernel, especially on the first three data sets with linearly generated labels. This suggests that *the use of a nonlinear kernel may be important for FairSVM to achieve its goal.*

However, we caution that the results in Tab. 4 may not be easily comparable across classifiers as distances are computed in the induced feature spaces which are either low-dimensional (in case of a linear kernel), high-dimensional (in case of a polynomial kernel), or infinite-dimensional (in case of an RBF kernel), which is also why we chose to report results based on the same kernel type in § 4.

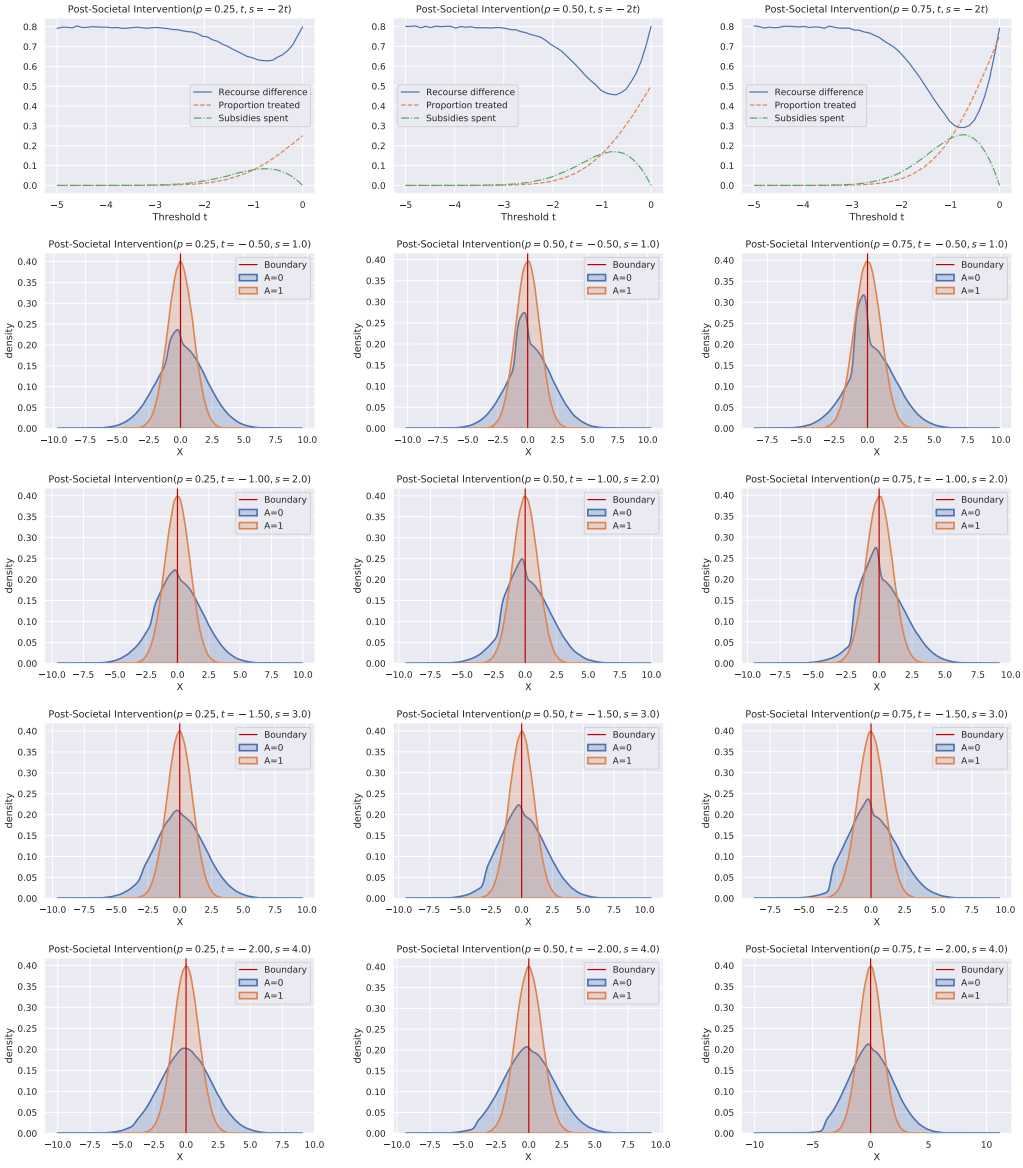


Figure 4: Plots for additional societal interventions $i_k = (p, t, s)$ in the context of the credit card approval example. We consider budgets of $p = 0.25$ (left column), $p = 0.5$ (middle column), and $p = 0.75$ (right column). In the top row, we show the difference across groups in the average distance to the decision boundary for negatively classified individuals (recourse difference), the proportion of negatively-classified individuals in the disadvantaged group $A = 0$ who received the treatment (proportion treated), and the amount of subsidies actually paid out to these individuals (subsidies spent) as a function of the threshold t. Note that the subsidy amount is fixed to its maximum amount without affecting the label distribution, i.e., $s = -2t$. In rows 2-5, we show the feature distribution resulting from $i_k = (p, t, s)$ with $t = -2s \in \{-0.5, -1, -1.5, -2\}$, see the plot titles for the exact values.

