On the Fairness of Causal Algorithmic Recourse

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

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

Abstract

While many recent works have studied the problem of algorithmic fairness from the perspective of *predictions*, here we investigate the fairness of *recourse* actions recommended to individuals to recover from an unfavourable classification. To this end, 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—are based on a causal framework that explicitly models relationships between input features, thereby allowing 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 is complementary to fairness of prediction. We then investigate how to enforce fair recourse in the training of the classifier. Finally, we discuss whether fairness violations in the data generating process revealed by our criteria may be better addressed by societal interventions and structural changes to the system, 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 [1–18]. Parallel to this, *algorithmic recourse* is concerned with offering explanations and recommendations to individuals who were unfavorably treated by ML-based decision-making systems to overcome their adverse situation [19–23]. 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. [21] 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 (namely, *flipsets*) require relatively more effort for individuals of a particular sub-group. Along these lines, Sharma et al. [24] evaluate group fairness via aggregating and comparing the cost of recourse (namely, *burden*) over individuals of different sub-populations. Karimi et al. [25] show that the addition of feasibility constraints (e.g., non-decreasing AGE) resulting in an increase in the cost of recourse indicates a reliance of the fixed model on the sensitive attribute AGE, which is often 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, a dataset comprising of two sub-groups (corresponding to the protected attribute $A \in \{0, 1\}$), with feature X distributed as $\mathcal{N}(0,1)$ and $\mathcal{N}(0,10)$, respectively. Moreover, consider a fair binary classifier $h(x,a) = \text{sign}(x)$. While the distribution of predictions satisfies, e.g., demographic parity (and other fairness criteria), the cost of recourse actions required of negatively affected individuals in $A = 1$ is much larger than of those in $A = 0$. This suggest to consider a new and distinct notion of fairness, i.e., *fairness of recourse*, that does not imply or is not implied by the *fairness of prediction*.

In this regard, *Equalising Recourse* was recently presented by Gupta et al. [26], offering the first recourse-based and prediction-independent notion of fairness. In their definition, the authors demonstrate that one can directly calibrate for the average distance to the decision boundary to be equalised across different subgroups during the training of both linear and nonlinear classifiers. This formulation, however, does not take causal relations between variables into account when generating recourse, which has been argued for by a number of authors [19, 21, 23, 25, 27–29], generally based on the intuition that changing some variables may have effects on others.

Our contributions. In this work, we consider the fairness of recourse from a causal perspective, 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. [19] where a causal model of the world in which actions are undertaken is used to generate a set of minimal interventions for recourse, we propose two new definitions for fair recourse. The first is a group-level criterion similar to that of Gupta et al. [26], but considering the average of interventional cost instead of distance across groups. The second is an individualised notion inspired by counterfactual fairness [14]. We show that fairness of recourse is complementary to fairness of prediction, and explore further links to counterfactual fairness. Finally, we investigate different paths towards achieving fair causal recourse and critically discuss these in the context of societal interventions.

2 Preliminaries: explainable AI, recourse, causality, and fairness

Notation Throughout, let $\mathbf{X} = \{X_1, ..., X_d\} \in \mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d \subseteq \mathbb{R}^d$ denote a set of observed (non-protected) features, $A \in \mathcal{A} = \{1, \dots, K\}$ the protected attribute denoting which social salient group each individual belongs to (based, e.g., on her age, sex, race, etc), and $h : \mathcal{X} \to \mathcal{Y}$ a 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 of n i.i.d. observations $\mathcal{D} = \{\mathbf{v}^i\}_{i=1}^n$ where $\mathbf{v}^i = (\mathbf{x}^i, a^i)$.[1]

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

Counterfactual explanations For explaining decisions made by (black-box) machine learning models, Wachter et al. [27] proposed the framework of nearest counterfactual explanations (CEs). Intuitively, a CE is a closest feature vector lying 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 formally defined as a solution to the following optimisation problem:

