Algorithmic Recourse: from Counterfactual Explanations to Interventions

Amir-Hossein Karimi
MPI-IS, Germany
ETH Zürich, Switzerland

Bernhard Schölkopf
MPI-IS, Germany

Isabel Valera
MPI-IS, Germany
Saarland University, Germany

ABSTRACT

As machine learning is increasingly used to inform consequential decision-making (e.g., pre-trial bail and loan approval), it becomes important to explain how the system arrived at its decision, and also suggest actions to achieve a favorable decision. Counterfactual explanations –"how the world would have (had) to be different for a desirable outcome to occur"– aim to satisfy these criteria. Existing works have primarily focused on designing algorithms to obtain counterfactual explanations for a wide range of settings. However, it has largely been overlooked that ultimately, one of the main objectives is to allow people to act rather than just understand. In layman's terms, counterfactual explanations inform an individual where they need to get to, but not how to get there. In this work, we rely on causal reasoning to caution against the use of counterfactual explanations as a recommendable set of actions for recourse. Instead, we propose a shift of paradigm from *recourse via nearest counterfactual explanations* to *recourse through minimal interventions*, shifting the focus from explanations to interventions.

KEYWORDS

algorithmic recourse, counterfactual explanations, minimal interventions, interpretable machine learning

1 INTRODUCTION

Predictive models are being increasingly used to support consequential decision-making in a number of contexts, e.g., denying a loan, rejecting a job applicant, or prescribing life-altering medication. As a result, there is mounting social and legal pressure [51] to provide explanations that help the affected individuals to understand "why a prediction was output", as well as "how to act" to obtain a desired outcome. Answering these questions, for the different stakeholders involved, is one of the main goals of explainable machine learning [7, 14, 20, 27, 32, 41, 42].

In this context, several works have proposed to explain a model's predictions of an affected individual using *counterfactual explanations*, which are defined as statements of "how the world would have (had) to be different for a desirable outcome to occur" [52]. Of specific importance are *nearest counterfactual explanations*, presented as the most similar *instances* to the feature vector describing the individual, that result in the desired prediction from the model [18, 26]. A closely related term is *algorithmic recourse* – the actions required for, or "the systematic process of reversing unfavorable decisions by algorithms and bureaucracies across a range of counterfactual scenarios" – which is argued as the underwriting factor for temporally extended agency and trust [50].

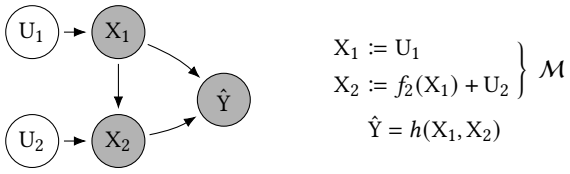

Figure 1: Illustration of an example causal generative process governing the world, showing both the graphical model, $\mathcal{G}$, and the structural causal model, $\mathcal{M}$, [34]. In this example, X_1 represents an individual's annual salary, X_2 is bank balance, and $\hat{Y}$ is the output of a fixed deterministic predictor h, predicting the eligibility of an individual to receive a loan.

Counterfactual explanations have shown promise for practitioners and regulators to validate a model on metrics such as fairness and robustness [18, 45, 49]. However, in their raw form, such explanations do not seem to fulfill one of the primary objectives of "explanations as a means to help a data-subject *act* rather than merely *understand*" [52].

The translation of counterfactual explanations to recourse actions, i.e., to a recommendable set of actions to help an individual to achieve a favourable outcome, was first explored in [49], where additional *feasibility* constraints were imposed to support the concept of actionable features (e.g., prevent asking the individual to reduce their age or change their race). While a step in the right direction, this work and others that followed [18, 31, 38, 45] implicitly assume that the set of actions resulting in the desired output would directly follow from the counterfactual explanation. This arises from the assumption that "what would *have had to be* in the past" (retrodiction) not only translates to "what *should be* in the future" (prediction) but also to "what *should be done* in the future" (recommendation) [47]. We challenge this assumption and attribute the shortcoming of existing approaches to their lack of consideration for real-world properties, specifically the *causal relationships* governing the world in which actions will be performed.

For ease of exposition, we present the following examples (see [3] for additional examples).

Example 1: Consider, for example, the setting in Figure 1 where an individual has been denied a loan and seeks an explanation and recommendation on how to proceed. This individual has an annual salary (X_1) of \$75,000 and an account balance (X_2) of \$25,000 and the predictor grants a loan based on the binary output of $h = \text{sgn}(X_1 + 5 \cdot X_2 - \$225,000)$. Existing approaches may identify nearest counterfactual explanations as another individual with an annual salary of \$100,000 (+%33) or a bank balance of \$30,000 (+%20), therefore encouraging the individual to reapply when either of these conditions are met. On the other hand, bearing in mind that

actions take place in a world where home-seekers save %30 of their salary (i.e., $X_2 := 3/10 \cdot X_1 + U_2$), a salary increase of only %14 to \$85,000 would automatically result in \$3,000 additional savings, with a net positive effect on the loan-granting algorithm's decision.

Example 2: Consider now another setting of Figure 1 where an agricultural team wishes to increase the yield of their rice paddy. While many factors influence yield = h(temperature, solar radiation, water supply, seed quality, ...), the primary actionable capacity of the team is their choice of paddy location. Importantly, the altitude at which the paddy sits has an effect on other variables. For example, the laws of physics may imply that a $100m$ increase in elevation results in a 1°C decrease in temperature on average. Therefore, it is conceivable that a counterfactual explanation suggesting an increase in elevation for optimal yield, without consideration for downstream effects of the elevation increase on other variables, may actually result in the prediction *not* changing.

The two examples above illustrate the pitfalls of generating recourse actions directly from counterfactual explanations without consideration for the structure of the world in which the actions will be performed. Actions derived directly from counterfactual explanations may ask too much effort from the individual (**Example 1**) or may not even result in the desired output (**Example 2**).

In this paper, we remedy this situation via a fundamental reformulation of the recourse problem, where we rely on causal reasoning to incorporate knowledge of causal dependencies into the process of recommending recourse actions, that if acted upon would result in a counterfactual instance that favourably changes the output of the predictive model. In more detail, we first provide a causal analysis to illuminate the intrinsic limitations of the setting in which actions directly follow counterfactual explanations. Importantly, we show that even when equipped with knowledge of causal dependencies after-the-fact, the actions derived from pre-computed (nearest) counterfactual explanations may prove sub-optimal, or directly, unfeasible. Second, to address the above limitations, we emphasize that, from a causal perspective, actions correspond to interventions which not only model the change in the intervened-upon variable, but also the downstream effects of this intervention on the rest of the (non-intervened-upon) variables. This insight allows us to propose a *recourse through minimal interventions* problem, whose solution informs stakeholders on how to act in addition to understand. We complement this result with a commentary on the form of interventions, and with a more general definition of feasibility beyond actionability. Finally, we provide a detailed discussion on both the importance and the practical limitations of incorporating causal reasoning in the formulation of recourse.

2 ALGORITHMIC RECOURSE VIA COUNTERFACTUAL EXPLANATIONS

Counterfactual explanations (CFE) are statements of "how the world would have (had) to be different for a desirable outcome to occur" [52]. In the context of explainable machine learning, the literature has focused on finding *nearest counterfactual explanations* (i.e., instances),[1] which result in the desired prediction while incurring the smallest change to the individual's feature vector, as measured by a context-dependent dissimilarity metric, $\text{dist}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. This problem has been formulated as the following optimization problem [52]:

$$\boldsymbol{x}^{*\text{CFE}} \in \underset{\boldsymbol{x}}{argmin} \quad \text{dist}(\boldsymbol{x}, \boldsymbol{x}^{\text{F}}) \quad \text{s.t.} \quad h(\boldsymbol{x}) \neq h(\boldsymbol{x}^{\text{F}}), \boldsymbol{x} \in \mathcal{P}, \tag{1}$$

where $\boldsymbol{x}^{\text{F}} \in \mathcal{X}$ is the factual instance; $\boldsymbol{x}^{*\text{CFE}} \in \mathcal{X}$ is a (perhaps not unique) nearest counterfactual instance; h is the fixed binary predictor; and $\mathcal{P}$ is an optional set of *plausibility* constraints, e.g., the counterfactual instance be from a relatively high-density region of the input space [17, 38].

