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, Germany
[2] University of Cambridge
[3] ETH Zürich
[4] Saarland University
[5] The Alan Turing Institute

jvk@tue.mpg.de, amir@tue.mpg.de, usb20@cam.ac.uk, ivalera@cs.uni-saarland.de, aw665@cam.ac.uk, bs@tue.mpg.de

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 (Dwork et al. 2012; Zemel et al. 2013; Hardt, Price, and Srebro 2016; Chouldechova 2017). Given a dataset comprising individuals from multiple legally protected groups (defined, e.g., 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 a pre-defined (statistical or causal) criterion, 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, to overcome their adverse situation (Joshi et al. 2019; Ustun, Spangher, and Liu 2019; Sharma, Henderson, and Ghosh 2019; Mahajan, Tan, and Sharma 2019; Mothilal, Sharma, and Tan 2020; Venkatasubramanian and Alfano 2020; Karimi et al. 2020c,b; Karimi, Schölkopf, and Valera 2021; Upadhyay, Joshi, and Lakkaraju 2021). 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*.

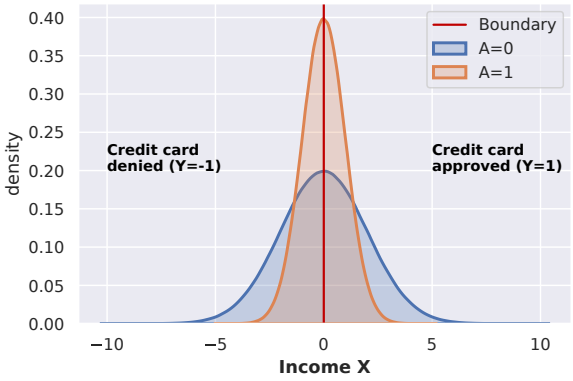


Figure 1: Example demonstrating the difference between fair *prediction* and fair *recourse*: here, only the variance of (centered income) X differs across two protected groups $A \in \{0, 1\}$, while the true and predicted label (whether an individual is approved for a credit card) are determined by $\text{sign}(X)$. This scenario would be considered fair from the perspective of *prediction*, but the cost of *recourse* (here, the distance to the decision boundary, set at $X = 0$) is much larger for individuals in the blue group with $A = 0$.

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 (flipsets) suggested to otherwise similar male and female individuals (Ustun, Spangher, and Liu 2019), or comparing the aggregated cost of recourse (burden) across different protected groups (Sharma, Henderson, and Ghosh 2019)—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* (Karimi et al. 2020a). 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) (Barocas, Selbst, and Raghavan 2020). While this scenario satisfies 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 (Arneson 2015).[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. (2019) 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 (Mahajan, Tan, and Sharma 2019; Karimi, Schölkopf, and Valera 2021; Karimi et al. 2020c; Mothilal, Sharma, and Tan 2020), cf. also (Barocas, Selbst, and Raghavan 2020; Wachter, Mittelstadt, and Russell 2017; Ustun, Spangher, and Liu 2019). By not reasoning about causal relations between features, the distance-based approach (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 to the minimal intervention-based framework of recourse (Karimi, Schölkopf, and Valera 2021), we introduce *causal* notions of fair recourse which capture the true differences in recourse cost more faithfully if features are not independently manipulable, 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 highlight 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 (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 which do not imply each other (Prop. 3.4);
- we establish sufficient conditions that allow for individually-fair causal recourse (Prop. 3.6);
- we evaluate different fair recourse metrics for several classifiers (§ 4.1), verify our main results, and demonstrate that non-causal metrics misrepresent recourse unfairness;
- in a case study on the Adult dataset, we detect recourse discrimination at the group and individual level (§ 4.2), demonstrating its relevance for real world settings;
- we propose societal interventions as an alternative to altering a classifier to address unfairness (§ 5).