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

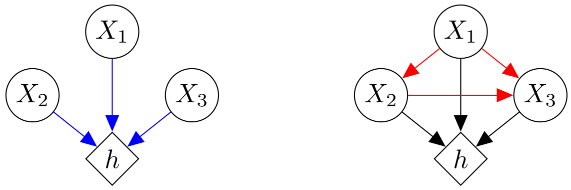
Recourse with independent features CEs are useful to understand the behaviour of a classifier, but they generally do not lead to *actionable recommendations*: while CEs inform an individual of where she should be to obtain a more favourable prediction, they do not suggest *feasible* changes she could perform to get there.[2] Ustun et al. [21] refer to the latter as *recourse* and propose to solve

$$\text{IW-rec}(\mathbf{x}^\text{F}) \in \arg\min_{\boldsymbol{\delta} \in \Delta(\mathbf{x}^\text{F})} \text{cost}(\boldsymbol{\delta}; \mathbf{x}^\text{F}) \qquad \text{subject to} \qquad h(\mathbf{x}^\text{F} + \boldsymbol{\delta}) = 1. \tag{2}$$

where $\Delta(\mathbf{x}^\text{F})$ is a set of feasible change vectors and $\text{cost}(\cdot; \mathbf{x}^\text{F})$ is a cost function defined over these actions, both of which may depend on the individual. As pointed out by Karimi et al. [19], the framework in (2) has the shortcoming that it treats features as manipulable independently of each other (see Figure 1a) and thus fails to capture causal relations that may exist between them (see

[1] We will use $\mathbf{v}$ when there is an explicit distinction between protected attribute and other features, i.e., in the context of fairness considerations, and $\mathbf{x}$ otherwise, i.e., in the context of explainability and (vanilla) recourse.

[2] E.g., a CE may correspond to a feature vector with decreased age, which is infeasible.



(a) IW assumption

(b) Causal view

$$\begin{aligned} X_1 &:= f_1(U_1), \\ X_2 &:= f_2(X_1, U_2), \\ X_3 &:= f_3(X_1, X_2, U_3) \\ \mathbb{P}_{\mathbf{U}} &= \mathbb{P}_{U_1} \times \mathbb{P}_{U_2} \times \mathbb{P}_{U_3} \end{aligned}$$

(c) SCM

Figure 1: (a) The independent world (IW) assumption underlying the frameworks of counterfactual explanations (CEs) [27] and distance-based recourse [21, 26] treats features X_i as independently manipulable inputs to a classifier h. (b) The causal perspective instead considers an underlying causal generative process in the form of asymmetric functional relationships between variables, thus allowing to model downstream effects of changing some features on others. (c) General form of a structural causal model (SCM) over $\{X_1, X_2, X_3\}$ corresponding to the causal graph in (b); here, U_i are latent background variables which are assumed fixed when performing counterfactual reasoning.

Figure 1b). We therefore refer to this view inherent to both CEs and the approach in (2) as *independent world* (IW) assumption. While the IW-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.[3] As a consequence, the IW-recourse approach of (2) only guarantees recourse if the set of variables which are causally dependent on the acted upon variables is empty [19].

Structural causal models (SCMs) An SCM [30, 31] over a set of observed variables $\mathbf{V} = \{V_i\}_{i=1}^d$ is defined as a pair $\mathcal{M} = (\mathbf{S}, \mathbb{P}_{\mathbf{U}})$, where the structural equations $\mathbf{S}$ are a set of assignments

$$\mathbf{S} = \{V_i := f_i(\text{PA}_i, U_i)\}_{i=1}^d,$$

computing 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 $\mathbf{U} = \{U_i\}_{i=1}^d$, incorporating the assumption that there is no unobserved confounding (*causal sufficiency*). Provided that 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 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 background variables $\mathbb{P}_{\mathbf{U}|\mathbf{v}^{\text{F}}}$ (*abduction*) and then proceeding as in the interventional case, i.e., it is induced by $\mathcal{M}^{do(\boldsymbol{\theta}_{\mathcal{I}})|\mathbf{v}^{\text{F}}} = (\mathbf{S}^{do(\boldsymbol{\theta}_{\mathcal{I}})}, \mathbb{P}_{\mathbf{U}|\mathbf{v}^{\text{F}}})$.

