ALGORITHMIC RECOURSE: FROM COUNTERFACTUAL EXPLANATIONS TO INTERVENTIONS

Amir-Hossein Karimi, Bernhard Schölkopf, Isabel Valera
Max Planck Institute for Intelligent Systems
Tübingen, Germany
{amir, bs, isabel.valera}@tuebingen.mpg.de

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, one of the main objectives of "explanations as a means to help a data-subject *act* rather than merely *understand*" has been overlooked. 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*, moving the focus from explanations to recommendations.

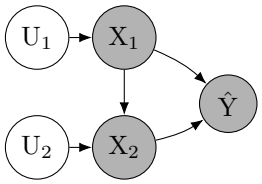
1 INTRODUCTION

Predictive models are being increasingly used to support consequential decision-making in contexts, e.g., denying a loan, rejecting a job applicant, or prescribing life-altering medication. As a result, there is increasing social and legal pressure Voigt & Von dem Bussche (2017) 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 focuses of explainable machine learning Kodratoff (1994); Rüping (2006); Doshi-Velez & Kim (2017); Lipton (2018); Rudin (2018); Gunning (2019).

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" Wachter et al. (2017). 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 Laugel et al. (2017); Karimi et al. (2019). A closely related term is *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 Venkatasubramanian & Alfano (2020).

Counterfactual explanations have shown promise for practitioners and regulators to validate a model on metrics such as fairness and robustness Ustun et al. (2019); Sharma et al. (2019); Karimi et al. (2019). 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*" Wachter et al. (2017).

The translation of counterfactual explanations to a recommendable set of actions (recourse) was first explored by Ustun et al. (2019), 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 Sharma et al. (2019);



$$\left.\begin{aligned} \mathrm{X}_1 &:= \mathrm{U}_1 \\ \mathrm{X}_2 &:= f_2(\mathrm{X}_1) + \mathrm{U}_2 \end{aligned}\right\} \mathcal{M}$$
$$\hat{\mathrm{Y}} = h_\theta(\mathrm{X}_1, \mathrm{X}_2)$$

Figure 1: Illustration of an example causal data generative process governing the world, showing both the graphical model, $\mathcal{G}$, and the structural causal model, $\mathcal{M}$, Pearl (2000). In this example, X_1 represents an individual's annual salary, X_2 is bank balance, and $\hat{\mathrm{Y}}$ is the output of a fixed deterministic predictor h_θ, predicting the eligibility of an individual to receive a loan.

Karimi et al. (2019); Mothilal et al. (2019); Poyiadzi et al. (2019) 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). 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.[1]

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 $h_\theta = \mathrm{sgn}(\mathrm{X}_1 + 5 \cdot \mathrm{X}_2 - \$225,000)$. 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., $\mathrm{X}_2 := 3/10 \cdot \mathrm{X}_1 + \mathrm{U}_2$), a salary increase of only $\%14$ to $\$85,000$ would 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_θ(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 state 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 indeed result in the prediction *not* changing.

The two examples above show the pitfalls of generating a recommendable set of actions directly from counterfactual explanations without consideration for the world structure in which the actions will be performed. Acting on the recommendations derived directly from counterfactual explanations, is asking the individual in Example #1 for *too much effort*, and for effort that *does not* even result in the desired output in Example #2. We remedy this situation via a fundamental reformulation of the recourse problem, and incorporate knowledge of causal dependencies into the process of generating recommendations, that if acted upon would result in a counterfactual explanation, i.e., an instance that favourably changes the output of h_θ.

Our Contributions: In this paper, we first provide a causal analysis to illuminate the intrinsic limitations of the setting in which recommendations directly follow counterfactual explanations. Importantly, we show that even when equipped with knowledge of causal dependencies after-the-fact, generating recommendations from pre-computed (nearest) counterfactual explanations may prove sub-optimal. Second, in order to solve the above limitations, we propose a fundamental reformulation of the recourse problem, which relies on tools of structural counterfactuals to directly incorporate causal dependencies for a broad class of causal models. The resulting *Recourse through Minimal Interventions* thus informs stakeholders on how to act in addition to understand.

[1] See Barocas et al. (2020) for additional examples.

2 OVERVIEW OF COUNTERFACTUAL EXPLANATIONS

Counterfactual explanations are statements of "how the world would have (had) to be different for a desirable outcome to occur" Wachter et al. (2017). In the context of explainable machine learning, the literature has focused on finding *nearest counterfactual explanations* (i.e., instances),[2] 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, $\mathrm{dist}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. This problem has been formulated as the following optimization problem Wachter et al. (2017):

$$\begin{aligned} \boldsymbol{x}_*^{\mathrm{CFE}} \in \arg\min_{\boldsymbol{x}} \quad & \mathrm{dist}(\boldsymbol{x}, \boldsymbol{x}^{\mathrm{F}}) \\ \text{s.t.} \quad & h_\theta(\boldsymbol{x}) \neq h_\theta(\boldsymbol{x}^{\mathrm{F}}) \\ & \boldsymbol{x} \in \mathcal{P}\mathrm{lausible} \ , \end{aligned} \tag{1}$$

where $\boldsymbol{x}^{\mathrm{F}} \in \mathcal{X}$ is the factual instance; $\boldsymbol{x}_*^{\mathrm{CFE}} \in \mathcal{X}$ is a (perhaps not unique) nearest counterfactual instance; h_θ is the fixed predictor; and $\mathcal{P}$ is an optional set of *plausibility* constraints, e.g., the instance be from a high density region of the input space Joshi et al. (2019); Poyiadzi et al. (2019).