Table 3: **Complete account of experimental results corresponding to the setting described in § 4 of the main paper, where we additionally consider using the true SCM $\mathcal{M}^\star$ or a linear ($\hat{\mathcal{M}}_{\mathbf{LIN}}$) estimate thereof to infer the latent variables U and solve the recourse optimisation problem.** We compare different classifiers with respect to accuracy and different recourse fairness metrics on our three synthetic data sets with ground truth labels drawn from either a linear or a nonlinear logistic regression. For ease of comparison, we use the same kernel for all SVM variants for a given dataset: a linear kernel for linearly generated ground truth labels and a polynomial kernel for non-linearly generated ground truth labels. Moreover, linear (resp. nonlinear) logistic regression classifiers are used for linearly (resp. nonlinearly) generated ground truth labels. All other hyper-parameters are chosen by 10-fold cross-validation. We use a dataset of 500 observations for all experiments and make sure that it is roughly balanced, both with respect to the protected attribute A and the label Y. Accuracies (higher is better) are computed on a separate i.i.d. test set of equal size. Fairness metrics (lower is better) are computed based on randomly selecting 50 negatively-classified samples from each of the two protected groups and using these to compute the difference between group-wise averages ($\Delta_{\mathbf{dist}}$ and $\Delta_{\mathbf{cost}}$) and maximum individual unfairness. When using an SCM estimate for recourse, we only consider valid actions to compute $\Delta_{\mathbf{cost}}$ and $\Delta_{\mathbf{ind}}$, where the validity of an action is determined by whether it results in a changed prediction according the oracle $\mathcal{M}^\star$. For each experiment and metric, the best performing method is highlighted in **bold**.

SCM	**Classifier**	**GT labels from *linear* log. reg. → using *linear* kernel / *linear* log. reg.**												**GT labels from *nonlinear* log. reg. → using *polynomial* kernel / *nonlinear* log. reg.**											
		IMF				**CAU-LIN**				**CAU-ANM**				**IMF**				**CAU-LIN**				**CAU-ANM**			
		Acc	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$	**Acc**	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$	**Acc**	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$	**Acc**	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$	**Acc**	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$	**Acc**	$\Delta_{\mathbf{dist}}$	$\Delta_{\mathbf{cost}}$	$\Delta_{\mathbf{ind}}$
$\mathcal{M}^\star$	$\text{SVM}(\mathbf{X}, A)$	86.5	0.96	0.40	1.63	89.5	1.18	0.43	2.11	**88.2**	0.65	**0.12**	2.41	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.04	1.06	90.6	0.04	0.07	1.40
	$\text{LR}(\mathbf{X}, A)$	**86.7**	0.48	0.50	1.91	89.5	0.63	0.49	2.11	87.7	0.40	0.22	2.41	90.5	0.08	0.03	1.06	90.6	0.09	**0.02**	1.00	90.6	0.19	0.18	1.28
	$\text{SVM}(\mathbf{X})$	86.4	0.99	0.42	1.80	89.4	1.61	0.61	2.11	88.0	0.56	**0.12**	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.09	1.09	**91.0**	**0.02**	**0.02**	1.64
	$\text{LR}(\mathbf{X})$	86.6	0.47	0.53	1.80	89.5	0.64	0.52	2.11	87.7	0.41	0.31	2.79	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	90.9	0.08	**0.02**	1.16
	$\text{FairSVM}(\mathbf{X}, A)$	68.1	**0.04**	0.28	1.36	66.8	0.26	**0.12**	0.78	66.3	0.25	0.21	1.50	90.1	**0.02**	**0.00**	1.15	90.7	0.06	0.04	1.16	90.3	0.37	0.03	1.64
