ON THE FAIRNESS OF CAUSAL ALGORITHMIC RECOURSE

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

Isabel Valera [1,4] **Adrian Weller** [2,5] **Bernhard Schölkopf** [1,3]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[3] ETH Zürich, Zürich, Switzerland
[4] Department of Computer Science, Saarland University, Saarbrücken, Germany
[5] The Alan Turing Institute, London, United Kingdom
{jvk,amir,ivalera,bs}@tue.mpg.de, {usb20,aw665}@cam.ac.uk

February 23, 2021

ABSTRACT

Many recent works have studied the problem of algorithmic fairness from the perspective of *predictions*. Instead, here we investigate the fairness of *recourse* actions recommended to individuals to recover from an unfavourable classification. We propose two new fairness criteria at the group and individual level which—unlike prior work on equalising the average distance from the decision boundary across protected groups—explicitly account for the causal relationships between input features, thereby allowing us to capture 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 (both causal and non-causal) is complementary to fairness of prediction. We then investigate how to enforce fair causal recourse in the training of a classifier. 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 in machine learning (ML) is a primary area of study for researchers concerned with uncovering and correcting for potentially discriminatory behavior of ML models (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Chouldechova, 2017). Parallel to this, *algorithmic recourse* is concerned with offering explanations and recommendations to individuals who were unfavourably treated by ML-based decision-making systems to overcome their adverse situation (Joshi et al., 2019; Ustun et al., 2019; Sharma et al., 2019; Venkatasubramanian and Alfano, 2020; Karimi et al., 2020b,c). Thus far, the literature has only informally studied the *relation between fairness and recourse*, e.g., recourse methods have been used as proxies for evaluating the fairness of a trained prediction system. For example, Ustun et al. (2019) look at comparable male/female individuals that were denied a loan and show that a disparity can be detected in that the suggested recourse actions (*flipsets*) require relatively more effort for individuals of a particular sub-group. Similarly, Sharma et al. (2019) evaluate group fairness via aggregating and comparing the cost of recourse (*burden*) over individuals of different sub-populations, and Karimi et al. (2020a) show that if the cost of recourse increases as a result of adding feasibility constraints on a protected attribute (e.g., non-decreasing age), then this indicates a reliance of the classifier on the protected attribute, which is typically considered as legally and socially unfair.

The examples above seem to suggest that evidence of *discriminatory recourse* always involves (i.e., logically implies) *discriminatory prediction*; however, this is not the case. Consider, for example, the data set shown in Figure 1 which comprises two sub-groups with feature X distributed as $\mathcal{N}(0,4)$ and $\mathcal{N}(0,1)$, respectively, i.e., only the variances differ. Moreover, consider a binary classifier $h(x) = \text{sign}(x)$ which perfectly predicts the true labels. While the distribution of

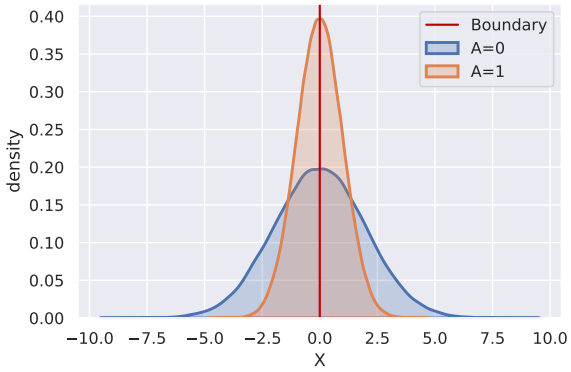

Figure 1: A simple toy example demonstrates the difference between fair *prediction* and fair *recourse*: here, only the variance of feature X differs across two protected groups $A = 0$ and $A = 1$, while the true and predicted label are determined by $\text{sign}(X)$. This scenario would be considered fair from the perspective of prediction, but the cost of recourse (here, distance to the decision boundary) is much larger for individuals in the group with $A = 0$.

predictions satisfies several fairness criteria (e.g., demographic parity, equalised odds, calibration), the effort required by negatively classified individuals to achieve recourse (i.e., cross the boundary) is much larger for the first group. This suggests to consider fairness of recourse as a *distinct notion of fairness* that is not necessarily implied by the fairness of prediction.

In this regard, *Equalising Recourse* was recently presented by Gupta et al. (2019), offering the first recourse-based and prediction-independent notion of fairness. Gupta et al. demonstrate that one can directly calibrate for the average distance to the decision boundary for those getting a bad outcome to be equalised across different subgroups during the training of linear and nonlinear classifiers. This formulation, however, does not take causal relations between variables into account when generating recourse, which has been addressed by several recent studies (Mahajan et al., 2019; Karimi et al., 2020c,d; Mothilal et al., 2020), c.f. also (Wachter et al., 2017; Ustun et al., 2019), generally based on the intuition that actively changing some variables may have downstream effects on other variables.

Our Contributions In this work, we study the fairness of recourse from a causal perspective, and generalise the idea of *Equalising Recourse* by considering the *interventional cost of recourse*—as opposed to the distance to the decision boundary—in flipping the prediction across subgroups. Building on the framework of Karimi et al. (2020c), which uses a causal model of the world in which actions are undertaken to generate a set of minimal interventions for recourse, we propose two new definitions for fair recourse (§3). The first is a group-level criterion (§3.1) similar to that of Gupta et al. (2019), the second is an individualised notion (§3.2) inspired by counterfactual fairness (Kusner et al., 2017). We show that fairness of recourse is complementary to fairness of prediction, and explore further links to counterfactual fairness (§3.3). We then investigate different paths towards achieving fair causal recourse at the level of the classifier (§4) and critically discuss these in the larger context of societal interventions (§5). Finally, we empirically validate our main claims by comparing a number of (fair) classifiers with respect to different definitions of fair recourse in experiments on synthetic data sets (§6).