Intervention-based recourse To capture causal relations between features, Karimi et al. [19] propose to approach the actionable recourse task within the framework of SCMs and to shift the focus from nearest CEs to minimal interventions (MINT), leading to the following optimisation problem,

$$\text{MINT-rec}(\mathbf{x}^{\text{F}}) \in \underset{\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{x}^{\text{F}})}{\arg\min} \; \text{cost}(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{x}^{\text{F}}) \qquad \text{subject to} \qquad h(\mathbf{x}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})) = 1, \tag{3}$$

where $\mathbf{x}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})$ denotes the "counterfactual twin" of $\mathbf{x}^{\text{F}}$ had $\mathbf{X}_{\mathcal{I}}$ been $\boldsymbol{\theta}_{\mathcal{I}}$. In practice, the SCM is unknown and needs to be inferred from domain knowledge and data, leading to probabilistic versions of (3) that overcome a lack of guarantees via actions that achieve recourse with high probability [23].

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

[3] E.g., an increase in income will likely have a positive downstream effect on the individual's savings balance.

lead to potential discrimination against protected groups. While there are many different statistical notions of fairness [1–7], these are sometimes mutually incompatible [4] and it has been argued that discrimination, at its heart, corresponds to a causal influence of a protected attribute on the prediction, thus making fairness a fundamentally causal problem [8–13]. Of particular interest to our work is the notion of *counterfactual fairness* [14] (and extensions [15–18]), which calls a (probabilistic) classifier h over $\mathbf{V} = \mathbf{X} \cup A$ counterfactually fair if it satisfies

$$h(\mathbf{v}^{\texttt{F}}) = h(\mathbf{v}_a(\mathbf{u}^{\texttt{F}})) \qquad \text{for all} \qquad a \in \mathcal{A} \qquad \text{and} \qquad \mathbf{v}^{\texttt{F}} = (\mathbf{x}^{\texttt{F}}, a^{\texttt{F}}) \in \mathcal{X} \times \mathcal{A},$$

where $\mathbf{v}_a(\mathbf{u}^{\texttt{F}})$ denotes the "counterfactual twin" of $\mathbf{v}^{\texttt{F}}$ had the attribute been a instead of $a^{\texttt{F}}$.

Fairness by equalising recourse cost across groups As opposed to the fairness of a model's predictions, Gupta et al. [26] are, to the best of our knowledge, the first to study the *fairness of recourse actions*. They advocate for equalising the average cost of recourse across protected groups and to incorporate this as a constraint when training a classifier. Defining the following subgroups,

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

they take a distance-based approach in line with the view of CEs and define the cost of recourse for individual $\mathbf{x}^{\texttt{F}}$ with $h(\mathbf{x}^{\texttt{F}}) = -1$ as the minimum achieved in (1), i.e.,

$$r_{\text{IW}}(\mathbf{x}^{\texttt{F}}) = \min_{\mathbf{x} \in \mathcal{X}} d(\mathbf{x}^{\texttt{F}}, \mathbf{x}) \qquad \text{subject to} \qquad h(\mathbf{x}) = 1. \tag{4}$$

Note that this is equivalent to the IW-recourse (2) of Ustun et al. [21] if $\text{cost}(\boldsymbol{\delta}; \mathbf{x}^{\texttt{F}}) = d(\mathbf{x}^{\texttt{F}}, \mathbf{x}^{\texttt{F}} + \boldsymbol{\delta})$ is chosen as cost function. The group-level cost of recourse is then simply given by the average,[4]

$$r_{\text{IW}}(G_a^-) = \frac{1}{|G_a^-|} \sum_{\mathbf{v}^i \in G_a^-} r_{\text{IW}}(\mathbf{x}^i). \tag{5}$$

We remark that while Gupta et al. [26] only consider classifiers (and distances) defined over $\mathcal{X}$, we use the notation $\mathbf{v}^i = (\mathbf{x}^i, a^i)$ to allow for the more general case of classifiers defined over $\mathcal{V} = \mathcal{X} \times \mathcal{A}$, as is common in the (causal) fairness literature.[5]

The idea of fair recourse by equalising cost (i.e., distance from the decision boundary) across negatively classified subgroups [26] can then be summarised as follows.