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)})$	65.5	0.05	0.06	**0.00**	67.4	**0.15**	0.17	**0.00**	65.9	0.31	0.37	**0.00**	66.7	0.10	0.06	**0.00**	58.4	0.05	0.06	**0.00**	62.0	0.13	0.11	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)})$	65.3	0.05	**0.05**	**0.00**	67.3	0.18	0.18	**0.00**	65.6	0.31	0.31	**0.00**	64.7	**0.02**	0.04	**0.00**	58.4	**0.02**	**0.02**	**0.00**	61.1	**0.02**	0.03	**0.00**
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	86.5	0.96	0.58	**0.00**	**89.6**	1.07	0.70	**0.00**	88.0	**0.21**	0.14	**0.00**	90.7	**0.02**	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	90.1	0.15	0.12	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	**86.7**	0.43	0.90	**0.00**	89.5	0.35	0.77	**0.00**	87.8	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**
$\hat{\mathcal{M}}_{\text{LIN}}$	$\text{SVM}(\mathbf{X}, A)$	86.5	0.96	0.40	1.63	89.5	1.18	0.44	2.11	**88.2**	0.65	0.30	3.77	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.04	1.06	90.6	0.04	0.04	1.49
	$\text{LR}(\mathbf{X}, A)$	**86.7**	0.48	0.50	1.91	89.5	0.63	0.51	2.11	87.7	0.40	0.43	3.77	90.5	0.08	0.03	1.06	90.6	0.09	**0.01**	1.00	90.6	0.19	0.20	1.28
	$\text{SVM}(\mathbf{X})$	86.4	0.99	0.42	1.80	89.4	1.61	0.61	2.11	88.0	0.56	0.20	3.48	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.10	1.09	**91.0**	**0.02**	0.03	1.49
	$\text{LR}(\mathbf{X})$	86.6	0.47	0.53	1.80	89.5	0.64	0.58	2.11	87.7	0.41	0.55	3.48	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	90.9	0.08	0.04	1.66
	$\text{FairSVM}(\mathbf{X}, A)$	68.1	**0.04**	0.28	1.36	66.8	0.26	**0.12**	0.78	66.3	0.25	0.21	1.50	90.1	**0.02**	**0.00**	1.15	90.7	0.06	0.05	1.16	90.3	0.37	**0.01**	1.64
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)})$	65.5	0.05	0.06	**0.00**	67.4	**0.15**	0.17	**0.00**	65.9	0.31	0.37	**0.00**	66.7	0.10	0.06	**0.00**	58.4	0.05	0.06	**0.00**	62.0	0.13	0.11	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)})$	65.3	0.05	**0.05**	**0.00**	67.3	0.18	0.18	**0.00**	65.6	0.31	0.31	**0.00**	64.7	**0.02**	0.04	**0.00**	58.4	**0.02**	0.02	**0.00**	61.1	**0.02**	0.03	**0.00**
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	86.5	0.96	0.58	**0.00**	**89.6**	1.07	0.70	**0.00**	88.0	**0.21**	**0.14**	**0.00**	90.7	**0.02**	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	90.1	0.15	0.12	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	**86.7**	0.43	0.90	**0.00**	89.5	0.35	0.77	**0.00**	87.8	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**
$\hat{\mathcal{M}}_{\text{KR}}$	$\text{SVM}(\mathbf{X}, A)$	86.5	0.96	0.40	1.63	89.5	1.18	0.44	2.11	**88.2**	0.65	0.27	2.32	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.03	1.06	90.6	0.04	0.03	1.40
	$\text{LR}(\mathbf{X}, A)$	**86.7**	0.48	0.50	1.91	89.5	0.63	0.53	2.11	87.7	0.40	0.34	2.32	90.5	0.08	0.03	1.06	90.6	0.09	**0.01**	1.00	90.6	0.19	0.22	1.28
	$\text{SVM}(\mathbf{X})$	86.4	0.99	0.42	1.80	89.4	1.61	0.61	2.11	88.0	0.56	0.29	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.08	1.09	**91.0**	**0.02**	0.03	1.64
	$\text{LR}(\mathbf{X})$	86.6	0.47	0.53	1.80	89.5	0.64	0.57	2.11	87.7	0.41	0.43	2.79	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	90.9	0.08	0.06	1.66
	$\text{FairSVM}(\mathbf{X}, A)$	68.1	**0.04**	0.28	1.36	66.8	0.26	**0.12**	0.78	66.3	0.25	0.21	1.50	90.1	**0.02**	**0.00**	1.15	90.7	0.06	0.04	1.16	90.3	0.37	**0.02**	1.64