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, $A \in \mathcal{A} = \{1, \dots, K\}$ the protected attribute indicating which social salient 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 $Y \in \mathcal{Y} = \{\pm 1\}$ denoting the predicted label (e.g., whether a loan is denied or 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 $\mathbf{v}^i = (\mathbf{x}^i, a^i)$.[1]

Our work builds on the explainability and fairness literature, as well as on causal modelling and counterfactual reasoning. We review the most important concepts below.

[1] 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).

(a) IMF assumption (b) Causal view

Figure 2: (a) 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. (b) The setting studied in this work generalises the IMF assumption by allowing causal influences between the X_i, thus considering downstream effects of changing some features on others.

Counterfactual Explanations For explaining decisions made by (black-box) machine learning models, Wachter et al. (2017) proposed the framework of nearest counterfactual explanations (CEs). A CE is a closest feature vector on the other side of the decision boundary. Given a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, 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 the following optimisation problem:

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

While CEs are a useful tool 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; e.g., a CE with decreased age would be infeasible.

Recourse with Independently Manipulable Features Ustun et al. (2019) refer to "the ability of a person to change the decision of a model by altering actionable input 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. (2020c), the framework in (2) treats features as manipulable independently of each other (see Figure 2a) and does not consider causal relations that may exist between them (see Figure 2b). In other words, (2) considers feasibility constraints on actions, but assumes that variables which are not acted-upon ($\delta_i = 0$) remain unchanged. We refer to this assumption underlying both CEs and the approach in (2) 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: for example, an increase in income will likely also have a positive downstream effect on the individual's savings balance. 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 (Karimi et al., 2020c).

Structural Causal Models A structural causal model (SCM) (Pearl, 2009; Peters et al., 2017) 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 (*causal sufficiency*). If the associated 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 also 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 (atomic) 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. (2020c) 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

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

where $\mathbf{x}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\mathrm{F}})$ denotes the "counterfactual twin" of $\mathbf{x}^{\mathrm{F}}$ 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 (Karimi et al., 2020d). 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 of Ustun et al. (2019) (2) as a special case.

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 (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Zafar et al., 2017a,b), these are sometimes mutually incompatible (Chouldechova, 2017) 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 (Kilbertus et al., 2017; Russell et al., 2017; Loftus et al., 2018; Zhang and Bareinboim, 2018a,b; Nabi and Shpitser, 2018; Nabi et al., 2019; Chiappa, 2019; Wu et al., 2019; Kilbertus et al., 2020). Of particular interest to our work is the notion of *counterfactual fairness* introduced by Kusner et al. (2017) which calls a (probabilistic) classifier h over $\mathbf{V} = \mathbf{X} \cup A$ counterfactually fair if it satisfies

$$h(\mathbf{v}^{\mathrm{F}}) = h(\mathbf{v}_a(\mathbf{u}^{\mathrm{F}})) \qquad \forall a \in \mathcal{A}, \mathbf{v}^{\mathrm{F}} = (\mathbf{x}^{\mathrm{F}}, a^{\mathrm{F}}) \in \mathcal{X} \times \mathcal{A},$$

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}.

2.1 Equalising 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. (2019). They advocate for equalising 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), i.e.,

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

which is equivalent to the IMF-recourse (2) of Ustun et al. (2019) 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 following protected subgroups,

$$G_a = \{\mathbf{v}^i \in \mathcal{D} : a^i = a\}, \qquad G_a^- = \{\mathbf{v} \in G_a : h(\mathbf{v}) = -1\},$$

the group-level cost of recourse is then given by,

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

The idea of fair recourse by equalising cost (i.e., distance from the decision boundary) across protected subgroups by Gupta et al. (2019) can then be summarised as follows.

Definition 1 (Group-level fair IMF-recourse; adapted from Gupta et al. (2019))**.** *The group-level unfairness of* recourse with independently-manipulable features *(IMF) for a dataset $\mathcal{D}$, classifier h, and distance metric d is given by*

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

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

3 Fair causal recourse

Since the notion of fair recourse of Gupta et al. (2019) rests on the assumption of independently manipulable features (IMF) inherent to the view of CEs (1) (Wachter et al., 2017) and IMF-based recourse (2) (Ustun et al., 2019), it exhibits the shortcomings pointed out by Karimi et al. (2020c,d): it does not model causal relationships between variables, fails to account for downstream effects of actions on other relevant features, and may thus incorrectly estimate the true cost of recourse. In the present work, we argue that considerations regarding the fairness of recourse actions should be based

on an underlying causal model capturing the effect of interventions performed in the physical world where features may be causally related to each other.

In §3.1, as an alternative to Definition 1, we consider a group-level fairness criterion building on the causal view of recourse (3) (Karimi et al., 2020c), before moving in §3.2 to an individualised notion of fair recourse inspired by counterfactual fairness (Kusner et al., 2017). Throughout we consider an SCM $\mathcal{M}$ over $\mathbf{V} = (\mathbf{X}, A)$ to model causal relationships between the protected attribute and the remaining features. For generality, we assume that the classifier h is defined over $\mathcal{X} \times \mathcal{A}$, though most considerations will equally apply to classifiers blind to the protected attribute.