2 Preliminaries & Background

Notation. Let the random vector $\mathbf{X} = (X_1, ..., X_n)$ taking values $\mathbf{x} = (x_1, ..., x_n) \in \mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_n \subseteq \mathbb{R}^n$ denote observed (non-protected) features. Let the random variable A taking values $a \in \mathcal{A} = \{1, \dots, K\}$ for some $K \in \mathbb{Z}_{>1}$ denote a (legally) protected attribute/feature indicating which group each individual belongs to (based, e.g., on her age, sex, ethnicity, religion, etc). And let $h : \mathcal{X} \to \mathcal{Y}$ be a *given* binary classifier with $Y \in \mathcal{Y} = \{\pm 1\}$ 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 variable $\mathbf{V} = (\mathbf{X}, A)$ with $\mathbf{v}^i := (\mathbf{x}^i, a^i)$.[3]

Counterfactual Explanations. A common framework for explaining decisions made by (black-box) ML models is that of counterfactual explanations (CE; Wachter, Mittelstadt, and Russell 2017). 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:

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

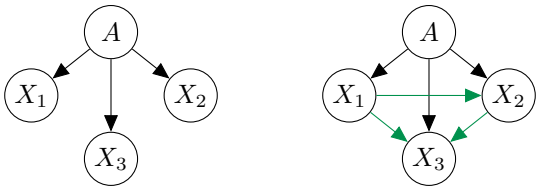
While CEs are useful to *understand the behaviour of a classifier*, they do not generally lead to *actionable recommendations*: they inform an individual of where she should be to obtain a more favourable prediction, but they may not suggest *feasible* changes she could perform to get there.

Recourse with Independently-Manipulable Features. Ustun, Spangher, and Liu (2019) refer to a person's ability to change the decision of a model by altering actionable

[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

[2] This differs from the commonly-used purely predictive, statistical criterion of equal opportunity (Hardt, Price, and Srebro 2016).

[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).



(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 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. It also provides a framework for studying alternative routes to achieve fair recourse beyond changing the classifier.

variables as *recourse* and propose to solve

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

where $\mathcal{F}(\mathbf{x}^{\mathrm{F}})$ is a set of feasible change vectors and $c(\cdot; \mathbf{x}^{\mathrm{F}})$ is a cost function defined over these actions, both of which may depend on the individual.[4] As pointed out by Karimi, Schölkopf, and Valera (2021), (2) implicitly treats features as manipulable independently of each other (see Fig. 2a) and does not account for causal relations that may exist between them (see Fig. 2b): while allowing feasibility constraints on actions, variables which are not acted-upon ($\delta_i = 0$) are assumed to remain unchanged. We refer to this as the *independently-manipulable features* (IMF) assumption. While the IMF-view may be appropriate when only analysing the behaviour of a classifier, it falls short of capturing effects of interventions performed in the real world, as is the case in actionable recourse; e.g., an increase in income will likely also positively affect the individual's savings balance. As a consequence, (2) only guarantees recourse if the acted-upon variables have no causal effect on the remaining variables (Karimi, Schölkopf, and Valera 2021).

Structural Causal Models. A structural causal model (SCM) (Pearl 2009; Peters, Janzing, and Schölkopf 2017) over 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 $\mathbf{S} = \{V_i := f_i(\mathrm{PA}_i, U_i)\}_{i=1}^n$, which compute each V_i as a deterministic function f_i of its direct causes (causal parents) $\mathrm{PA}_i \subseteq \mathbf{V} \setminus V_i$ and an unobserved variable U_i. In this work, we make the common assumption that the distribution $\mathbb{P}_{\mathbf{U}}$ factorises over the latent $\mathbf{U} = \{U_i\}_{i=1}^n$, meaning that there is no unobserved confounding (causal sufficiency). If the causal graph $\mathcal{G}$ associated with $\mathcal{M}$ (obtained by drawing a directed edge from each variable in PA_i to V_i, see Fig. 2 for examples) is acyclic, $\mathcal{M}$ induces a unique "observational" distribution over $\mathbf{V}$, defined as the push forward of $\mathbb{P}_{\mathbf{U}}$ via $\mathbf{S}$.