Definition 1 (Group IW-fair recourse; adapted from Gupta et al. [26]). *We say that the pair $(\mathcal{D}, h)$ satisfies group* IW*-fair recourse at level ϵ if*

$$\left| r_{\text{IW}}(G_a^-) - r_{\text{IW}}(G_{a'}^-) \right| \leq \epsilon \qquad \forall a, a' \in \mathcal{A}.$$

We say that $(\mathcal{D}, h)$ is group IW*-fair if it satisfies the above for any $\epsilon > 0$.*

3 Fair causal recourse

Since the notion of fair recourse of Gupta et al. [26] rests on the assumption of independently manipulable features inherent to the view of CEs (1) [27] and IW-recourse (2) [21], it exhibits the shortcomings pointed out by Karimi et al. [19, 23]: it does not model causal relationships between variables and thus fails to account for downstream effects of actions on other relevant features, thus potentially incorrectly estimating the true cost of recourse. In the present work, we therefore argue that considerations regarding the fairness of recourse actions should be based on an underlying causal model that can capture the effect of interventions performed in the physical world where features may be causally related to each other.

In §3.1, we will consider an alternative group-level fairness criterion to Definition 1, using instead the MINT-based view of recourse (3) [19], before moving to individualised notions of fair recourse inspired by counterfactual fairness [14] in §3.2.

Throughout we consider an underlying 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.

[4] One could also define recourse cost with reference to the true distribution; since, in practice, we generally only observe a finite sample $\mathcal{D}$ from it, we choose to work with the empirical distribution instead.

[5] Forcing the classifier to be agnostic to the protected attribute is aligned with the notion of fairness through unawareness [2], but it has been shown that such blindness is insufficient for various fairness criteria, e.g., because the protected attribute may be correlated with features used for classification.

3.1 Group-level fair causal recourse

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

$$r_{\text{MINT}}(\mathbf{v}^{\text{F}}) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{v}^{\text{F}})} \text{cost}(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\text{F}}) \qquad \text{subject to} \qquad h(\mathbf{v}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})) = 1. \tag{6}$$

Let $r_{\text{MINT}}(G_a^-)$ be the average of $r_{\text{MINT}}(\mathbf{v}^{\text{F}})$ across G_a^-, analogously to (5). We can then define group-level causally-fair recourse as follows.

Definition 2 (Group MINT-fair recourse). *We say that $(\mathcal{D}, h, \mathcal{M})$ satisfies group MINT-fair recourse at level ϵ if*

$$\left| r_{\text{MINT}}(G_a^-) - r_{\text{MINT}}(G_{a'}^-) \right| \leq \epsilon \qquad \forall a, a' \in \mathcal{A}.$$

We say that $(\mathcal{D}, h, \mathcal{M})$ is group MINT-fair if it satisfies the above for any $\epsilon > 0$.

Definition 2 has the conceptual advantage over Definition 1 that it takes causal relationships into account and thus reflects the true effect of actions and the necessary cost of recourse more faithfully, provided that the IW-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 both only consider the average cost of recourse across all individuals sharing the same protected attribute. However, it has been argued both from causal [10, 11, 14, 16] and non-causal [2] perspectives that fairness is fundamentally a concept at the level of the individual. After all, for an individual who was unfairly given an unfavourable prediction it is not much consolation finding out that other members of the same group were treated more favourably in return. 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 the main idea behind counterfactual fairness [14]: 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 [14], 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 group, i.e., under a counterfactual change to the protected attribute.

Definition 3 (Individually MINT-fair recourse). *We say that $(\mathcal{D}, h, \mathcal{M})$ satisfies individually MINT-fair recourse at level ϵ if*

$$\left| r_{\text{MINT}}(\mathbf{v}^{F}) - r_{\text{MINT}}(\mathbf{v}_a(\mathbf{u}^{F})) \right| \leq \epsilon \qquad \forall a \in \mathcal{A}, \mathbf{v}^{F} \in \mathcal{D}.$$

We say that $(\mathcal{D}, h, \mathcal{M})$ is individually MINT-fair if it satisfies the above for any $\epsilon > 0$.

We note that it is possible to satisfy both group IW-fair (Definition 1) and group MINT-fair recourse (Definition 2), without satisfying individually MINT-fair recourse.

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

$$\textit{Group IW-fair} \;\not\Longrightarrow\; \textit{Individually MINT-fair} \quad \text{and} \quad \textit{Group MINT-fair} \;\not\Longrightarrow\; \textit{Individually MINT-fair}.$$

The proof is given by the following counterexample.

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

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

It can easily be shown that $\mathbb{P}_{X|A=0} = \mathbb{P}_{X|A=1} = \textit{Bernoulli}(0.5)$, which implies that the distance to the decision boundary at $X = 0.5$ is the same across both subgroups, so that the criterion for group-level IW-fairness of recourse (Definition 1) is satisfied by $(\mathcal{D}, h)$ for any finite sample $\mathcal{D}$.