3.1 Group-level fair causal recourse

A direct adaptation of Definition 1 to the causal (CAU) recourse framework (3) is obtained by replacing the minimum distance in (4) with the cost of recourse actions performed within a causal model, i.e., the minimum achieved in (3),

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

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

Definition 2 (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*

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

We say that recourse for $(\mathcal{D}, h, c, \mathcal{M})$ is "group CAU-fair at level $\epsilon > 0$" if $\Delta_{\text{cost}}(\mathcal{D}, h, c, \mathcal{M}) < \epsilon$, and that it is "group CAU-fair" if $\Delta_{\text{cost}}(\mathcal{D}, h, c, \mathcal{M}) = 0$.

While Definition 1 is agnostic to the (causal) generative process that gave rise to the data (note the absence of a reference SCM $\mathcal{M}$ from Definition 1), Definition 2 has the conceptual advantage that it 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 Definitions 1 and 2 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 both from causal (Kusner et al., 2017; Zhang and Bareinboim, 2018b,a; Chiappa, 2019) and non-causal (Dwork et al., 2012) perspectives that fairness is fundamentally a concept at the level of the individual. 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. Put differently, group-level definitions of fairness still allow for unfairness at the individual level, provided that positive and negative discrimination cancel out across the group. This is also a motivation behind counterfactual fairness (Kusner et al., 2017): a decision is considered fair at the individual level if it would not have changed, had the individual belonged to a different protected group. Next, we apply these ideas to fairness of recourse.

3.2 Individually fair causal recourse

Inspired by counterfactual fairness (Kusner et al., 2017), 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 (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*

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

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

It is possible to satisfy both group IMF-fair (Definition 1) and group CAU-fair recourse (Definition 2), without satisfying individually CAU-fair recourse.

Proposition 4. *Neither of the group-level notions of fair recourse (Definitions 1 and 2) are sufficient conditions for individually* CAU*-fair recourse (Definition 3), i.e.,*

$$\textit{Group IMF-fair} \not\Longrightarrow \textit{Individually CAU-fair}$$
$$\textit{Group CAU-fair} \not\Longrightarrow \textit{Individually CAU-fair}.$$

The proof is given by the following counterexample.

Example 5 (Group-level fair, but individually unfair)**.** *Consider the following simple* SCM $\mathcal{M}$*:*

$$\begin{aligned} A &:= U_A, & U_A &\sim \text{Bernoulli}(0.5), \\ X &:= AU_X + (1-A)(1-U_X), & U_X &\sim \text{Bernoulli}(0.5), \end{aligned}$$

with $Y := \text{sign}(X - 0.5) =: h(X)$*. We have* $\mathbb{P}_{X|A=0} = \mathbb{P}_{X|A=1} = \text{Bernoulli}(0.5)$*, so the distance to the decision boundary at* $X = 0.5$ *is the same across both subgroups. The criterion for "group IMF-fair" recourse (Definition 1) is therefore satisfied, and, since there is only one actionable feature* X *without descendants, recourse is also "group CAU-fair" (Definition 2). However, for all* $\mathbf{v}^F = (x^F, a^F)$ *and any* $a \neq a^F$*, we have* $h(x^F) \neq h(x_a(u_X^F)) = 1 - h(x^F)$*, 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 Example 5 is *not* counterfactually fair. This suggests to investigate the relationship between counterfactual fairness and individually CAU-fair recourse: *does a counterfactually fair classifier imply fair (causal) recourse?*

Proposition 6. *Counterfactual fairness of* h *is not sufficient for any of the three notions of fair recourse , i.e.,*

$$\begin{aligned} h \textit{ counterfactually fair} &\not\Rightarrow \textit{Group IMF-fair} \\ h \textit{ counterfactually fair} &\not\Rightarrow \textit{Group CAU-fair} \\ h \textit{ counterfactually fair} &\not\Rightarrow \textit{Individually CAU-fair}. \end{aligned}$$

The above is proven by the following counter-example.

Example 7 (Counterfactually fair classifier, but individually unequal cost of recourse)**.** *Consider the following* SCM*,*

$$\begin{aligned} A &:= U_A, & U_A &\sim \textit{Bernoulli}(0.5), \\ X &:= (2-A)U_X, & U_X &\sim \mathcal{N}(0,1), \end{aligned}$$

with $Y := \text{sign}(X) =: h(X)$*, which was used to generate the data shown in Figure 1. Then* $h(X) = \text{sign}(X) = \text{sign}(U_X)$*, and hence* h *is counterfactually fair as* U_X *is assumed fixed when counterfactually reasoning about a change of the protected attribute. 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 decision boundary differs significantly both at the group level and when counterfactually changing* A*: specifically, the cost of recourse for members of* G_0^- *is twice that of members of* G_1^-*.*

Remark 8. *An important characteristic of Example 7 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.*

4 Achieving fair causal recourse algorithmically

We now discuss approaches for achieving fair causal recourse algorithmically by altering the underlying classifier.

4.1 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. (2019). This has the advantage that a potential trade-off between accuracy and fairness can be controlled with a hyperparameter. However, the causal recourse optimisation problem in (3) involves optimising over the combinatorial space of intervention targets $\mathcal{I} \subseteq \{1, ..., n\}$, so it is not clear whether fairness of causal recourse may easily be included as a differentiable constraint when training a classifier.