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 (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 $\{V_i := f_i(\mathrm{PA}_i, U_i)\}_{i \in \mathcal{I}}$ by their new assignments $\{V_i := \theta_i\}_{i \in \mathcal{I}}$ to obtain the modified structural equations $\mathbf{S}^{do(\boldsymbol{\theta}_{\mathcal{I}})}$ and then computing the distribution induced by the interventional SCM $\mathcal{M}^{do(\boldsymbol{\theta}_{\mathcal{I}})} = (\mathbf{S}^{do(\boldsymbol{\theta}_{\mathcal{I}})}, \mathbb{P}_{\mathbf{U}})$, i.e., the push-forward of $\mathbb{P}_{\mathbf{U}}$ via $\mathbf{S}^{do(\boldsymbol{\theta}_{\mathcal{I}})}$.

Similarly, SCMs allow reasoning about (structural) *counterfactuals*: statements about interventions performed in a hypothetical world where all unobserved noise terms $\mathbf{U}$ are kept unchanged and fixed to their factual value $\mathbf{u}^{\mathrm{F}}$. The counterfactual distribution for a hypothetical intervention $do(\boldsymbol{\theta}_{\mathcal{I}})$ given a factual observation $\mathbf{v}^{\mathrm{F}}$, denoted $\mathbf{v}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\mathrm{F}})$, can be obtained from $\mathcal{M}$ using a three step procedure: first, inferring the posterior distribution over the unobserved variables $\mathbb{P}_{\mathbf{U}|\mathbf{v}^{\mathrm{F}}}$ (*abduction*); second, replacing some of the structural equations as in the interventional case (*action*); third, computing the distribution induced by the counterfactual SCM $\mathcal{M}^{do(\boldsymbol{\theta}_{\mathcal{I}})|\mathbf{v}^{\mathrm{F}}} = (\mathbf{S}^{do(\boldsymbol{\theta}_{\mathcal{I}})}, \mathbb{P}_{\mathbf{U}|\mathbf{v}^{\mathrm{F}}})$ (*prediction*).

Causal Recourse. To capture causal relations between features, Karimi, Schölkopf, and Valera (2021) 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 $\boldsymbol{\theta}_{\mathcal{I}}$.[5] 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. 2020c). If the IMF assumptions holds (i.e., the set of descendants of all actionable variables is empty), then (3) reduces to IMF recourse (2) as a special case.

Algorithmic and Counterfactual Fairness. While there are many statistical notions of fairness (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, Malinsky, and Shpitser 2019; Chiappa 2019; Salimi et al. 2019; Wu et al. 2019).

[4] For simplicity, (2) assumes that all X_i are continuous; we do not make this assumption in the remainder of the present work.

[5] For an interventional notion of recourse related to conditional average treatment effects (CATE) for a specific subpopulation, see (Karimi et al. 2020c); in the present work, we focus on the individualised counterfactual notion of causal recourse.

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

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

where $\mathbf{v}_a(\mathbf{u}^{\text{F}})$ denotes the "counterfactual twin" of $\mathbf{v}^{\text{F}}$ 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. (2019). 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 CEs, they define the cost of recourse for $\mathbf{x}^{\text{F}}$ with $h(\mathbf{x}^{\text{F}}) = -1$ as the minimum achieved in (1):

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

which is equivalent to IMF-recourse (2) if $c(\boldsymbol{\delta}; \mathbf{x}^{\text{F}}) = d(\mathbf{x}^{\text{F}} + \boldsymbol{\delta}, \mathbf{x}^{\text{F}})$ 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,

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

The idea of *Equalizing Recourse* across groups (Gupta et al. 2019) can then be summarised as follows.

Definition 2.1 (Group-level fair IMF-recourse, 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:

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

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

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. 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):

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

where we recall that the constraint $h(\mathbf{v}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})) = 1$ ensures that the counterfactual twin of $\mathbf{v}^{\text{F}}$ in $\mathcal{M}$ 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 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:

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

Recourse for $(\mathcal{D}, h, c, \mathcal{M})$ is "group CAU-fair" if $\Delta_{\text{cost}} = 0$.