Considering that protected attributes are generally immutable (thus making any recourse actions involving changes to A infeasible) and that 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 seems to be a reasonable choice of cost function for* MINT*-recourse. In this case,* $(\mathcal{D}, h, \mathcal{M})$ *also satisfies group-level* MINT*-fair recourse (Definition 2).*

However, for all $\mathbf{v}^{\texttt{F}} = (x^{\texttt{F}}, a^{\texttt{F}})$ *and any* $a \neq a^{\texttt{F}}$*, we have* $h(x^{\texttt{F}}) \neq h(x_a(u_X^{\texttt{F}})) = 1 - h(x^{\texttt{F}})$ *, 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 classification would have been the opposite.*

3.3 Relation to counterfactual fairness

Note that in Example 5 h is *not* counterfactually fair. This suggests to investigate the relationship between counterfactual fairness and individually MINT-fair recourse; in particular, one may wonder: *does a counterfactually fair classifier imply individually* MINT*-fair recourse?*

Proposition 6. *Counterfactual fairness of* h *is not sufficient for any of the three notions of fair recourse in Definitions 1, 2, and 3, i.e.,*

$$h \text{ counterfactually fair } \not\Rightarrow \text{ [Group IW-fair / Group MINT-fair / Individually MINT-fair] recourse}$$

The above statement is proven by the following counter-example.

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

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

Then $h(X) = \operatorname{sign}(X) = \operatorname{sign}(U_X)$*, 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* $\mathbb{P}_{X|A=1} = \mathcal{N}(0,1)$*, hence the distance to the decision boundary (which is a reasonable cost for* MINT*-recourse in this one-variable toy example) differs significantly both at the group level and when counterfactually changing* A*: specifically, the cost of recourse for members of* G_0^- *is twice as high as that for members of* G_1^-*.*

Remark 8. *An important characteristic of Example 7 is that the classifier* h *is deterministic, which makes it possible that* h *is counterfactually fair, even though it depends on a descendant of the protected attribute* A*. This would generally not be the case if* h *were probabilistic with the probability of a positive classification decreasing with the distance from the decision boundary, e.g., by replacing the structural equation for* Y *with*

$$Y := \mathbb{I}[U_Y > h(X)], \qquad h(X) = \left(1 + e^{0.5 - X}\right)^{-1}, \qquad U_Y \sim U[0,1].$$

4 Achieving fair causal recourse

Having introduced two new causally-motivated definitions of fair recourse, we now discuss approaches for achieving such fair causal recourse algorithmically, i.e., by altering the underlying classifier.

4.1 ... through constrained optimisation

A first possible approach to achieving fair recourse is to take fairness constraints into account when training the classifier, as suggested for IW-recourse by Gupta et al. [26]. However, it is worth pointing out that the optimisation problem in (6) (upon which our causal notions of fair recourse are built) is considerably more involved than the distance-based IW-recourse optimisation problem in (4). This is because the former involves optimising both over the combinatorial space of intervention targets $\mathcal{I} \subseteq \{1, ..., d\}$ and the corresponding values $\boldsymbol{\theta}_{\mathcal{I}}$, subject to a constraint on the resulting counterfactual computed from the SCM $\mathcal{M}$. A second challenge is that the true SCM $\mathcal{M}$ is generally not known, but instead needs to be learnt from data based on additional (domain-specific) assumptions in order to compute minimal interventions for recourse [23]. As a result, it is not immediately clear whether MINT-fairness of recourse may easily be included as a differentiable constraint when training a classifier, so that brute-force approaches may be necessary.

4.2 ... through restricting the set of classifier inputs

A different approach to achieve MINT-fair 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 during training to only contain non-descendants of the protected attribute. In this case, (and subject to some additional assumptions discussed in more detail below) individually MINT-fair recourse can be guaranteed, as summarised in the following proposition.