4.2 Restricting the Classifier Inputs

An approach to achieve fair causal recourse that only requires *qualitative* knowledge in form of the causal graph (but not a fully-specified SCM), is to a priori restrict the set of features to which the classifier has access 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:

Proposition 9. *Assume h only depends on a subset $\tilde{\mathbf{X}}$ which are non-descendants of A in $\mathcal{M}$, $\tilde{\mathbf{X}} \subseteq \mathbf{V} \setminus (A \cup \text{desc}(A))$; and that the set of feasible actions and their cost remain the same under a counterfactual change of A, $\mathcal{F}(\mathbf{v}^{\text{F}}) = \mathcal{F}(\mathbf{v}_a(\mathbf{u}^{\text{F}}))$ and $c(\cdot\,; \mathbf{v}^{\text{F}}) = c(\cdot\,; \mathbf{v}_a(\mathbf{u}^{\text{F}}))\ \forall a \in \mathcal{A}, \mathbf{v}^{\text{F}} \in \mathcal{D}$. Then $(\mathcal{D}, h, c, \mathcal{M})$ is "individually CAU-fair".*

Proof. According to Definition 3, it suffices to show that $r^{\text{CAU}}(\mathbf{v}^{\text{F}}) = r^{\text{CAU}}(\mathbf{v}_a(\mathbf{u}^{\text{F}}))\ \forall a \in \mathcal{A}, \mathbf{v}^{\text{F}} \in \mathcal{D}$. Substituting our assumptions in the definition of r^{CAU} from §3.1, we get:

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

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 (Pearl, 2009) since $\tilde{\mathbf{X}}$ does not contain any descendants of A by assumption, and is thus not influenced by counterfactual changes to A. □

Remark 10. *Relying exclusively on non-descendants of the protected attribute is also sufficient to ensure counterfactual fairness of h, see Lemma 1 of Kusner et al. (2017).*

The assumption of Proposition 9 that both the set of feasible actions $\mathcal{F}(\mathbf{v}^{\text{F}})$ and the cost function $c(\cdot\,; \mathbf{v}^{\text{F}})$ 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.

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 Proposition 9, can therefore lead to a substantial drop in accuracy which can be a strong restriction in practice.

4.3 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 SCM-background 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, since $\mathbf{U}$ is not observed, it first needs to be inferred from the observed $\mathbf{v}^{\text{F}}$, 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, c.f. the discussion of this point by Kusner et al. (2017, § 4.1).

Note that our notions of fair causal recourse are defined in reference to $(\mathcal{D}, h, c, \mathcal{M})$. In this section we have discussed different ways of changing the classifier h for overcoming unfair recourse. Next, we consider instead the possibility of altering the underlying data-generating process, as captured by $\mathcal{M}$ and manifested in the form of the observed data $\mathcal{D}$, as a viable alternative towards fair recourse.

5 On Societal interventions

In fair ML, typically we attempt to enforce a notion of fairness by requiring a learned classifier to satisfy some constraint which implicitly places the cost of an intervention on the deployer. For example, a bank might need to modify their classifier so as to offer loans to some individuals who would not otherwise receive them. Another possibility is to suggest an intervention to a customer to allow them to change their outcome, e.g., per (3). In our approach, we are already explicitly modelling the cost to an individual of adjusting their own properties in response to a fixed classifier, h. We suggest another perspective is to consider how best to absorb the "costs" of fairness across different agents such that, as a society, we best enjoy the associated "benefits." The bank may not be the right stakeholder to absorb all the costs of societal disparities: (i) it may not be a 'fair' allocation; and (ii) it may not create good incentives across society to lead to desirable outcomes.

Our causal approach here is perhaps particularly well suited to exploring this perspective. Such societal interventions may manifest themselves as changes to the SCM which result in causally-fair recourse across subgroups. 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

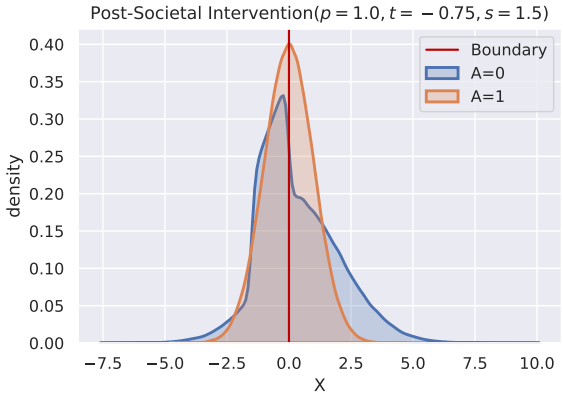

Figure 3: Distribution after applying a societal intervention to the setting from Figure 1. We randomly select a proportion $p = 1$ of individuals from the disadvantaged group ($A = 0$) to receive a *subsidy* $s = 1.5$ if U_X is below the *threshold* $t = -0.75$. As a result, the distribution of negatively-classified individuals ($X < 0$) shifts towards the boundary which makes it more similar to those in $A = 1$ resulting in more fair recourse, whereas the distribution of positively-classified individuals remains unchanged.

individual level). This is akin to mechanism design in game theory, asking the question which societal intervention (e.g., supporting one of the subgroups) would enable fairer recourse, similar to Kusner et al. (2019), but for recourse.