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 (Chiappa 2019; Wu et al. 2019) and non-causal (Dwork et al. 2012) perspectives that fairness is fundamentally an individual-level concept:[6] 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 (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.

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 3.2 (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

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

Recourse is "individually CAU-fair" if $\Delta_{\text{ind}} = 0$.

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 Defn. 3.2:

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.,*

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

[6] 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

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

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

We have $\mathbb{P}_{X|A=0} = \mathbb{P}_{X|A=1} = \text{Bernoulli}(0.5)$, so the distance to the boundary at $X = 0.5$ is the same across groups. The criterion for "group IMF-fair" recourse (Defn. 2.1) is thus satisfied.

Since protected attributes are generally immutable (thus making any recourse actions involving changes to A infeasible) and since there is only a single feature in this example (so that causal downstream effects on descendant features can be ignored), the distance between the factual and counterfactual value of X is a reasonable choice of cost function also for causal recourse. In this case, $(\mathcal{D}, h, \mathcal{M})$ also satisfies group-level CAU-fair recourse (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: for any individual, 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

The classifier h used 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?* The answer is no.

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

$$\begin{aligned} &h \textit{ counterfactually fair} \implies \textit{Group IMF-fair} \\ &h \textit{ counterfactually fair} \implies \textit{Group CAU-fair} \\ &h \textit{ counterfactually fair} \implies \textit{Individually CAU-fair} \end{aligned}$$

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

$$\begin{aligned} A &:= U_A, & U_A &\sim \text{Bernoulli}(0.5), \\ X &:= (2-A)U_X, & U_X &\sim \mathcal{N}(0,1), \\ Y &:= h(X) = \text{sign}(X) \end{aligned} \tag{6}$$

which we used to generate Fig. 1. 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 (which is a reasonable cost for CAU-recourse in this one-variable toy example) differs at the group level. Moreover, X either doubles or halves when counterfactually changing A. □

Remark 3.5. An important characteristic of the counterexample used in the proof of Prop. 3.4 is that h is *deterministic*, which makes it possible that h is counterfactually fair, even though it depends on a descendant of A. This is generally not be the case if h is *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

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). 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.

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.

Proposition 3.6. *Assume h only depends on a subset $\tilde{\mathbf{X}} \subseteq \mathbf{V} \setminus (A \cup \text{desc}(A))$ which are non-descendants of A in $\mathcal{M}$; 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 recourse for $(\mathcal{D}, h, c, \mathcal{M})$ is "individually CAU-fair".*

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:

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

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. □

The assumption of Prop. 3.6 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.

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 (Wu, Zhang, and Wu 2019).

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 $\mathbf{v}^{\text{F}}$, 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 (Kusner et al. 2017, § 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 A for further experimental details and to Appendix B for additional results and analyses.[7] Code to reproduce our experiments is available at https://github.com/amirhk/recourse.

4.1 Numerical Simulations

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

- IMF: the setting underlying IMF recourse where features do not causally influence each other, but may depend on the protected attribute A.
- CAU: features causally depend on each other and on A. 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.

The corresponding causal graphs are included in Fig.3 of (von Kügelgen et al. 2022). 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 described in more detail in Appendix A.1. The ground truth (GT) labels y^i 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, as detailed in Appendix A.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; Schölkopf and Smola 2002) (for ease of comparison with Gupta et al. (2019)), 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. (2019), designed to equalise the average distance to the decision boundary across different protected groups;
- LR/SVM$(\mathbf{X}_{\text{nd}})$: trained only on features $\mathbf{X}_{\text{nd}(A)}$ which are non-descendants of A, see § 3.4;
- LR/SVM$(\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}})$: trained on non-descendants $\mathbf{X}_{\text{nd}(A)}$ 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 § 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 cross-validation are also provided in Tab. 4 in Appendix B.3. Linear (nonlinear, resp.) LR is used when the GT labels are generated using linear (nonlinear, resp.) logistic regression, as detailed in Appendix A.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 Karimi et al. (2020c), see Appendix B.2 for details), and select the least costly valid action resulting in a favourable 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 B.2; the trends are mostly 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) (Gupta et al. 2019), 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 generally exhibit high accuracy and rather poor performance in terms of fairness metrics, but achieve surprisingly low $\Delta_{\textbf{cost}}$ on some datasets. We observe no clear preference of one baseline over the other, consistent with prior work showing that blindness to protected attributes is not necessarily beneficial for fair *prediction* (Dwork et al. 2012); our results suggest this is also true for fair *recourse*.