Most of the existing approaches in the counterfactual explanations literature have focused on providing solutions to the optimization problem in (1), by exploring semantically meaningful distance/dissimilarity functions $\text{dist}(\cdot,\cdot)$ between individuals (e.g., $\ell_0, \ell_1, \ell_\infty$, percentile-shift), accommodating different predictive models h (e.g., random forest, multilayer perceptron), and realistic plausibility constraints, $\mathcal{P}$. In particular, [6, 31, 52] solve (1) using gradient-based optimization; [43, 49] employ mixed-integer linear program solvers to support mixed numeric/binary data; [38] use graph-based shortest path algorithms; [26] use a heuristic search procedure by growing spheres around the factual instance; [13, 45] build on genetic algorithms for model-agnostic behavior; and [18] solve (1) using satisfiability solvers with closeness guarantees.

Although nearest counterfactual explanations provide an *understanding* of the most similar set of features that result in the desired prediction, they stop short of giving explicit *recommendations* on how to act to realize this set of features. The lack of specification of the actions required to realize $\boldsymbol{x}^{*\text{CFE}}$ from $\boldsymbol{x}^{\text{F}}$ leads to uncertainty and limited agency for the individual seeking recourse. To shift the focus from explaining a decision to providing recommendable actions to achieve recourse, Ustun et al. [49] reformulated (1) as:

$$\begin{aligned} \boldsymbol{\delta}^* \in \underset{\boldsymbol{\delta}}{argmin}\ \text{cost}(\boldsymbol{\delta}; \boldsymbol{x}^{\text{F}}) \quad & \text{s.t. } h(\boldsymbol{x}^{\text{CFE}}) \neq h(\boldsymbol{x}^{\text{F}}), \\ & \boldsymbol{x}^{\text{CFE}} = \boldsymbol{x}^{\text{F}} + \boldsymbol{\delta}, \\ & \boldsymbol{x}^{\text{CFE}} \in \mathcal{P},\ \boldsymbol{\delta} \in \mathcal{F}, \end{aligned} \tag{2}$$

where $\text{cost}(\cdot; \boldsymbol{x}^{\text{F}}): \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a user-specified cost that encodes preferences between feasible actions from $\boldsymbol{x}^{\text{F}}$, and $\mathcal{F}$ and $\mathcal{P}$ are optional sets of feasibility and plausibility constraints,[2] restricting the actions and the resulting counterfactual explanation, respectively. The feasibility constraints in (2), as introduced in [49], aim at restricting the set of features that the individual may act upon. For instance, recommendations should not ask individuals to change their gender or reduce their age. Henceforth, we refer to the optimization problem in (2) as the *CFE-based recourse* problem.

3 A CAUSAL PERSPECTIVE OF ALGORITHMIC RECOURSE

The seemingly innocent reformulation of the counterfactual explanation problem in (1) as a recourse problem in (2) is founded on two assumptions:

Assumption 1: the feature-wise difference between factual and nearest counterfactual instances, $\boldsymbol{\delta}^* = \boldsymbol{x}^{*\text{CFE}} - \boldsymbol{x}^{\text{F}}$, directly translates

[1] A counterfactual instance can be from the dataset [38, 53] or generated [18, 49, 52].

[2] Here, "feasible" means *possible to do*, whereas "plausible" means *possibly true, believable or realistic*. Optimization terminology refers to both as *feasibility* sets.

to the minimal action set, $\mathbf{A}^{\mathsf{CFE}}$, such that performing the actions in $\mathbf{A}^{\mathsf{CFE}}$ starting from $\boldsymbol{x}^{\mathsf{F}}$ will result in $\boldsymbol{x}^{*\mathsf{CFE}}$; and

Assumption 2: there is a 1-1 mapping between $\text{dist}(\cdot,\cdot)$ and $\text{cost}(\cdot;\cdot)$, whereby larger actions incur larger distance and higher cost.

Unfortunately, these assumptions only hold in restrictive settings, rendering the solution of (2) *sub-optimal* or *infeasible* in many real-world scenarios. Specifically, **Assumption 1** holds only if (i) the individual applies effort in a world where changing a variable does not have downstream other variables (i.e., features are independent from each other); or if (ii) the individual changes the value of a subset of variables while simultaneously enforcing that the value of all other variables remain unchanged (i.e., breaking dependencies between features). Beyond the *sub-optimality* that arises from assuming/reducing to an independent world in (i), and disregarding the *feasibility* of non-altering actions in (ii), non-altering actions may naturally incur a cost which is not captured in the current definition of cost, and hence **Assumption 2** does not hold either. Therefore, except in trivial cases where the model designer actively inputs pair-wise independent features to h, generating recommendations from counterfactual explanations in this manner, i.e., ignoring the dependencies between features, warrants reconsideration. Next, we formalize these shortcomings using causal reasoning.

3.1 Actions as Interventions

Let $\mathcal{M} \in \Pi$ denote the structural causal model (SCM) capturing all inter-variable causal dependencies in the real world. $\mathcal{M} = \langle \mathbb{F}, \mathbb{X}, \mathbb{U} \rangle$ is characterized by the endogenous (observed) variables, $\mathbb{X} \in \mathcal{X}$, the exogenous variables, $\mathbb{U} \in \mathcal{U}$, and a sequence of structural equations $\mathbb{F}$: $\mathcal{U} \to \mathcal{X}$, describing how endogenous variables can be (deterministically) obtained from the exogenous variables [34, 46]. Often, $\mathcal{M}$ is illustrated using a directed graphical model, $\mathcal{G}$ (see, e.g., Figure 1).

From a causal perspective, actions may be carried out via *structural interventions*, $\mathbf{A}$: $\Pi \to \Pi$, which can be thought of as a transformation between SCMs [33, 34]. A set of interventions can be constructed as $\mathbf{A} = \text{do}(\{\mathrm{X}_i := a_i\}_{i \in I})$ where I contains the indices of the subset of endogenous variables to be intervened upon. In this case, for each $i \in I$, the do-operator replaces the structural equation for the i-th endogenous variable X_i in $\mathbb{F}$ with $\mathrm{X}_i := a_i$. Correspondingly, graph surgery is performed on $\mathcal{G}$, severing graph edges incident on an intervened variable, X_i. Thus, performing the actions $\mathbf{A}$ in a world $\mathcal{M}$ yields the post-intervention world model $\mathcal{M}_{\mathbf{A}}$ with structural equations $\mathbb{F}_{\mathbf{A}} = \{F_i\}_{i \notin I} \cup \{\mathrm{X}_i := a_i\}_{i \in I}$. Structural interventions are illustrated in Figure 2.

Structural interventions are used to predict the effect of actions on the world as a whole (i.e., how $\mathcal{M}$ becomes $\mathcal{M}_{\mathbf{A}}$). In the context of recourse, we aim to model the effect of actions on one individual's situation (i.e., how $\boldsymbol{x}^{\mathsf{F}}$ becomes $\boldsymbol{x}^{\mathsf{SCF}}$) to ascertain whether or not the desirable outcome is achieved (i.e., $h(\boldsymbol{x}^{\mathsf{F}}) \neq h(\boldsymbol{x}^{\mathsf{SCF}})$). We compute individual-level effects using *structural counterfactuals* [36].