As an example of such societal interventions that attempt to lessen the cost of recourse for specific subgroups, consider the selection process for a job or research grant, where work experience (since the PhD) is often an important factor in the evaluation. In such scenarios, women, or more specifically, mothers (the protected group) are at a disadvantage. A common societal intervention in this scenario is to alter the mechanism through which work experience is counted, e.g., by taking pregnancies and maternity leave into account.

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 awarded a loan, over time, an individual might not be able to repay the loan and this could imply costs to them, to the bank, or to society. Leveraging the economics literature which studies the effect of policy interventions on society, institutions, and individuals (Heckman and Vytlacil, 2005; Heckman, 2010), future work could formalise the effect of these interventions to the SCM. Such a framework can help trade off the costs and benefits for individuals, companies and society.

5.1 Formalism & simple demonstration

The role of societal interventions in overcoming (algorithmic) discrimination is a complex topic which does not only apply to fair recourse but also to other notions of fairness, and it deserves further study outside the scope of the present work. Nevertheless, we attempt to provide a formalism to study this idea and provide a simple toy demonstration for how it may be applied in the context of fair recourse.

Our framework consists of three main elements. First, let $\{i_1, ..., i_K\}$ be a set of societal interventions where each i_k represents a change to the original SCM $\mathcal{M}$ which gives rise to a new SCM $\mathcal{M}'_k = i_k(\mathcal{M})$, thus capturing the way in which i_k modifies the original data generating process. For example, i_k may introduce additional variables or modify a subset of the original structural equations. Secondly, we associate with each i_k a non-negative scalar c_k which measures the societal cost of implementing i_k (relative to a given sample size). For example, in the case of giving subsidies or tax-cuts to underpriviliged individuals, c_k may represent the (expected) amount of money spent by the government or institution. Thirdly, we require a metric m which measures the effectiveness or level of success m_k of societal intervention i_k at achieving its goal. The choice of m will depend on the context, e.g., it may measure a trade-off between accuracy, fairness of prediction, and fairness of recourse.

With these three elements, we may reason about different societal interventions i_k by first simulating the proposed change via sampling data from the resulting SCM $\mathcal{M}'_k$. We then compute the corresponding metric m_k based on simulated data and form the difference to its value m_0 under the original generative process. To decide which societal intervention to implement, we may then compare the societal benefit $(m_k - m_0)$ and cost (c_k) of i_k for different k and choose the one with the most favourable trade-off.

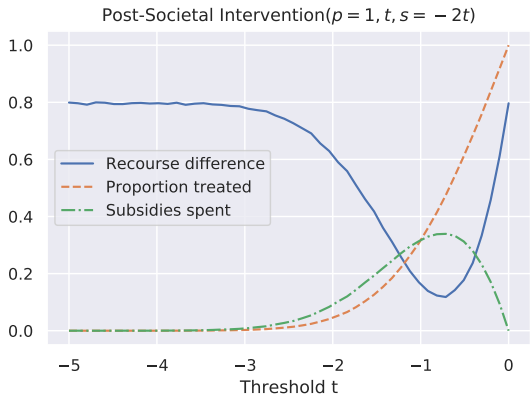

Figure 4: Comparison of different societal interventions $i_k = (1, t, -2t)$ with respect to their benefit (reduction in recourse difference) and cost (paid-out subsidies). The threshold $t \approx 0.75$ with resulting distribution shown in Figure 3 leads to the largest reduction in recourse difference, but also incurs the highest cost. Smaller reductions can be achieved using two different thresholds with one corresponding to giving a larger subsidy to fewer individuals and the other to giving a smaller subsidy to more individuals.

Example 7 Revisited. To make things more concrete, we now demonstrate our ideas for the simple toy Example 7 with different variances across groups as shown in Figure 1. Here, the difference in recourse cost across groups cannot easily be resolved by changing the classifier $h(X)$ (e.g., per the techniques discussed in §4) without substantially altering the distribution of predictions: to achieve perfectly fair recourse, we would have to use a constant classifier which would severely reduce accuracy and is thus undesirable. 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. Specifically, we consider paying subsidies to particular eligible individuals. To this end, we introduce a new treatment variable T which randomly selects a proportion $0 \leq p \leq 1$ of individuals from $A = 0$. Selected individuals are then awarded a subsidy s if their latent variable U_X is below a threshold t, as captured by the modified structural equations

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

Here, each societal intervention i_k corresponds to a particular way of setting the triple (p, t, s). To avoid changing the predictions $Y = \text{sgn}(X)$, 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 Figure 3, see the caption for details. Since we are interested in fairness of recourse and since i_k are designed such that the predicted labels remain unchanged, we can choose, e.g., the reduction in average difference in recourse cost across groups as our metric m. Moreover, the cost c_k of implementing i_k can reasonably be chosen as the total amount of paid-out subsidies. We show these two quantities for $i_k = (1, t, -2t)$ with varying t in Figure 4 and refer to the caption for further details. Plots similar to Figures 3 and 4 for different choices of (p, t, s) are shown in Figure 6 in B.1.

Note that it also seems important to examine how the societal cost c_k should be shared across different people and institutions. In particular, if X in our example represents some sort of 'ability' then arguably the particularly able (perhaps the extreme positive blue members) might bear more burden—though we recognise there are many complexities involved (e.g., this might reduce their incentive to work hard) and therefore leave this for future work.

We also remark that, 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) may provide a useful way to inform whom to target with treatments/societal interventions as it can be used to identify individuals for whom the counterfactual difference in recourse cost is particularly high.