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

Proof. According to Definition 3, it suffices to show that $r_{\text{MINT}}(\mathbf{v}^{\text{F}}) = r_{\text{MINT}}(\mathbf{v}_a(\mathbf{u}^{\text{F}}))$ for all $a \in \mathcal{A}$ and $\mathbf{v}^{\text{F}} \in \mathcal{D}$. Substituting our assumptions in the definition of r_{MINT} (6), we get:

$$r_{\text{MINT}}(\mathbf{v}^{\text{F}}) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{v}^{\text{F}})} \text{cost}(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\text{F}}) \qquad \text{subject to} \qquad h(\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})) = 1,$$

$$r_{\text{MINT}}(\mathbf{v}_a(\mathbf{u}^{\text{F}})) = \min_{\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{v}^{\text{F}})} \text{cost}(\boldsymbol{\theta}_{\mathcal{I}}; \mathbf{v}^{\text{F}}) \qquad \text{subject to} \qquad h(\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}},a}(\mathbf{u}^{\text{F}})) = 1.$$

It thus only remains to show that $\tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}},a}(\mathbf{u}^{\text{F}}) = \tilde{\mathbf{x}}_{\boldsymbol{\theta}_{\mathcal{I}}}(\mathbf{u}^{\text{F}})$ for all $\boldsymbol{\theta}_{\mathcal{I}} \in \Theta(\mathbf{v}^{\text{F}}), a \in \mathcal{A}$, which follows from the application of do-calculus 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. *Note that the condition of relying exclusively on non-descendants of the protected attribute in Proposition 9 is also sufficient to ensure counterfactual fairness of h, see [14, Lemma 1].*

We point out that the assumption of Proposition 9 that both the set of possible actions $\Theta(\mathbf{v}^{\text{F}})$ and the cost function $\text{cost}(\cdot\,; \mathbf{v}^{\text{F}})$ 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 the assumptions of Proposition 9, can therefore be a strong restriction for many applications. Consequently, the cost of achieving fair recourse in this way is likely a significantly lower classification accuracy.

4.3 ... through representation learning / abduction

We have shown that considering only non-descendants of A is a way to achieve individually MINT-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 a 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. [14, § 4.1].

4.4 ... through altering the data generating process

Note that our notion of MINT-fair recourse is defined in reference to the triple $(\mathcal{D}, h, \mathcal{M})$. 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 Societal interventions

In fair ML, typically we attempt to enforce a notion of fairness by requiring a learned classifier to satisfy some constraint [6]. This implicitly places the cost of an intervention on the deployer. For example, a bank might need to modify their approach so as to make 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 Definition 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 in order 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.

Returning to Example 7 where the variance of each subgroup differs, the higher variance group may incur a high cost of recourse on an individual level, which results in costs for the bank that might not be assumable. Furthermore, without additional support for individuals, the high uncertainty in repaying the loan of this group will not be reduced by changing h (per the techniques discussed in Section 4) but instead by altering the population distribution. One way to alter the population distribution could be external interventions, e.g., subsidies for particular eligible individuals. We can consider the relative merits of a welfare system which distributes money to particular individuals, or a regulatory regime which can place requirements on banks that might reduce their profitability or affect prices or state tax receipts.

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 individual level). This is akin to mechanism design in game theory: what societal intervention (e.g., supporting one of the subgroups) would enable more fair recourse, along the lines of Kusner et al. [32] 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 generally at a disadvantage due to having spent a portion of their potential time for gaining work experience (e.g., time since graduation) with family instead. A common societal intervention in this scenario is to alter the process by which work experience is counted or normalised, e.g., by taking possible pregnancies and maternity leaves into account.

Note that when thinking in these terms, we might also question whether it is always good 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 have significant negative costs to them, to the bank, and to society. Leveraging the economics literature which studies the effect of policy interventions on society, institutions, and individuals [33, 34], 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. *Is it better to sacrifice accuracy in enforcing fairness within the current system or to invest (e.g., via subsidies) to change the underlying system to enable more fair and accurate prediction?*

6 Conclusion

In this work, we considered the fairness of recourse, as opposed to the fairness of predictions. Following similar lines in earlier work on both fairness and recourse, 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. While our fairness criteria may help assess the fairness or recourse more faithfully, 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.

References

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

[2] 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.

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

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

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

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

[7] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. The variational fair autoencoder. In *ICLR*, 2016.

[8] 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.

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

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

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

[12] 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.

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

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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*, 2020.

[24] 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.

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

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

[27] 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.

[28] 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.

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

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

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

[32] 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.

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

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