FairSVM generally performs well in terms of $\Delta_{\textbf{dist}}$ (which is what it is trained for), especially on the two IMF datasets, and sometimes (though not consistently) outperforms the baselines on the causal fairness metrics. However, this comes at decreased accuracy, particularly on 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 due to access to fewer predictive features (see § 3.4),

[7] All Appendix mentions refer to the arXiv version (von Kügelgen et al. 2022) containing the supplement of this work.

Classifier	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}
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
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
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
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
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
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**
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**
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**
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**
SVM($\mathbf{X}, A$)	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($\mathbf{X}, A$)	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($\mathbf{X}$)	**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($\mathbf{X}$)	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($\mathbf{X}, A$)	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($\mathbf{X}_{\text{nd}}$)	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($\mathbf{X}_{\text{nd}}$)	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($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$)	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($\mathbf{X}_{\text{nd}}, \mathbf{U}_{\text{d}}$)	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 1: Results for our numerical simulations from § 4.1, comparing various classifiers differing mostly in their input sets with respect to accuracy (Acc, higher is better) and different recourse fairness metrics (Δ., lower is better) on a number of synthetic datasets (columns). SVM: support vector machine, LR: logistic regression. The first subtable (first nine rows) corresponds to ground truth labels drawn from a *linear* LR (and a linear kernel is used) and the second subtable to labels from a *nonlinear* LR (and a polynomial kernel is used). The first four rows in each subtable are baselines, the middle row corresponds to the method of Gupta et al. (2019), and the last four rows are methods taking causal structure into account. For each dataset and metric, the best performing methods are highlighted in bold. As can be seen, only our causally-motivated methods (last four rows) achieve individually fair recourse ($\Delta_{\text{ind}} = 0$) throughout.

the latter maintains high accuracy by additionally relying on (the true) $\mathbf{U}_{\text{d}(A)}$ for prediction. Its accuracy should be understood as an upper bound on what is possible while preserving "individually CAU-fair" recourse if abduction is done correctly, see the discussion in § 3.4.

Generally, we observe no clear relationship between the different fairness metrics: e.g., 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* $\Delta_{\textbf{\textit{dist}}}$ *nor small* $\Delta_{\textbf{\textit{cost}}}$ *imply small* $\Delta_{\textbf{\textit{ind}}}$*, consistent with* Prop. 3.3, and, empirically, the converse does not hold either.

Summary of Main Findings from § 4.1: The non-causal metric $\Delta_{\textbf{dist}}$ does not accurately capture recourse unfairness on the CAU-datasets where causal relations are present, thus necessitating our new causal metrics $\Delta_{\textbf{cost}}$ and $\Delta_{\textbf{ind}}$. Methods designed in accordance with Prop. 3.6 indeed guarantee individually fair recourse, and group fairness does not imply individual fairness, as expected per Prop. 3.3.

4.2 Case Study on the Adult Dataset

Data. We use the Adult dataset (Lichman et al. 2013), which consists of 45k$+$ samples without missing data. We process the dataset similarly to Chiappa (2019) and Nabi and Shpitser (2018) and adopt the causal graph assumed therein (see also Fig. 3c of (von Kügelgen et al. 2022)). The eight heterogeneous variables include the three binary protected attributes sex (m=male, f=female), age (binarised as $\mathbb{I}\{\text{age} \geq 38\}$; y=young, o=old), and nationality (Nat; US vs non-US), as well as 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 treat the protected attributes and marital status as non-actionable, and the remaining variables as actionable when searching for recourse actions.