6 Experiments

We now validate our main claims empirically. We describe the main aspects of our experimental setup and refer to Appendix A for further details and B for additional results.

Data Since computing recourse actions in the general case requires knowledge (or estimation) of the true SCM, we compare different approaches for fair recourse in the following controlled settings with synthetically generated data:

- Independently-manipulable features (IMF): the setting underlying counterfactual explanations and the recourse approaches of Ustun et al. (2019) and Gupta et al. (2019); features do not causally depend on each other, but may depend on the protected attribute A, c.f. Figure 2a.
- Causally-dependent features (CAU): features may causally depend on each other and on the protected attribute A, c.f. Figure 2b. We use $\{X_i := f_i(A, \text{PA}_i) + U_i\}_{i=1}^n$ 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. The ground-truth labels y^i which are used to train different classifiers are sampled as $Y^i \sim \text{Bernoulli}(h(\mathbf{x}^i))$ where $h(\mathbf{x}^i)$ is a linear or nonlinear logistic regression, independently of A.

Classifiers On each data set, we train several ("fair") classifiers. For ease of comparison with Gupta et al. (2019), we consider different support vector machines (SVMs) (Schölkopf and Smola, 2002) trained on varying input sets:

- SVM$(\mathbf{X}, A)$: trained on all features (*naïve baseline*);
- SVM$(\mathbf{X})$: trained on non-protected features $\mathbf{X}$ (*unaware*);
- FairSVM$(\mathbf{X}, A)$: the method of Gupta et al. (2019), designed to equalise the average distance to the decision boundary across different protected groups;
- SVM$(\mathbf{X}_{\text{nd}(A)})$: trained only on features which are non-descendants of the protected attribute, see §4.2;
- SVM$(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$: trained on non-descendants of A and on the unobserved variables $\mathbf{U}_{\text{d}(A)}$ corresponding to features $\mathbf{X}_{\text{d}(A)}$ which are descendants of A, see §4.3.

To make distances comparable across classifiers, we use either a linear or polynomial kernel for all SVMs (depending on the ground truth labels) and select all remaining hyperparameters (including the trade-off parameter λ for FairSVM) using 5-fold cross validation. Results for also selecting the kernel by cross-validation are given in Table 3 in B.3.

Solving the Causal Recourse Optimisation Problem We treat A and all U_i as non-actionable and all X_i as actionable. We discretise the space of feasible actions, compute their effect using the true oracle SCM $\mathcal{M}^\star$, and select the valid action with lowest cost. Results using a linear or nonlinear estimate of $\mathcal{M}^\star$ instead are included in Tables 2 and 3 in Appendix B.2; the trends are largely the same as for $\mathcal{M}^\star$.

Metrics We report (a) accuracy (**Acc**) on a held out test set of size 3000; and (b) the fairness measure of Gupta et al. (2019) from Definition 1 (Δ_{dist}), our group-level fairness measure from Definition 2 (Δ_{cost}), and our individual level measure from Definition 3 (Δ_{ind}). For the fairness metrics (b), we select 50 negatively classified individuals from each protected group and report the difference in group-wise means (Δ_{dist} and Δ_{cost}) or the maximum difference over all 100 individuals (Δ_{ind}). To facilitate a comparison between the different SVMs, Δ_{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 $\mathbf{x}^{\text{F}}_{\mathcal{I}}$.

6.1 Results

The results are shown in Table 1. We find that the naïve and unaware baselines SVM$(\mathbf{X}, A)$ and SVM$(\mathbf{X})$ generally exhibit high accuracy and rather poor performance in terms of fairness metrics, though they achieve surprisingly low Δ_{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 (Dwork et al., 2012; Kilbertus et al., 2017).

The FairSVM generally performs well in terms of Δ_{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 SVMs achieve $\Delta_{\text{ind}} = 0$ throughout as expected per Proposition 9, and they are the only methods to do so. In the case of SVM$(\mathbf{X}_{\text{nd}(A)})$ this comes at a substantial drop in accuracy, as discussed in §4.2, whereas SVM$(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$ 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. §4.3.

Table 1: Experimental results for the setting described in §6; the best performing method is highlighted in **bold**.

Classifier	**Ground truth labels from *linear* logistic regression → using *linear* kernel**												**Ground truth labels from *nonlinear* logistic regression → using *polynomial* kernel**											
	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**	$\Delta_{\textbf{dist}}$	$\Delta_{\textbf{cost}}$	$\Delta_{\textbf{ind}}$
$\text{SVM}(\mathbf{X}, A)$	**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
$\text{SVM}(\mathbf{X})$	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
$\text{FairSVM}(\mathbf{X}, A)$	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**
$\text{SVM}(\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)})$	**86.5**	0.96	0.60	**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**	90.0	0.15	0.12	**0.00**

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 Proposition 4, and, empirically, the converse does not hold either.

7 Conclusion

Fairness and appropriate ways to address recourse are rapidly gaining major significance as data-driven decision systems pervade our societies. However, there is still much progress to be made in identifying and understanding the best conceptual paths forward. We are glad in this work to make progress by applying tools of graphical causality, and are hopeful that this approach will continue to be fruitful for coming up with the right concepts and definitions, as well as for assaying interventions on societal mechanisms.

Here, we 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 features. To address this limitation, we introduced new causal notions of fair recourse at the group- and individual level and showed that they are complementary to fairness of prediction.