Assuming *causal sufficiency* of $\mathcal{M}$ (i.e., no hidden confounders), and full specification of an invertible $\mathbb{F}$ (such that $\mathbb{F}(\mathbb{F}^{-1}(\boldsymbol{x})) = \boldsymbol{x}$), $\mathbb{X}$ can be uniquely determined given the value of $\mathbb{U}$ (and vice-versa). Hence, one can determine the distinct values of exogenous variables that give rise to a particular realization of the endogenous variables,

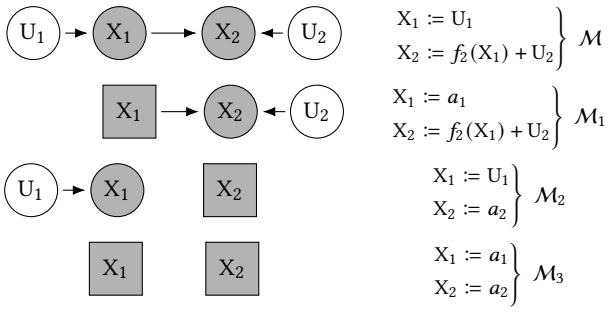

Figure 2: Given world model, $\mathcal{M}$, intervening on X_1 and/or on X_2 result in different post-intervention models: $\mathcal{M}_1 = \mathcal{M}_{\mathbf{A}=\{\text{do}(\mathrm{X}_1 := a_1)\}}$ corresponds to interventions only on X_1 with consequential effects on X_2; $\mathcal{M}_2 = \mathcal{M}_{\mathbf{A}=\{\text{do}(\mathrm{X}_2 := a_2)\}}$ shows the result of structural interventions only on X_2 which in turn dismisses ancestral effects on this variable; and, $\mathcal{M}_3 = \mathcal{M}_{\mathbf{A}=\{\text{do}(\mathrm{X}_1 := a_1, \mathrm{X}_2 := a_2)\}}$ is the resulting (independent world) model after intervening on both variables, i.e., the type of interventions generally assumed in the CFE-based recourse problem.

$\{\mathrm{X}_i = x_i^{\mathsf{F}}\}_i \subseteq \mathcal{X}$, as $\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})$ [36].[3] As a result, we can compute *any* structural counterfactual query $\boldsymbol{x}^{\mathsf{SCF}}$ for an individual $\boldsymbol{x}^{\mathsf{F}}$ as $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$. In our context, that is: "if an individual $\boldsymbol{x}^{\mathsf{F}}$ observed in world $\mathcal{M}$ performs the set of actions $\mathbf{A}$, what *will be* the resulting individual's feature vector $\boldsymbol{x}^{\mathsf{SCF}}$".[4]

3.2 Limitations of CFE-based recourse

Next, we use causal reasoning to formalize the limitations of the CFE-based recourse approach in (2). To this end, we first reinterpret the actions resulting from solving the CFE-based recourse problem, i.e., $\boldsymbol{\delta}^*$, as structural interventions by defining the set of indices of observed variables that are intervened upon, I. We remark that, given $\boldsymbol{\delta}^*$, an individual seeking recourse may intervene on any arbitrary subset of observed variables I, as long as the intervention contains the variable indices for which $\delta_i^* \neq 0$. Now, we are in a position to define CFE-based actions as interventions, i.e.,

Definition 3.1 (CFE-based actions)**.** Given an individual $\boldsymbol{x}^{\mathsf{F}}$ in world $\mathcal{M}$, the solution of (2), $\boldsymbol{\delta}^*$, and the set of indices of observed variables that are acted upon, I, a *CFE-based action* refers to a set of structural interventions of the form $\mathbf{A}^{\mathsf{CFE}} := \text{do}(\{\mathrm{X}_i := x_i^F + \delta_i^*\}_{i \in I})$.

Using Definition 3.1, we can derive the following key results that provide necessary and sufficient conditions for CFE-based actions to guarantee recourse.

Proposition 3.1. *A CFE-based action, $\mathbf{A}^{\mathsf{CFE}}$, where $I = \{i \mid \delta_i^* \neq 0\}$, performed by individual $\boldsymbol{x}^{\mathsf{F}}$, in general results in the structural counterfactual, $\boldsymbol{x}^{\mathsf{SCF}} = \boldsymbol{x}^{*\mathsf{CFE}} := \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^*$, and thus guarantees recourse (i.e., $h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}})$), if and only if, the set of descendants of the acted upon variables, determined by I, is the empty set.*

[3] For notational simplicity, we interchangeably use sets and vectors, e.g., $\{\mathrm{X}_i = x_i^{\mathsf{F}}\}_i \subseteq \mathcal{X}$ and $\boldsymbol{x}^{\mathsf{F}} \in \mathcal{X}$.

[4] Queries such as this subsume both *retrospective/subjunctive/counterfactual* ("what would have been the value of") and *prospective/indicative/predictive* ("what will be the value of") conditionals [11, 25, 48], as long as we assume that the laws governing the world, $\mathbb{F}$, are stationary.

Corollary 3.1. *If the true world $\mathcal{M}$ is independent, i.e, all the observed features are root-nodes, then CFE-based actions always guarantee recourse.*

While the above results are formally proven in Appendix A, we provide a sketch of the proof below. If the intervened-upon variables do not have descendants, then by definition $\boldsymbol{x}^{\mathsf{SCF}} = \boldsymbol{x}^{*\mathsf{CFE}}$. Otherwise, the value of the descendants will depend on the counterfactual value of their parents, leading to a structural counterfactual that does not resemble the nearest counterfactual explanation, $\boldsymbol{x}^{\mathsf{SCF}} \neq \boldsymbol{x}^{*\mathsf{CFE}}$, and thus may not result in recourse. Moreover, in an independent world the set of descendants of all the variables is by definition the empty set.

Unfortunately, the independent world assumption is not realistic, as it requires all the features selected to train the predictive model h to be independent of each other. Moreover, limiting changes to only those variables without descendants may unnecessarily limit the agency of the individual, e.g., in **Example 1**, restricting the individual to only changing bank balance without e.g., pursuing a new/side job to increase their income would be limiting. Thus, for a given non-independent $\mathcal{M}$ capturing the true causal dependencies between features, CFE-based actions require the individual seeking recourse to enforce (at least partially) an independent post-intervention model $\mathcal{M}_{\mathbf{A}^{\mathsf{CFE}}}$ (so that **Assumption 1** holds), by intervening on all the observed variables for which $\delta_i \neq 0$ as well as on their descendants (even if their $\delta_i = 0$). However, such requirement suffers from two main issues. First, it conflicts with **Assumption 2**, since holding the value of variables may still imply potentially *infeasible* and costly interventions in $\mathcal{M}$ to sever all the incoming edges to such variables, and even then it may not change the prediction (see **Example 2**). Second, as will be proven in the next section (see also, **Example 1**), CFE-based actions may still be *suboptimal*, as they do not benefit from the causal effect of actions towards changing the prediction. Thus, even when equipped with knowledge of causal dependencies, recommending actions directly from counterfactual explanations in the manner of existing approaches is not satisfactory.

4 ALGORITHMIC RECOURSE VIA MINIMAL INTERVENTIONS

In the previous section, we learned that actions which immediately follow from counterfactual explanations may require unrealistic assumptions, or alternatively, result in sub-optimal or even infeasible recommendations. To solve such limitations we rewrite the recourse problem so that instead of finding the minimal (independent) shift of features as in (2), we seek the minimal cost set of actions (in the form of structural interventions) that results in a counterfactual instance yielding the favourable output from h:

$$\begin{aligned} \mathbf{A}^* \in \underset{\mathbf{A}}{argmin} \quad & \text{cost}(\mathbf{A}; \boldsymbol{x}^{\mathsf{F}}) \\ \text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\ & \boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) \\ & \boldsymbol{x}^{\mathsf{SCF}} \in \mathcal{P}, \quad \mathbf{A} \in \mathcal{F}, \end{aligned} \tag{3}$$

where $\mathbf{A}^* \in \mathcal{F}$ directly specifies the set of feasible actions to be performed for minimally costly recourse, with $\text{cost}(\cdot; \boldsymbol{x}^{\mathsf{F}}) : \mathcal{F} \times \mathcal{X} \to \mathbb{R}_+$,