Experimental Setup. We extend the probabilistic framework of Karimi et al. (2020c) to consider causal recourse in the presence of heterogeneous features, see Appendix B.2 for more details. We use a nonlinear LR($\mathbf{X}$) as a classifier (i.e., fairness through unawareness) which attains 78.4% accuracy, and (approximately) solve the recourse optimisation problem (3) using brute force search as in § 4.1. We compute the best recourse actions for 10 (uniformly sampled) negatively predicted individuals from each of the eight different protected groups (all 2^3 combinations of the three protected attributes), as well as for each of their seven counterfactual twins, and evaluate using the same metrics as in § 4.1.

Results. At the group level, 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 be-

	Sex	Age	Nat	MS	Edu	WC	Occ	Hrs	Recourse action	Cost
CF	m	y	US	married	Some Collg.	Private	Sales	32.3	do(Edu: Prof-school, WC: Private)	6.2
CF	m	y	non-US	married	HiSch. Grad	Private	Sales	27.8	do(WC: Self-empl., Hrs: 92.0)	64.2
CF	m	o	US	married	Some Collg./Bachelors	Private	Cleaner	36.2	do(Edu: Prof-school, WC: Private)	5.5
CF	m	o	non-US	married	HiSch. Grad	Private	Sales	30.3	do(WC: Self-empl., Hrs: 92.0)	61.7
CF	f	y	US	married	Some Collg.	Self-empl.	Sales	27.3	do(Hrs: 92.0)	64.7
CF	f	y	non-US	married	HiSch. Grad	Self-empl.	Sales	24.0	do(Edu: Some Collg., WC: Self-empl., Hrs: 92.0)	68.0
CF	f	o	US	married	HiSch./Some Collg.	Private	Sales	28.8	do(Edu: Prof-school, WC: Private)	6.4
F	f	o	non-US	married	HiSch. Grad	W/o pay	Sales	25	do(Hrs: 92.0)	67.0

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

tween *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.

At the individual-level, we find an average difference in recourse cost to the counterfactual twins of 24.32 and a maximum difference ($\Delta_{\textbf{ind}}$) of 61.53. The corresponding individual/factual observation for which this maximum is obtained is summarised along with its seven counterfactual twins in Tab. 2, see the caption for additional analysis.

Summary of Main Findings from § 4.2: Our causal fairness metrics reveal qualitative and quantitative aspects of recourse discrimination at both the group and individual level. In spite of efforts to design classifiers that are predictively fair, recourse unfairness remains a valid concern on real datasets.

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. (2019), 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. (2019), 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.

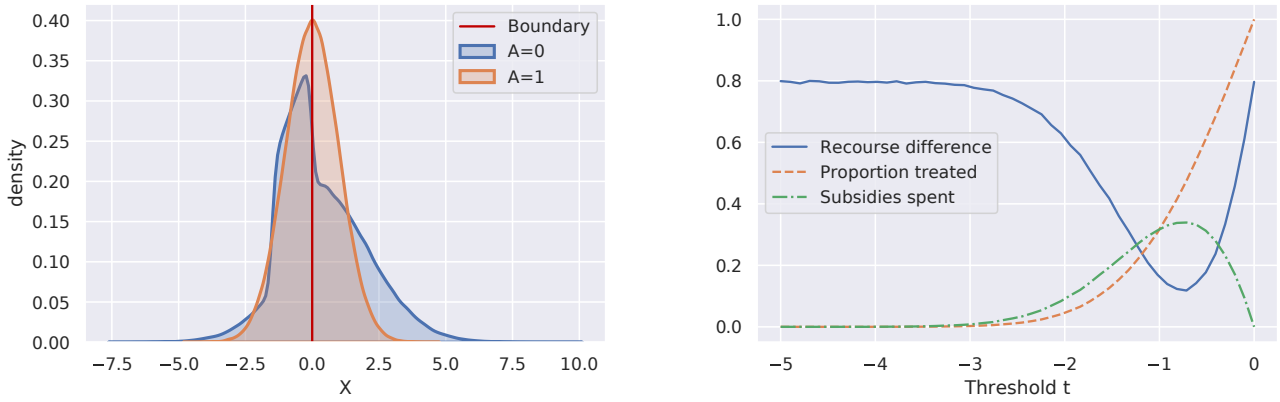
We demonstrate our ideas for the toy example with different variances across groups from Fig. 1. Here, the difference in recourse cost across groups cannot easily be resolved by changing the classifier h (e.g., per the techniques in § 3.4): to achieve perfectly fair recourse, we would have to use a constant 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 data generating process, $X := (2 - A)U_X$, $U_X \sim \mathcal{N}(0, 1)$, 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.[8] This is captured by the modified structural equations