Our fairness criteria may help assess the fairness of recourse more faithfully, but it is still unclear how best to achieve fair causal recourse algorithmically. We believe 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.

Acknowledgements

We thank Adrián Javaloy Bornás and the anonymous reviewers of the NeurIPS2020 Workshop "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)" 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.

References

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

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

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.

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

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

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

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

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.

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, 2020a.

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*, 2020b.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *arXiv preprint arXiv:2002.06278*, 2020c.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv preprint arXiv:2006.06831*, 2020d.

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.

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.

Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*, pages 3591–3600, 2019.

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

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

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

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.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

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.

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

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.

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

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.

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.

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.

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.

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.

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, 2017a.

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, 2017b.

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

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

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

A Experimental Details

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

A.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 Figure 5. 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 SVM($\mathbf{X}_{\text{nd}(A)}$) and 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.

A.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.1.2 CAU-LIN

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

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

A.1.3 CAU-ANM

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

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

A.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

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

(a) IMF (b) CAU

Figure 5: Causal graphs used to generate synthetic data for our experiments.

A.3 SVM Hyper-parameters

We use the implementation of Gupta et al. (2019) for the FairSVM and the `sklearn SVC` class (Pedregosa et al., 2011) for all other SVM variants. We consider the following values of hyperparameters (which are the same as those reported by Gupta et al. 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 Gupta et al. (2019), we also pick the fairness trade-off parameter $\lambda = \{0.2, 0.5, 1, 2, 10, 50, 100\}$ by cross-validation.

A.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 Karimi et al. (2020d).

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 $[x_i^{\text{F}} - 2(x_i^{\text{F}} - \min_i), x_i^{\text{F}} + 2(\max_i - x_i^{\text{F}})]$. We then consider all possible combinations of intervention values over all subsets $\mathcal{I}$ of the actionable variables. We note that for SVM($\mathbf{X}_{\text{nd}(A)}$) and SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$), only X_2 is actionable, while for the other SVMs all of $\{X_1, X_2, X_3\}$ are actionable.

B Additional Results

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

B.1 Additional Societal Interventions

In §5.1, 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 Figures 3 and 4 for different choices of (p, t, s) in Figure 6 in B.1.

B.2 Using SCM Estimates for Recourse

The results presented in the main paper assume access to the ground truth SCM $\mathcal{M}^\star$ to solve the recourse optimisation problem. In practice, this is unrealistic and the SCM instead needs to be learnt from data and assumptions based on domain knowledge (Peters et al., 2017; Karimi et al., 2020d).

In Table 2 we show a more complete account of results which also includes the cases where a linear ($\hat{\mathcal{M}}_{\text{LIN}}$) or nonlinear kernel-ridge ($\hat{\mathcal{M}}_{\text{KR}}$) regression is used to estimate the SCM and 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 an estimate of the SCM does not affect **Acc** or $\Delta_{\textbf{dist}}$ since these metrics are, by definition, agnostic to the underlying 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. Definitions 2 and 3.