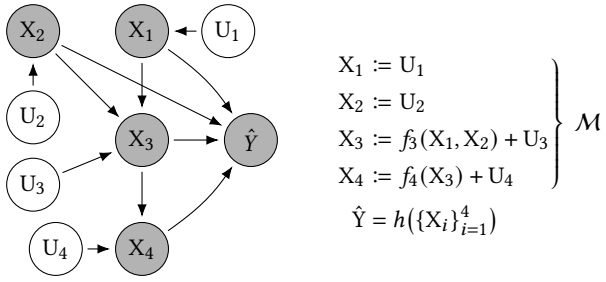

Figure 3: The structural causal model (graph and equations) for the working example and demonstration in Section 4.

and $\boldsymbol{x}^{*\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$ denotes the resulting structural counterfactual. We recall that, although $\boldsymbol{x}^{*\mathsf{SCF}}$ is a counterfactual instance, it does not need to correspond to the nearest counterfactual explanation, $\boldsymbol{x}^{*\mathsf{CFE}}$, resulting from (2) (see, e.g., **Example 1**). Importantly, using the formulation in (3) it is now straightforward to show the suboptimality of CFE-based actions, as shown next (proof in Appendix A):

Proposition 4.1. *Given an individual $\boldsymbol{x}^{\mathsf{F}}$ observed in world $\mathcal{M} \in \Pi$, a family of feasible actions $\mathcal{F}$, and the solution of* (3)*, $\mathbf{A}^* \in \mathcal{F}$. Assume that there exists CFE-based action $\mathbf{A}^{\mathsf{CFE}} \in \mathcal{F}$ that achieves recourse, i.e., $h(\boldsymbol{x}^{\mathsf{F}}) \neq h(\boldsymbol{x}^{*\mathsf{CFE}})$. Then,* $\text{cost}(\mathbf{A}^*; \boldsymbol{x}^{\mathsf{F}}) \leq \text{cost}(\mathbf{A}^{\mathsf{CFE}}; \boldsymbol{x}^{\mathsf{F}})$.

Thus, for a known causal model capturing the dependencies among observed variables, and a family of feasible interventions, the optimization problem in (3) yields *Recourse through Minimal Interventions* (MINT). Generating minimal interventions through solving (3) requires that we be able to compute the structural counterfactual, $\boldsymbol{x}^{\mathsf{SCF}}$, of the individual $\boldsymbol{x}^{\mathsf{F}}$ in world $\mathcal{M}$, given *any* feasible action, $\mathbf{A}$. To this end, we consider that the SCM $\mathcal{M}$ falls in the class of additive noise models (ANM), so that we can deterministically compute the counterfactual $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$ by performing the *Abduction-Action-Prediction* steps proposed by Pearl et al. [36].

4.1 Working example

Consider the model in Figure 3, where $\{\mathrm{U}_i\}_{i=1}^4$ are mutually independent exogenous variables, and $\{f_i\}_{i=1}^4$ are structural (linear or nonlinear) equations. Let $\boldsymbol{x}^{\mathsf{F}} = [x_1^{\mathsf{F}}, x_2^{\mathsf{F}}, x_3^{\mathsf{F}}, x_4^{\mathsf{F}}]^T$ be the observed features belonging to the (factual) individual, for whom we seek a counterfactual explanation and recommendation. Also, let I denote the set of indices corresponding to the subset of endogenous variables that are intervened upon according to the action set $\mathbf{A}$. Then, we obtain a structural counterfactual, $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$, by applying the Abduction-Action-Prediction steps [35] as follows:

Step 1. Abduction uniquely determines the value of all exogenous variables, $\{u_i\}_{i=1}^4$, given evidence, $\{\mathrm{X}_i = x_i^{\mathsf{F}}\}_{i=1}^4$:

$$\begin{aligned} u_1 &= x_1^{\mathsf{F}}, \\ u_2 &= x_2^{\mathsf{F}}, \\ u_3 &= x_3^{\mathsf{F}} - f_3(x_1^{\mathsf{F}}, x_2^{\mathsf{F}}), \\ u_4 &= x_4^{\mathsf{F}} - f_4(x_3^{\mathsf{F}}). \end{aligned} \tag{4}$$

Step 2. Action modifies the SCM according to the hypothetical interventions, $\text{do}(\{\text{X}_i := a_i\}_{i \in I})$ (where $a_i = x_i^F + \delta_i$), yielding $\mathbb{F}_{\mathbf{A}}$:

$$\begin{aligned}
\text{X}_1 &:= [1 \in I] \cdot a_1 + [1 \notin I] \cdot \text{U}_1, \\
\text{X}_2 &:= [2 \in I] \cdot a_2 + [2 \notin I] \cdot \text{U}_2, \\
\text{X}_3 &:= [3 \in I] \cdot a_3 + [3 \notin I] \cdot \big(f_3(\text{X}_1, \text{X}_2) + \text{U}_3\big), \\
\text{X}_4 &:= [4 \in I] \cdot a_4 + [4 \notin I] \cdot \big(f_4(\text{X}_3) + \text{U}_4\big),
\end{aligned} \tag{5}$$

where $[\cdot]$ denotes the Iverson bracket.

Step 3. Prediction recursively determines the values of all endogenous variables based on the computed exogenous variables $\{u_i\}_{i=1}^4$ from Step 1 and $\mathbb{F}_{\mathbf{A}}$ from Step 2, as:

$$\begin{aligned}
x_1^{\mathsf{SCF}} &:= [1 \in I] \cdot a_1 + [1 \notin I] \cdot \big(u_1\big), \\
x_2^{\mathsf{SCF}} &:= [2 \in I] \cdot a_2 + [2 \notin I] \cdot \big(u_2\big), \\
x_3^{\mathsf{SCF}} &:= [3 \in I] \cdot a_3 + [3 \notin I] \cdot \big(f_3(x_1^{\mathsf{SCF}}, x_2^{\mathsf{SCF}}) + u_3\big), \\
x_4^{\mathsf{SCF}} &:= [4 \in I] \cdot a_4 + [4 \notin I] \cdot \big(f_4(x_3^{\mathsf{SCF}}) + u_4\big).
\end{aligned} \tag{6}$$

4.2 General assignment formulation

As we have not made any restricting assumptions about the structural equations (only that we operate with additive noise models[5] where noise variables are pairwise independent), the solution for the working example naturally generalizes to SCMs corresponding to other DAGs with more variables. The assignment of structural counterfactual values can generally be written as:

$$\begin{aligned}
x_i^{\mathsf{SCF}} = {} & [i \in I] \cdot (x_i^{\mathsf{F}} + \delta_i) \\
& + [i \notin I] \cdot \big(x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}})\big).
\end{aligned} \tag{7}$$

In words, the counterfactual value of the i-th feature, x_i^{SCF}, takes the value $x_i^{\mathsf{F}} + \delta_i$ if such feature is intervened upon (i.e., $i \in I$). Otherwise, x_i^{SCF} is computed as a function of both the factual and counterfactual values of its parents, denoted respectively by $f_i(\mathbf{pa}_i^{\mathsf{F}})$ and $f_i(\mathbf{pa}_i^{\mathsf{SCF}})$. The closed-form expression in (7) can replace the counterfactual constraint in (3), i.e., $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$, after which the optimization problem may be solved by building on existing frameworks for generating nearest counterfactual explanations, including gradient-based, evolutionary-based, heuristics-based, or verification-based approaches as referenced in Section 2. While out of scope of the current work, for the demonstrative examples below, we extended the open-source code of MACE [18]; we will submit a pull-request to the respective repository.

4.3 Demonstration

We showcase our proposed formulation by comparing the actions recommended by existing (nearest) counterfactual explanation methods, as in (2), to the ones generated by the proposed minimal intervention formulation in (3). We recall that prior literature has focused on generating counterfactual explanations or CFE-based actions, which as shown above lack optimally or feasibility guarantees in non-independent worlds. Thus, to the best of our knowledge, there exists no baseline approach in the literature that guarantees algorithmic recourse. The experiments below serve as an illustration of the sub-optimality of existing approaches relative to our proposed formulation of recourse via minimal intervention. Section 6 presents a detailed discussion on practical considerations.