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

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. 3a, see the caption for details.

[8] E.g., for interventions with minimum quantum size and a fixed budget, it makes sense to spread interventions across a *randomly* chosen subset since it is not possible to give everyone a very small amount, see (Grgić-Hlača et al. 2017) for broader comments on the potential benefits of randomness in fairness. Note that $p = 1$, i.e., deterministic interventions are included as a special case.



(a) Post-intervention distribution

(b) Comparison of different societal interventions

Figure 3: (a) Distribution after applying a societal intervention to the credit-card example from Fig. 1. 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. (b) 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.

To evaluate the effectiveness of different societal interventions i_k in reducing recourse unfairness, we 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. 3b and refer to the caption for further details. Plots similar to Fig. 3 for different choices of (p, t, s) are shown in Fig. 5 in Appendix B.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 (Defn. 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.

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, as well as different costs and benefits to individuals, institutions, and society.

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 (Heckman and Vytlacil 2005; Heckman 2010). 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 grateful to Chris Russell for insightful feedback on connections to existing fairness notions within machine learning and philosophy, and to Matthäus Kleindessner, Adrián Javaloy Bornás, and the anonymous reviewers for helpful comments and suggestions. AHK is appreciative of NSERC, CLS, and Google 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

Arneson, R. 2015. Equality of Opportunity. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2015 edition.

Barocas, S.; Selbst, A. D.; and Raghavan, M. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89.

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

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

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K.; and Weller, A. 2017. On Fairness, Diversity, and Randomness in Algorithmic Decision Making. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Gupta, V.; Nokhiz, P.; Roy, C. D.; and Venkatasubramanian, S. 2019. Equalizing Recourse across Groups. *arXiv preprint arXiv:1909.03166*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

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

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

Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.

Karimi, A.-H.; Barthe, G.; Balle, B.; and Valera, I. 2020a. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, 895–905.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020b. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020c. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, 265–277.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076.

Lichman, M.; et al. 2013. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/adult.

Loftus, J. R.; Russell, C.; Kusner, M. J.; and Silva, R. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Mahajan, D.; Tan, C.; and Sharma, A. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *arXiv preprint arXiv:1912.03277*.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

Nabi, R.; Malinsky, D.; and Shpitser, I. 2019. Learning optimal fair policies. In *International Conference on Machine Learning*, 4674–4682. PMLR.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference*. MIT Press.

Russell, C.; Kusner, M. J.; Loftus, J.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, 6414–6423.

Salimi, B.; Rodriguez, L.; Howe, B.; and Suciu, D. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, 793–810.

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

Sharma, S.; Henderson, J.; and Ghosh, J. 2019. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. *arXiv preprint arXiv:1905.07857*.

Upadhyay, S.; Joshi, S.; and Lakkaraju, H. 2021. Towards Robust and Reliable Algorithmic Recourse. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.

Venkatasubramanian, S.; and Alfano, M. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293.

von Kügelgen, J.; Karimi, A.-H.; Bhatt, U.; Valera, I.; Weller, A.; and Schölkopf, B. 2022. On the fairness of causal algorithmic recourse. *arXiv preprint arXiv:2010.06529v5*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2).

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Wu, Y.; Zhang, L.; Wu, X.; and Tong, H. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, 3404–3414.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970. PMLR.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International Conference on Machine Learning*, 325–333.

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

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