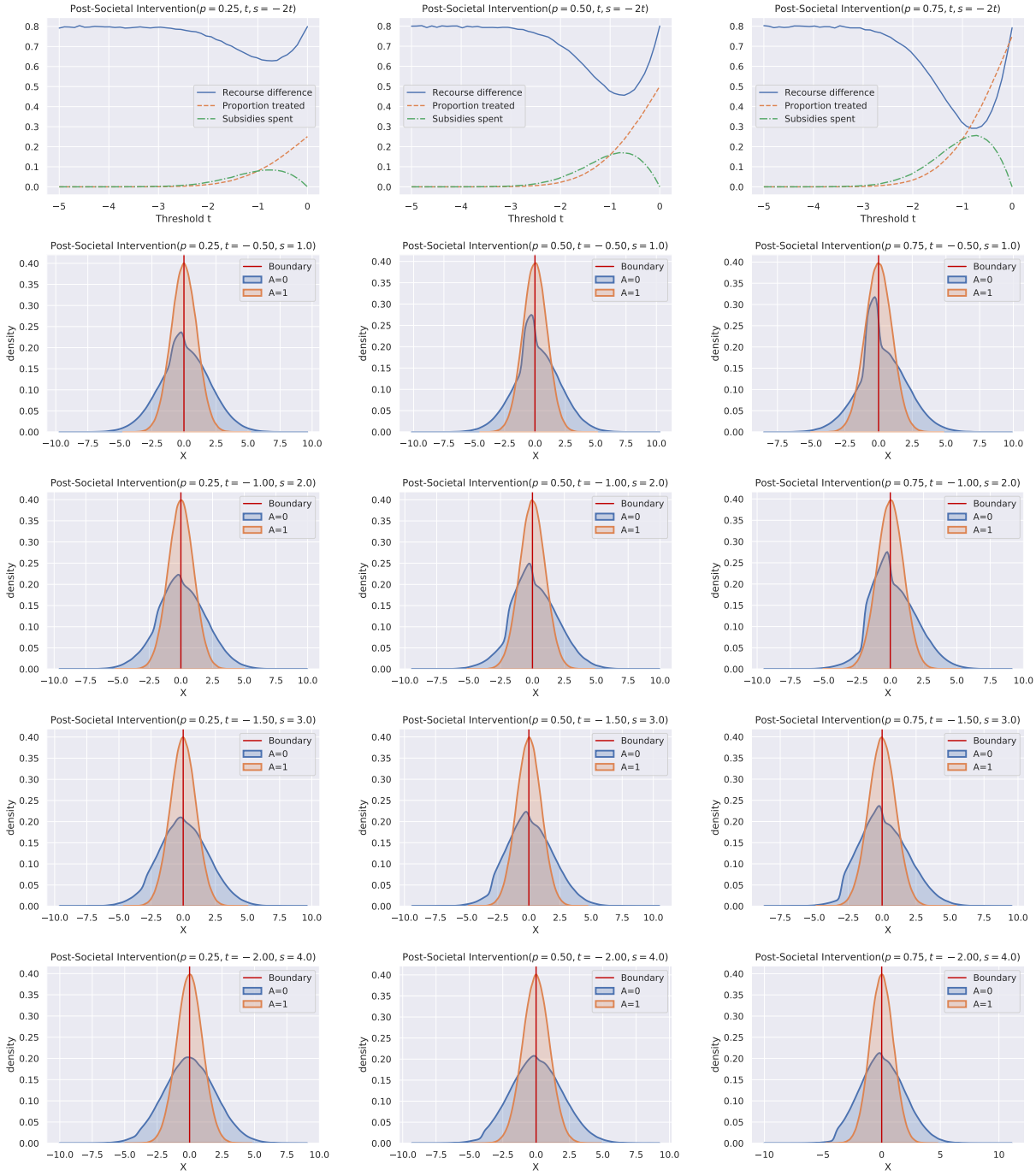

Figure 6: Plots for additional societal interventions $i_k = (p, t, s)$ in the context of Example 7 as discussed in §5.1. We consider budgets of $p = 0.25$ (left column), $p = 0.5$ (middle column), and $p = 0.75$ (right column). In the top row, we show the difference across groups in the average distance 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.

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 compared with using the ground truth SCM, the overall trends reported in §6 remain very much the same, and thus seem relatively robust to estimation errors in the SCM which is used to compute recourse actions.

B.3 Kernel selection by cross validation

For completeness, we also perform the same set of experiments shown in Table 2 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 Table 3 and the overall trends, again, remain largely the same.

As expected, we observe variations in accuracy compared to table 2 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 Table 3 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 §6.

Table 2: **Complete account of experimental results corresponding to the setting described in §6 of the main paper, where we additionally consider using a linear ($\hat{\mathcal{M}}_{\textbf{LIN}}$) or nonlinear ($\hat{\mathcal{M}}_{\textbf{KR}}$) estimate of the true SCM $\mathcal{M}^\star$ 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 across different classifiers, 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. 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 ($\Delta_{\textbf{dist}}$ and $\Delta_{\textbf{cost}}$) and maximum individual unfairness. 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 $\mathcal{M}^\star$. For each experiment and metric, the best performing method is highlighted in **bold**.

SCM	**Classifier**	**Labels from *linear* logistic regression → using *linear* kernel**												**Labels from *nonlinear* logistic regression → using *polynomial* kernel**											
		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	89.5	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
	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
	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
	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**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.5**	0.96	0.60	**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**	90.0	0.15	0.12	**0.00**
$\hat{\mathcal{M}}_{\text{LIN}}$	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
	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
	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
	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**
	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**	88.0	0.21	0.14	**0.00**
$\hat{\mathcal{M}}_{\text{KR}}$	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
	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	0.09	1.09	**91.0**	**0.02**	0.03	1.64
	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($\mathbf{X}_{\text{nd}(A)}$)	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**
	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**	88.0	0.21	0.14	**0.00**

Table 3: **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. 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. Apart from the ground truth SCM $\mathcal{M}^\star$ (as shown in Table 1 in the main paper), we also consider a linear ($\hat{\mathcal{M}}_{\text{LIN}}$) or nonlinear ($\hat{\mathcal{M}}_{\text{KR}}$) estimate of the true SCM for inferring the latent variables $\mathbf{U}$ and solving the causal recourse optimisation problem. 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 size 3000. 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 ($\Delta_{\textbf{dist}}$ and $\Delta_{\textbf{cost}}$) and maximum individual unfairness. 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 $\mathcal{M}^\star$. For each experiment and metric, the best performing method is highlighted in **bold**.

SCM	**Classifier**	**Labels from *linear* logistic regression → kernel selected by cross-validation**												**Labels from *nonlinear* logistic regression → kernel selected by cross-validation**											
		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	**0.09**	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
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.12	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
	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	**0.01**	0.04	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.07	**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**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.5**	0.96	0.60	**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**
$\hat{\mathcal{M}}_{\text{LIN}}$	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
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.12	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
	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	**0.01**	**0.03**	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.07	**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**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.5**	0.96	0.58	**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**
$\hat{\mathcal{M}}_{\text{KR}}$	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
	SVM($\mathbf{X}$)	86.4	0.99	0.42	1.80	**89.5**	1.12	0.52	2.14	87.6	0.43	0.44	2.79	**91.4**	0.13	**0.00**	0.92	91.0	0.17	0.09	1.09	89.4	0.16	0.14	1.16
	FairSVM($\mathbf{X}, A$)	86.4	**0.01**	0.20	1.61	60.5	**0.00**	0.26	1.05	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	**0.01**	0.01	1.73
	SVM($\mathbf{X}_{\text{nd}(A)}$)	64.6	0.07	**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**
	SVM($\mathbf{X}_{\text{nd}(A)}, \mathbf{U}_{\text{d}(A)}$)	**86.5**	0.96	0.58	**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**