We consider two settings: i) a synthetic setting where $\mathcal{M}$ follows Figure 1; and ii) a real-world setting based on the `german` credit dataset [1], where $\mathcal{M}$ follows Figure 3. We computed the cost of actions as the ℓ_1 norm over normalized feature changes to make effort comparable across features, i.e., $\text{cost}(\cdot; \boldsymbol{x}^{\mathsf{F}}) = \sum_{i \in I} |\delta_i| / R_i$, where R_i is the range of feature i.

For the *synthetic setting*, we generate data following the model in Figure 1, where we assume $\text{X}_1 := \text{U}_1$, $\text{X}_2 := 3/10 \cdot \text{X}_1 + \text{U}_2$, with $\text{U}_1 \sim \$10000 \cdot \text{Poission}(10)$ and $\text{U}_2 \sim \$2500 \cdot \mathcal{N}(0, 1)$; and the predictive model $h = \text{sgn}(\text{X}_1 + 5 \cdot \text{X}_2 - \$225000)$. Given $\boldsymbol{x}^{\mathsf{F}} = [\$75000, \$25000]^T$, solving our formulation, (3), identifies the optimal action set $\mathbf{A}^* = \text{do}(\text{X}_1 := x_1^{\mathsf{F}} + \$10000)$ which results in $\boldsymbol{x}^{*\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) = [\$85000, \$28000]^T$, whereas solving previous formulations, (2), yields $\boldsymbol{\delta}^* = [\$0, +\$5000]^T$ resulting in $\boldsymbol{x}^{*\mathsf{CFE}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^* = [\$75000, \$30000]^T$. Importantly, while $\boldsymbol{x}^{*\mathsf{SCF}}$ appears to be at a further distance from $\boldsymbol{x}^{\mathsf{F}}$ compared to $\boldsymbol{x}^{*\mathsf{CFE}}$, achieving the former is less costly than the latter, specifically, $\text{cost}(\boldsymbol{\delta}^*; \boldsymbol{x}^{\mathsf{F}}) \approx 2\ \text{cost}(\mathbf{A}^*; \boldsymbol{x}^{\mathsf{F}})$.

As a *real-world setting*, we consider a subset of the features in the `german` credit dataset. The setup is depicted in Figure 3, where X_1 is the individual's gender (treated as immutable), X_2 is the individual's age (actionable but can only increase), X_3 is credit given by the bank (actionable), X_4 is the repayment duration of the credit (non-actionable but mutable), and $\hat{\text{Y}}$ is the predicted customer risk, according to h (logisitic regression or decision tree). We learn the structural equations by fitting a linear regression model to the child-parent tuples. We will release the data, and the code used to learn models and structural equations.

Given the setup above, for instance, for the individual $\boldsymbol{x}^{\mathsf{F}} = [\texttt{Male}, 32, \$1938, 24]^T$ identified as a risky customer, solving our formulation, (3), yields the optimal action set $\mathbf{A}^* = \text{do}(\{\text{X}_2 := x_2^{\mathsf{F}} + 1, \text{X}_3 := x_3^{\mathsf{F}} - \$800\})$ which results in $\boldsymbol{x}^{*\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) = [\texttt{Male}, 33, \$1138, 22]^T$, whereas solving (2) yields $\boldsymbol{\delta}^* = [\texttt{N/A}, +6, 0, 0]^T$ resulting in $\boldsymbol{x}^{*\mathsf{CFE}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^* = [\texttt{Male}, 38, \$1938, 24]^T$. Similar to the toy setting, we observe a %42 decrease in effort required of the individual when using the action by our method, since our cost function states that waiting for six years to get the credit approved is more costly than applying the following year for a lower (−\$800) credit amount. We extend our analysis to a population level, and observe that for 50 negatively affected test individuals, previous approaches suggest actions that are on average %39 ± %24 and %65 ± %8 more costly than our approach when considering, respectively, a logistic regression and a decision tree as the predictive model h.

The demonstrations above confirm our theoretical analysis that MINT-based actions from (3) are less costly and thus more beneficial for affected individuals than existing CFE-based actions from (2) that fail to utilize the causal relations between variables.

[5] We remark that the presented formulation also holds for more general SCMs (for example where the exogenous variable contribution is not additive) as long as the sequence of structural equations $\mathbb{F}$ is invertible, i.e., there exists a sequence of equations $\mathbb{F}^{-1}$ such that $\boldsymbol{x} = \mathbb{F}(\mathbb{F}^{-1}(\boldsymbol{x}))$ (in other words, the exogenous variables are uniquely identifiable via the abduction step).

5 TOWARDS REALISTIC INTERVENTIONS

In Section 4, we formulated algorithmic recourse by considering the causal relations between features in the real world. Our formulation minimized the cost of actions, which were carried out as *structural* interventions on the corresponding graph. Each intervention proceeds by *unconditionally severing all edges* incident on the intervened node, fixing the post-manipulation distribution of a *single* variable to *one deterministic* value. While intuitive appealing and powerful, structural interventions are in many ways the simplest type of interventions, and their "simplicity comes at a price: foregoing the possibility of modeling many situations realisitically" [8, 22]. Below, we extend (3) and (7) to add flexibility and realism to the types of interventions performed by the individual. Notably, there is nothing inherent to an SCM that a priori determines the *form*, *feasibility*, or *scope* of intervention; instead, these choices are delegated to the individual and are made based on a semantic understanding of the modeled variables.

5.1 On the Form of Interventions

The demonstrations in Section 4.3 primarily focused on actions performed as *structural (a.k.a., hard)* interventions [34] where all incoming edges to the intervened node are severed (see (7)). Hard interventions are particularly useful for Randomized Control Trial (RCT) settings where one aims to evaluate (isolate) the causal effect of an action (e.g., effect of aspirin on patients with migraine) on the population by randomly assigning individuals to treatment/control groups, removing the influence of other factors (e.g., age).

In the context of algorithmic recourse, however, an individual performs actions in the real world, and therefore must play the rules governing the world. In earlier sections, these rules (captured in an SCM) guided the search for an optimal set of actions by modelling actions along with their consequences. The rules also determine the form of an intervention, e.g., specifying whether an intervention cancels out or complements existing causal relations.

For instance, consider **Example 1**, where an individual chooses to increase their bank balance (e.g., through borrowing money from family, i.e., a deliberate action/intervention on X_2 while continuing to put aside a portion of their income (i.e., retaining the relation $X_2 := 3/10 \cdot X_1 + U_2$). Indeed, it would be unwise for a recommendation to suggest abandoning saving habits. In such a scenario, the action would be carried out as an *additive (a.k.a., soft)* intervention [10]. Such interventions *do not* sever graphical edges incident on the intervened node and continue to allow for parents of the node to affect that node. Conversely, in **Example 2**, recourse recommendations may suggest performing a structural intervention on temperature, e.g., by creating a climate controlled green-house, to cancel the natural effect of altitude change on temperature.

The previous examples illustrate a scenario where an individual/agriculture team actually have the agency to choose which type of intervention to perform. However, it is easy to conceive of examples where such an option does not exist. For instance, as part of a medical system's recommendation, we might consider adding 5 mg/l of insulin to a patient with diabetes with a certain blood insulin level [36]. This action cannot disable pre-existing mechanisms regulating blood insulin levels and therefore, the action can only be performed additively. Conversely, one may also consider another example from the medical domain whereby the only treatment of malignancy may be through a surgical (structural) amputation.[6]

Just as structural interventions were supported in our framework via a closed-form expression (see (7)), additive interventions can be encoded through an analogous assignment formulation:

$$x_i^{\mathsf{SCF}} = [i \in I] \cdot \delta_i + \left(x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}})\right). \tag{8}$$

The choice of whether interventions should be applied in a additive/soft or structural/hard manner depends on the variable semantic [3], and should be decided prior to solving (3).