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)})$	65.5	0.05	0.06	**0.00**	67.4	**0.15**	0.17	**0.00**	65.9	0.31	0.37	**0.00**	66.7	0.10	0.06	**0.00**	58.4	0.05	0.06	**0.00**	62.0	0.13	0.11	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)})$	65.3	0.05	**0.05**	**0.00**	67.3	0.18	0.18	**0.00**	65.6	0.31	0.31	**0.00**	64.7	**0.02**	0.04	**0.00**	58.4	**0.02**	0.02	**0.00**	61.1	**0.02**	0.03	**0.00**
	$\text{SVM}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	86.5	0.96	0.58	**0.00**	**89.6**	1.07	0.70	**0.00**	88.0	**0.21**	**0.14**	**0.00**	90.7	**0.02**	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	90.1	0.15	0.12	**0.00**
	$\text{LR}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	**86.7**	0.43	0.90	**0.00**	89.5	0.35	0.77	**0.00**	87.8	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**

Table 4: **Additional results where also the kernel (linear, polynomial, or rbf) for each SVM is chosen by 5-fold cross-validation instead of being fixed based on the ground truth label distribution.** We remark that some metrics (e.g., $\Delta_{\textbf{dist}}$) may not be comparable across methods since they are computed in a different reference space when different kernels are selected. Otherwise the experimental setup is identical to that from Tab. 3, see the caption for details.

SCM	**Classifier**	**GT labels from *linear* log. reg. → using *cross-validated* kernel / *linear* log. reg.**												**GT labels from *nonlinear* log. reg. → using *cross-validated* kernel / *nonlinear* log. reg.**											
		IMF				**CAU-LIN**				**CAU-ANM**				**IMF**				**CAU-LIN**				**CAU-ANM**			
		Acc	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$	**Acc**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$	**Acc**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$	**Acc**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$	**Acc**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$	**Acc**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$
$\mathcal{M}^{\star}$	SVM($\mathbf{X}, A$)	86.5	0.96	0.40	1.63	89.2	1.33	0.55	2.10	**87.8**	0.36	**0.08**	2.79	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.04	1.06	90.6	0.04	0.07	1.40
	LR($\mathbf{X}, A$)	**86.7**	0.48	0.50	1.91	**89.5**	0.63	0.49	2.11	87.7	0.40	0.22	2.41	90.5	0.08	0.03	1.06	90.6	0.09	**0.02**	1.00	90.6	0.19	0.18	1.28
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.13	0.53	2.14	87.6	0.43	0.42	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.09	1.09	89.4	0.16	0.16	1.16
	LR($\mathbf{X}$)	86.6	0.47	0.53	1.80	**89.5**	0.64	0.52	2.11	87.7	0.41	0.31	2.79	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	**90.9**	0.08	**0.02**	1.16
	FairSVM($\mathbf{X}, A$)	86.4	**0.01**	0.20	1.61	60.5	**0.00**	0.33	1.05	57.6	**0.01**	0.13	1.76	90.1	0.02	**0.00**	1.15	90.7	0.06	0.04	1.16	78.0	**0.00**	0.04	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.06	0.09	**0.00**	67.3	0.17	0.25	**0.00**	65.9	0.28	0.31	**0.00**	65.7	**0.01**	0.02	**0.00**	55.6	0.04	0.03	**0.00**	61.6	0.04	0.03	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}$)	65.3	0.05	**0.05**	**0.00**	67.3	0.18	**0.18**	**0.00**	65.6	0.31	0.31	**0.00**	64.7	0.02	0.04	**0.00**	58.4	**0.02**	**0.02**	**0.00**	61.1	0.02	0.03	**0.00**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	86.6	0.84	0.64	**0.00**	89.4	0.81	0.54	**0.00**	87.4	0.21	0.35	**0.00**	90.7	0.02	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	89.0	0.31	0.13	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.7**	0.43	0.90	**0.00**	**89.5**	0.35	0.77	**0.00**	**87.8**	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**
$\hat{\mathcal{M}}_{\text{LIN}}$	SVM($\mathbf{X}, A$)	86.5	0.96	0.40	1.63	89.2	1.33	0.55	2.10	**87.8**	0.36	0.13	3.48	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.04	1.06	90.6	0.04	0.04	1.49
	LR($\mathbf{X}, A$)	**86.7**	0.48	0.50	1.91	**89.5**	0.63	0.51	2.11	87.7	0.40	0.43	3.77	90.5	0.08	0.03	1.06	90.6	0.09	**0.01**	1.00	90.6	0.19	0.20	1.28
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.13	0.51	2.14	87.6	0.43	0.42	4.05	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.10	1.09	89.4	0.16	0.11	1.16
	LR($\mathbf{X}$)	86.6	0.47	0.53	1.80	**89.5**	0.64	0.58	2.11	87.7	0.41	0.55	3.48	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	**90.9**	0.08	0.04	1.66
	FairSVM($\mathbf{X}, A$)	86.4	**0.01**	0.20	1.61	60.5	**0.00**	0.29	1.05	57.6	**0.01**	**0.12**	1.76	90.1	0.02	**0.00**	1.15	90.7	0.06	0.05	1.16	78.0	**0.00**	**0.03**	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.06	0.09	**0.00**	67.3	0.17	0.25	**0.00**	65.9	0.28	0.31	**0.00**	65.7	**0.01**	0.02	**0.00**	55.6	0.04	0.03	**0.00**	61.6	0.04	**0.03**	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}$)	65.3	0.05	**0.05**	**0.00**	67.3	0.18	**0.18**	**0.00**	65.6	0.31	0.31	**0.00**	64.7	0.02	0.04	**0.00**	58.4	**0.02**	0.02	**0.00**	61.1	0.02	**0.03**	**0.00**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	86.6	0.84	0.64	**0.00**	89.4	0.81	0.54	**0.00**	87.4	0.21	0.35	**0.00**	90.7	0.02	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	89.0	0.31	0.13	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.7**	0.43	0.90	**0.00**	**89.5**	0.35	0.77	**0.00**	**87.8**	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**
$\hat{\mathcal{M}}_{\text{KR}}$	SVM($\mathbf{X}, A$)	86.5	0.96	0.40	1.63	89.2	1.33	0.56	2.10	**87.8**	0.36	0.18	2.79	90.8	0.05	**0.00**	1.09	**91.1**	0.07	0.03	1.06	90.6	0.04	0.03	1.40
	LR($\mathbf{X}, A$)	**86.7**	0.48	0.50	1.91	**89.5**	0.63	0.53	2.11	87.7	0.40	0.34	2.32	90.5	0.08	0.03	1.06	90.6	0.09	**0.01**	1.00	90.6	0.19	0.22	1.28
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.13	0.52	2.14	87.6	0.43	0.44	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.08	1.09	89.4	0.16	0.14	1.16
	LR($\mathbf{X}$)	86.6	0.47	0.53	1.80	**89.5**	0.64	0.57	2.11	87.7	0.41	0.43	2.79	91.0	0.12	0.03	1.01	90.6	0.13	0.10	1.65	**90.9**	0.08	0.06	1.66
	FairSVM($\mathbf{X}, A$)	86.4	**0.01**	0.20	1.61	60.5	**0.00**	0.26	1.50	57.6	**0.01**	**0.12**	1.76	90.1	0.02	**0.00**	1.15	90.7	0.06	0.04	1.16	78.0	**0.00**	**0.01**	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.06	0.09	**0.00**	67.3	0.17	0.25	**0.00**	65.9	0.28	0.31	**0.00**	65.7	**0.01**	0.02	**0.00**	55.6	0.04	0.03	**0.00**	61.6	0.04	0.03	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}$)	65.3	0.05	**0.05**	**0.00**	67.3	0.18	**0.18**	**0.00**	65.6	0.31	0.31	**0.00**	64.7	0.02	0.04	**0.00**	58.4	**0.02**	0.02	**0.00**	61.1	0.02	0.03	**0.00**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	86.6	0.84	0.64	**0.00**	89.4	0.81	0.54	**0.00**	87.4	0.21	0.35	**0.00**	90.7	0.02	0.03	**0.00**	**91.1**	0.15	0.11	**0.00**	89.0	0.31	0.13	**0.00**
	LR($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.7**	0.43	0.90	**0.00**	**89.5**	0.35	0.77	**0.00**	**87.8**	0.14	0.34	**0.00**	90.9	0.28	0.05	**0.00**	90.9	0.49	0.07	**0.00**	90.2	0.43	0.21	**0.00**