5.2 On the Feasibility of Interventions

We saw in Section 3 that earlier works motivated the addition of *feasibility* constraints as a means to provide more actionable recommendations for the individual seeking recourse [49]. There, the *actionability* (a.k.a. *mutability*) of a feature was determined based on the feature semantic and value in the factual instance, marking those features which the individual has/lacks the agency to change (e.g., bank balance vs. race). While the interchangeable use of definition holds under an independent world, it fails when operating in most real-world settings governed by a set of causal dependencies. We study this subtlety below.

In an independent world, any change to variable X_i could come about only via an intervention on X_i itself. Therefore, immutable and non-actionable variables overlap. In a dependent world, however, changes to variable X_i may arise from an intervention on X_i or through changes to any of the ancestors of X_i. In this more general setting, we can tease apart the definition of *actionability* and *mutability*, and distinguish between three types of variables: (i) immutable (and hence non-actionable), e.g., race; (ii) mutable but non-actionable, e.g., credit score; and (iii) actionable (and hence mutable), e.g., bank balance. Each type requires special consideration which we show can be intuitively encoded as constraints amended to $\mathbf{A} \in \mathcal{F}$ from (3).

Immutable: We posit that the set of immutable (and hence non-actionable) variables should be closed under ancestral relationships given by the model, $\mathcal{M}$. This condition parallels the ancestral closure of *protected* attributions in [23]. This would ensure that under no circumstance would an intervention on an ancestor of an immutable variable change the immutable variable. Therefore, for an immutable variable X_i, the constraint $[i \notin I] = 1$ recursively necessitates the fulfillment of additional constraints $[j \notin I] = 1 \; \forall \; j \in \mathbf{pa}_i$ in $\mathcal{F}$. For instance, the immutability of race triggers the immutability of birthplace.

Mutable but non-actionable: To encode the conditions for mutable but non-actionable variables, we note that while a variable may not be directly actionable, it may still change as a result of changes to its parents. For example, the financial credit score in Figure 3 may change as a result of interventions to salary or savings, but is not itself directly intervenable. Therefore, for a non-actionable but mutable variable X_i, the constraint $[i \notin I] = 1$ is sufficient and does not induce any other constraints.

Actionable: In the most general sense, the actionable feasibility of an intervention on X_i may be contingent on a number of conditions, as follows: (a) the pre-intervention value of the intervened

[6] See, e.g., https://www.cancer.org/cancer/bone-cancer/treating/surgery.html.

variable (i.e., x_i^{F}); (b) the pre-intervention value of other variables (i.e., $\{x_j^{\mathsf{F}}\}_{j\subset[d]\setminus i}$); (c) the post-intervention value of the intervened variable (i.e., x_i^{SCF}); and (d) the post-intervention value of other variables (i.e., $\{x_j^{\mathsf{SCF}}\}_{j\subset[d]\setminus i}$). Such feasibility conditions can easily be encoded into $\mathcal{F}$; consider the following scenarios:

(a) an individual's age can only increase, i.e., $[x_{age}^{\mathsf{SCF}} \geq x_{age}^{\mathsf{F}}]$; (b) an individual cannot apply for credit on a temporary visa, i.e., $[x_{visa}^{\mathsf{F}} = \texttt{PERMANENT}] \geq [x_{credit}^{\mathsf{SCF}} = \texttt{TRUE}]$;

(c) an individual may undergo heart surgery (an additive intervention) only if they won't remiss due to sustained smoking habits, i.e., $[x_{heart}^{\mathsf{SCF}} \neq \texttt{REMISSION}]$; and

(d) an individual may undergo heart surgery only *after* their blood pressure is regularized due to medicinal intervention, i.e., $[x_{bp}^{\mathsf{SCF}} = \texttt{O.K.}] \geq [x_{heart}^{\mathsf{SCF}} = \texttt{SURGERY}]$.

In summary, while previous works on algorithmic recourse distinguished between actionable, conditionally actionable,[7] and immutable variables [49], we can now operate on a more realistic *spectrum* of variables, ranging from conditionally soft/hard actionable, to non-actionable but mutable, and finally to immutable and non-actionable variables. Finally, we remind that feasibility is a distinct notion from plausibility; whereas the former restricts actions $\mathbf{A} \in \mathcal{F}$ to those that can be performed by the individual, the latter determines the likeliness of the counterfactual instance $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}})) \in \mathcal{P}$ resulting from those actions. For instance, building on the earlier example, although an individual with similar attributes and higher credit score may exist in the dataset (i.e., plausible), directly acting on credit score is not feasible.

5.3 On the Scope of Interventions

One final assumption has been made throughout our discussion of actions as interventions which pertain to the one-to-one mapping between an action in the real world and an intervention on a endogenous variable in the structural causal model (which in turn are also input features to the predictive model). As exemplified in [3], it is possible for some actions (e.g., finding a higher-paying job) to simultaneously intervene on multiple variables in the model (e.g., income and length of employment). Alternatively, for **Example 2**, choosing a new paddy location is equivalent to intervening jointly on several input features of the predictive model (e.g., altitude, radiation, precipitation). Such confounded/correlated interventions, referred to as *fat-hand*/*non-atomic* interventions [10], will be explored further in follow-up work, by modelling the world at different causally consistent levels [4, 40].

6 DISCUSSION

In this paper, we have focused on the problem of algorithmic recourse, i.e., the process by which an individual can change their situation to obtain a desired outcome from a machine learning model. First, using the tools from causal reasoning (i.e., structural interventions and counterfactuals), we have shown that in their current form, counterfactual explanations only bring about agency for the individual to achieve recourse in unrealistic settings. In other words, counterfactual explanations do not translate to an *optimal* or *feasible* set of actions that would favourably change the prediction of h if acted upon. This shortcoming is primarily due to the lack of consideration of causal relations governing the world and thus, the failure to model the downstream effect of actions in the predictions of the machine learning model. In other words, although "counterfactual" is a term from causal language, we observed that existing approaches fall short in terms of taking causal reasoning into account when generating counterfactual explanations and the subsequent recourse actions. Thus, building on the statement by Wachter et al. [52] that counterfactual explanations "do not rely on knowledge of the causal structure of the world," it is perhaps more appropriate to refer to existing approaches as *contrastive*, rather than *counterfactual*, explanations [6, 30].

To directly take causal consequences of actions into account, we have proposed a fundamental reformulation of the recourse problem, where actions are performed as interventions and we seek to minimize the cost of performing actions in a world governed by a set of (physical) laws captured in a structural causal model. Our proposed formulation in (3), complemented with several examples and a detailed discussion, allows for *recourse through minimal interventions* (MINT), that when performed will result in a *structural counterfactual* that favourably changes the output of the model.

Next, we discuss the work most closely related to ours, the main limitation of the proposed recourse approach, and propose future venues for research to address such shortcomings.

Related work. A number of authors have argued for the need to consider causal relations between variables [18, 31, 49, 52], generally based on the intuition that changing some variables may have effects on others. In the original counterfactual explanations work, Wachter et al. [52] also suggest that "counterfactuals generated from an accurate causal model may ultimately be of use to experts (e.g., to medical professionals trying to decide which intervention will move a patient out of an at-risk group)". Despite this general agreement, to the best of our knowledge, only two works have attempted to technically formulate this requirement.

In the first work, Joshi et al. [17] study recourse in causal models under confounders and with predetermined treatment variables. In this work, a distribution over hidden confounders is first estimated along with a mapping from the attributes $\boldsymbol{x}$ to hidden confounders, i.e., $G_\theta^{-1}(\boldsymbol{x}) = \boldsymbol{z}$. Then, under each intervention on treatment variables, explanations are generated following (1) with the plausibility term constraining the inverse of the counterfactual instance (i.e., $G_\theta^{-1}(\boldsymbol{x})$) to the approximated confounding distribution. In this work, we instead optimize for recourse actions rather than counterfactual instances that result from those action.

In the second work, Mahajan et al. [28] present a modified version of the distance function in (1), amending the *standard proximity loss* between factual and counterfactual instances with a *causal regularizer* to encourage the counterfactual value of each endogenous variable to be close to the value of that variable had it been assigned via its structural equation. Beyond the uncertainty regarding the strength of regularization (which would mean causal relations may not be guaranteed), and why the standard proximity loss only iterates over the exogenous variables (which from a causal perspective, are characteristics that are shared across counterfactual worlds [23,

[7] Ustun et al. [49] also support conditionally actionable features (e.g., age or educational degree) with conditions derived only from x_i^{F} as in (a). We generalize the set of conditions to support actions conditioned on the value of other variables as in (b), additive interventions in (c), and sequential interventions as in (d).

footnote 4]), this approach suffers from a primary limitation in its causal treatment: the causal regularizer would penalize any variable whose value deviated away from its structurally assigned value. While on the surface this "preservation of causal relations" seems beneficial, such an approach would discourage interventions (additive or structural) on non-root variables, which would, by design, change the value of the intervened-upon variable away from its structurally assigned value. Instead, the regularizer would encourage interventions on variables that would not be penalized as such, i.e., root variables, which may not be contextually acceptable as root notes typically capture sensitive characteristic of the individual (e.g., birthplace, age, gender). The authors suggest (in the Appendix of [28]) that one may consider those variables, upon which (structural) interventions are to be performed, as exogenous. In this manner, interventions would not be penalized and down-stream effects of interventions would still be preserved when searching for the nearest counterfactual instance. We argue, however, that such an approach suffers from the same limitations as other CFE-based recourse approaches presented in Section 3.2 in that a returned counterfactual instance would not imply feasible or optimal actions for recourse. Finally, without an explicit abduction step and without assumptions on the form of structural equations, it is unclear how the authors infer and combine individual-specific characteristics (as embedded in the background variables) with the effect of ancestral changes to compute the counterfactual. We believe the problems above will be mostly resolved when minimizing over the cost of actions instead of distance over counterfactuals as we have done in this work.

Practical limitations. The primary limitation of our formulation in (3) is its reliance on the true causal model of the world, subsuming both the graph, and the structural equations. In practice, the underlying causal model is rarely known, which suggests that the counterfactual constraint in (3), i.e., $\boldsymbol{x}^{\mathsf{SCF}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathsf{F}}))$, may not be (deterministically) identifiable. We believe this is a valid criticism, not just of our work, but of any approach suggesting actions to be performed in the real world for consequential decision-making. Importantly, beyond recourse, the community on algorithmic fairness has echoed the need for causal counterfactual analysis for fair predictions, and have also voiced their concern about untestable assumptions when the true SCM is not available [2, 5, 19, 23, 44].

Perhaps more concerningly, our work highlights the implicit causal assumptions made by existing approaches (i.e., that of independence, or feasible and cost-free interventions), which may portray a false sense of recourse guarantees where one does not exists (see **Example 2** and all of Section 3.2). Our work aims to highlight existing imperfect assumptions, and to offer an alternative formulation, backed with proofs and demonstrations, which would guarantee recourse if assumptions about the causal structure of the world were satisfied. Future research on causal algorithmic recourse may benefit from the rich literature in causality that has developed methods to verify and perform inference under various assumptions [37]. Thus, we consider further discussion on causal identifiability to be out of scope of this paper, as it remains as an open and key question in the Ethical ML community.

This is not to say that counterfactual explanations should be abandoned altogether. On the contrary, we believe the counterfactual explanations hold promise for "guided audit of the data" [52] and evaluating various desirable model properties, such as robustness [16, 45] or fairness [15, 18, 45, 49]. Besides this, it has been shown that designers of interpretable machine learning systems use counterfactual explanations for predicting model behavior [24] or uncovering inaccuracies in the data profile of individuals [50]. Complementing these offerings of counterfactual explanations, we offer minimal interventions as a way to guarantee algorithmic recourse in general settings, which is not implied by counterfactual explanations.

Future work. In future work, we aim to focus on overcoming the main assumption of our formulation: the availability of the true world model, $\mathcal{M}$. An immediate first step involves learning the true world model (partially or fully) [9, 12, 29], and studying potential inefficiencies that may arise from partial or imperfect knowledge of the causal model governing the world. Furthermore, while additive noise models are a broadly used class of SCMs for modeling real-world systems, further investigation into the effects of confounders (non-independent noise variables), the presence of only the causal graph, as well as cyclic graphical models for time series data (e.g., conditional interventions), would extend the reach of algorithmic recourse to even broader settings.

In Section 5, we presented feasibility constraints for a wide range of settings, including dynamical settings in which one intervention enables the preconditions of another. An interesting line of future research would involve combining the causal intervention-based recourse framework, as presented in our work, with multi-stage planning strategies such as [39] to generate optimal sequential actions.

Finally, the examples presented in relation to the form and feasibility of intervention serve only to illustrate the flexibility of our formulation in supporting a variety of real-world constraints. They do not, however, aim to provide an authoritative definition of how to interpret variables and the context- and individual-dependent constraints for recourse as highlighted by other works [3, 21]. Future cross-disciplinary research would benefit from accurately defining the variables and relationships and types of permissible interventions in consequential decision-making settings. Relatedly, future research would also benefit from a study of properties that cost functions should satisfy (e.g., individual-based or population-based, monotonicity) as the primary means to measure the effort endured by the individual seeking recourse.

ACKNOWLEDGMENTS

The authors would like to thank Adrián Javaloy Bornás and Julius von Kügelgen for their valuable feedback on drafts of the manuscript.

REFERENCES

[1] Kevin Bache and Moshe Lichman. 2013. UCI machine learning repository.

[2] Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, and Jonathan Zittrain. 2017. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv preprint arXiv:1712.08238* (2017).

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

[4] Sander Beckers and Joseph Y Halpern. 2019. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2678–2685.

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

[6] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[8] Frederick Eberhardt. 2007. *Causation and intervention*. PhD dissertation. California Institute of Technology.

[9] Frederick Eberhardt. 2017. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics* 3, 2 (2017), 81–91.

[10] Frederick Eberhardt and Richard Scheines. 2007. Interventions and causal inference. *Philosophy of science* 74, 5 (2007), 981–995.

[11] Dorothy Edgington. 2014. Indicative Conditionals. In *The Stanford Encyclopedia of Philosophy* (winter 2014 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[12] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* 10 (2019).

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[14] David Gunning. 2019. DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, ii–ii.

[15] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. *arXiv preprint arXiv:1909.03166* (2019).

[16] Leif Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 640–647.

[17] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. REVISE: Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv preprint arXiv:1907.09615* (2019).

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

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

[20] Yves Kodratoff. 1994. The comprehensibility manifesto. *KDD Nugget Newsletter* 94, 9 (1994).

[21] Issa Kohler-Hausmann. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.* 113 (2018), 1163.

[22] Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. 2004. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 322–331.

[23] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[24] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[25] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. 2013. Causal responsibility and counterfactuals. *Cognitive science* 37, 6 (2013), 1036–1073.

[26] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse Classification for Comparison-based Interpretability in Machine Learning. *arXiv preprint arXiv:1712.08443* (2017).

[27] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[28] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *arXiv preprint arXiv:1912.03277* (2019).

[29] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.

[30] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[31] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2019. DiCE: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *arXiv preprint arXiv:1905.07697* (2019).

[32] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[33] Judea Pearl. 1994. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*. Elsevier, 454–462.

[34] Judea Pearl. 2000. *Causality: models, reasoning and inference*. Vol. 29. Springer.

[35] Judea Pearl. 2013. Structural counterfactuals: A brief introduction. *Cognitive Science* 37, 6 (2013), 977–985.

[36] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

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

[38] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2019. FACE: Feasible and Actionable Counterfactual Explanations. *arXiv preprint arXiv:1909.09369* (2019).

[39] Goutham Ramakrishnan, Yun Chan Lee, and Aws Albargouthi. 2019. Synthesizing Action Sequences for Modifying Model Decisions. *arXiv preprint arXiv:1910.00057* (2019).

[40] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2017. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819* (2017).

[41] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[42] Stefan Rüping. 2006. *Learning interpretable models*. PhD dissertation. Technical University of Dortmund.

[43] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 20–28. https://doi.org/10.1145/3287560.3287569

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

[45] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. *arXiv preprint arXiv:1905.07857* (2019).

[46] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[47] William Starr. 2019. Counterfactuals. In *The Stanford Encyclopedia of Philosophy* (fall 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[48] William Starr. 2019. Counterfactuals. In *The Stanford Encyclopedia of Philosophy* (fall 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

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

[50] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

[51] Paul Voigt and Axel Von dem Bussche. [n.d.]. The EU General Data Protection Regulation (GDPR). ([n. d.]).

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

[53] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

A PROOFS

A.1 Proof of Proposition 3.1

Proposition 3.1. *A CFE-based action, $\mathbf{A}^{\mathsf{CFE}}$, where $I = \{i \mid \delta_i^* \neq 0\}$, performed by individual $\boldsymbol{x}^{\mathsf{F}}$, in general results in the structural counterfactual, $\boldsymbol{x}^{\mathsf{SCF}} = \boldsymbol{x}^{*\mathsf{CFE}} := \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^*$, and thus guarantees recourse (i.e., $h(\boldsymbol{x}^{\mathsf{SCF}}) \neq h(\boldsymbol{x}^{\mathsf{F}})$), if and only if, the set of descendants of the acted upon variables, determined by I, is the empty set.*

Proof. The setting assumes that the causal graph $\mathcal{G}$ is available such that the parent set for each variable is known. Let $\mathsf{d}(X)$ and $\mathsf{nd}(X)$ denote the sets of descendants and non-descendants of the variable X according to $\mathcal{G}$, respectively. For multiple intervened-upon variables, we define:

$$\mathbb{X}_I := \{X_i\}_{i \in I},$$
$$\mathsf{nd}(\mathbb{X}_I) := \cap_{i \in I} \mathsf{nd}(X_i),$$
$$\mathsf{d}(\mathbb{X}_I) := \mathbb{X} \setminus (\mathbb{X}_I \cup \mathsf{nd}(\mathbb{X}_I)).$$

Note that, by definition, $\mathbb{X}_I$, $\mathsf{nd}(\mathbb{X}_I)$, and $\mathsf{d}(\mathbb{X}_I)$ form a partition of the set of all variables $\mathbb{X}$.

To prove the iff conditional, we prove each direction separately. For ease of exposition, we define

$$\underbrace{\boldsymbol{x}^{\mathsf{SCF}} = \boldsymbol{x}^{*\mathsf{CFE}} := \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^*}_{\mathrm{p}} \quad \iff \quad \underbrace{\mathsf{d}(\mathbb{X}_I) = \varnothing}_{\mathrm{q}}$$

where we recall the remark that given $\boldsymbol{\delta}^*$, an individual seeking recourse may intervene on any arbitrary subset of observed variables $\mathbb{X}_I$, as long as $(\delta_i^* \neq 0) \implies (i \in I)$.

$\mathbf{q} \implies \mathbf{p}$: Borrowing the closed-form expression of a structural counterfactual from (**??**), we have

$$x_i^{\mathsf{SCF}} = \begin{cases} x_i^{\mathsf{F}} + \delta_i^* & i \in I \\ x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}}) & i \notin I \end{cases} \tag{A.1}$$

which can be broken down further to specify the descendants and non-descendants of intervened upon variables, as

$$x_i^{\mathsf{SCF}} = \begin{cases} x_i^{\mathsf{F}} + \delta_i^* & i \in I \\ x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}}) & i \in \mathsf{d}(\mathbb{X}_I) \\ x_i^{\mathsf{F}} + f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}}) & i \in \mathsf{nd}(\mathbb{X}_I) \end{cases} \tag{A.2}$$

By assumption, $\mathsf{d}(\mathbb{X}_I) = \varnothing$, so the second case never holds.

Furthermore, since structural interventions leave non-descendant variables unaffected, we have that

$$\mathbf{pa}_i^{\mathsf{SCF}} = \mathbf{pa}_i^{\mathsf{F}} \quad \forall i \in \mathsf{nd}(\mathbb{X}_I).$$

Consequently,

$$f_i(\mathbf{pa}_i^{\mathsf{SCF}}) - f_i(\mathbf{pa}_i^{\mathsf{F}}) = f_i(\mathbf{pa}_i^{\mathsf{F}}) - f_i(\mathbf{pa}_i^{\mathsf{F}}) = 0 \quad \forall i \in \mathsf{nd}(\mathbb{X}_I).$$

In summary, we have

$$x_i^{\mathsf{SCF}} = \begin{cases} x_i^{\mathsf{F}} + \delta_i^* & i \in I \\ x_i^{\mathsf{F}} & i \in \mathsf{nd}(\mathbb{X}_I) \end{cases} \tag{A.3}$$

which, upon realising that $(\delta_i^* \neq 0) \implies (i \in I)$, reduces to $\boldsymbol{x}^{\mathsf{SCF}} = \boldsymbol{x}^{*\mathsf{CFE}} := \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^*$ as desired.

$\neg\mathbf{q} \implies \neg\mathbf{p}$: Starting with the negation of $\mathbf{q}$, we have the $\exists\, k \in I$ s.t. $\mathsf{d}(X_k) \neq \varnothing$. It is assumed that $\delta_k^* \neq 0$ (i.e., we are not performing a non-altering intervention on X_k), then using the same expression for structural counterfactuals in (A.2), there in general exists a descendant of X_k for which the value of its ancestors change under intervention, i.e., $\exists\, l \in \mathsf{d}(\mathbb{X}_I)$ s.t. $f_l(\mathbf{pa}_l^{\mathsf{SCF}}) - f_l(\mathbf{pa}_l^{\mathsf{F}}) \neq 0$. Thus, $x_l^{\mathsf{SCF}} \neq x_l^{\mathsf{F}}$ and thus $\boldsymbol{x}^{\mathsf{SCF}} \neq \boldsymbol{x}^{*\mathsf{CFE}} := \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}^*$. Our proof ignores special cases such as piece-wise constant structural equations, where for some $\delta_i^* \neq 0$, the descendant of X_i remains invariant. These rare cases can be thought of as locally violating causal minimality [37, Sec. 6.5] and are thus disregarded. □

A.2 Proof of Corollary 3.1

Corollary 3.1. *If the true world $\mathcal{M}$ is independent, i.e, all the observed features are root-nodes, then CFE-based actions always guarantee recourse.*

Proof. If the true world $\mathcal{M}$ is independent, then by definition the set of descendants for all variables is the empty set. Thus, the statement follows directly from Proposition 3.1. □

A.3 Proof of Proposition 4.1

Proposition 4.1. *Given an individual $\boldsymbol{x}^{\mathsf{F}}$ observed in world $\mathcal{M} \in \Pi$, a family of feasible actions $\mathcal{F}$, and the solution of* (3)*, $\mathbf{A}^* \in \mathcal{F}$. Assume that there exists CFE-based action $\mathbf{A}^{\mathsf{CFE}} \in \mathcal{F}$ that achieves recourse, i.e., $h(\boldsymbol{x}^{\mathsf{F}}) \neq h(\boldsymbol{x}^{*\mathsf{CFE}})$. Then,* $\mathrm{cost}(\mathbf{A}^*; \boldsymbol{x}^{\mathsf{F}}) \leq \mathrm{cost}(\mathbf{A}^{\mathsf{CFE}}; \boldsymbol{x}^{\mathsf{F}})$.

Proof. Having assumed that both $\mathbf{A}^{\mathsf{CFE}}, \mathbf{A}^* \in \mathcal{F}$, and considering that $\mathbf{A}^*$ is the optimal solution of (3) constrained to $\mathcal{F}$, it follows from definition of optimality that $\mathrm{cost}(\mathbf{A}^*; \boldsymbol{x}^{\mathsf{F}}) \leq \mathrm{cost}(\mathbf{A}^{\mathsf{CFE}}; \boldsymbol{x}^{\mathsf{F}})$. □