Most of the existing approaches in the counterfactual explanations literature have focused on providing solutions to the optimization problem in equation 1, by exploring semantically meaningful distance/dissimilarity functions $\mathrm{dist}(\cdot, \cdot)$ between individuals (e.g., $\ell_0, \ell_1, \ell_\infty$, percentile-shift), accommodating different predictive models h_θ (e.g., random forest, multilayer perceptron), and realistic plausibility constraints, $\mathcal{P}$. In particular, Wachter et al. (2017) and Mothilal et al. (2019) solve equation 1 using gradient-based optimization; Russell (2019) and Ustun et al. (2019) employ mixed-integer linear program solvers to support mixed numeric/binary data; Poyiadzi et al. (2019) use graph-based shortest path algorithms; Laugel et al. (2017) use a heuristic search procedure by growing spheres around the factual instance; Guidotti et al. (2018) and Sharma et al. (2019) build on genetic algorithms for model-agnostic behavior; and Karimi et al. (2019) solve equation 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, in counterfactual explanations, of the actions required to realize $\boldsymbol{x}_*^{\mathrm{CFE}}$ from $\boldsymbol{x}^{\mathrm{F}}$ leads to uncertainty and limited agency for the individual seeking recourse. In the next section, we elucidate the process of achieving a desired output, i.e., realizing a [nearest] counterfactual explanation via a [minimal] set of recommendable actions.

3 RECOURSE VIA COUNTERFACTUAL EXPLANATIONS

As the focus shifts away from finding [nearest] counterfactual explanations to obtaining the [minimal] set of recommendable actions that result in such explanations, we here follow Ustun et al. (2019) to rewrite equation 1 as:

$$\begin{aligned} \boldsymbol{\delta}^* \in \arg\min_{\boldsymbol{\delta}} \quad & \mathrm{cost}(\boldsymbol{\delta}; \boldsymbol{x}^{\mathrm{F}}) \\ \text{s.t.} \quad & h_\theta(\boldsymbol{x}^{\mathrm{CFE}}) \neq h_\theta(\boldsymbol{x}^{\mathrm{F}}) \\ & \boldsymbol{x}^{\mathrm{CFE}} = \boldsymbol{x}^{\mathrm{F}} + \boldsymbol{\delta} \\ & \boldsymbol{x}^{\mathrm{CFE}} \in \mathcal{P}\mathrm{lausible} \\ & \boldsymbol{\delta} \in \mathcal{F}\mathrm{easible} \ , \end{aligned} \tag{2}$$

where $\mathrm{cost}(\cdot; \boldsymbol{x}^{\mathrm{F}})\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a user-specified cost that encodes preferences between feasible actions from $\boldsymbol{x}^{\mathrm{F}}$, and $\mathcal{F}$ and $\mathcal{P}$ are optional sets of feasibility and plausibility constraints,[3] restricting the actions and the resulting counterfactual explanation, respectively. The feasibility constraints in equation 2, as introduced by Ustun et al. (2019), aim at restricting the set of features that the

[2] A counterfactual instance can be from the dataset Wexler et al. (2019); Poyiadzi et al. (2019), or generated as in Wachter et al. (2017); Ustun et al. (2019); Karimi et al. (2019) among others.

[3] Note the difference in definition: "feasible" means possible to do, whereas "plausible" means possibly true, believable or realistic.

individual may act upon.[4] For instance, recommendations should not ask an individual to reduce their age. The seemingly innocent reformulation of equation 1 as equation 2 is founded on two assumptions:

A1: the feature-wise vector difference between factual and nearest counterfactual instances, $\boldsymbol{\delta}^* = \boldsymbol{x}_*^{\text{CFE}} - \boldsymbol{x}^{\text{F}}$, directly translates to the minimal action set, $\mathbb{A}^*$, i.e., performing the actions in $\mathbb{A}^*$ starting from $\boldsymbol{x}^{\text{F}}$ will result in $\boldsymbol{x}_*^{\text{CFE}}$; and

A2: there is a 1-1 mapping between $\text{dist}(\cdot, \cdot)$ and $\text{cost}(\cdot; \cdot)$, whereby larger actions incur higher cost and larger distance.

Unfortunately, these assumptions only hold in restrictive settings, rendering $\mathbb{A}^*$ *sub-optimal* or *infeasible* in many real-world scenarios. Specifically, **A1** holds only if (i) the individual applies effort in a world where changing a variable does not affect other variables (i.e., features are independent); 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 **A2** does not hold either.

Therefore, except in trivial cases where the model designer actively inputs pair-wise independent features to h_θ, generating recommendations from counterfactual explanations in this manner, i.e., ignoring the dependencies between features, warrants reconsideration. We formalize these shortcomings using the language of causality.

3.1 A Causal Perspective: Actions as Interventions

Let $\mathcal{M} \in \Pi$ be a 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 variables, $\mathbb{X} \in \mathcal{X}$, the exogenous variables, $\mathbb{U} \in \mathcal{U}$, and a sequence of structural equations $\mathbb{F}\colon \mathcal{U} \to \mathcal{X}$, describing how endogenous variables can be (deterministically) obtained from the exogenous variables Pearl (2000); Spirtes et al. (2000). Often, a model, $\mathcal{M}$, is illustrated using a directed graphical model, $\mathcal{G}$ (see, e.g., Figure 1).

From a causal perspective, recommendable *actions* may be carried out via *structural interventions*, $\mathbb{A}\colon \Pi \to \Pi$, which can be thought of as a transformation between SCMs Pearl (1994; 2000). For instance, the set of interventions can be constructed as $\mathbb{A} = \text{do}(\{\text{X}_i := a_i\}_{i \in I})$ 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 variable X_i in $\mathbb{F}$ with $\text{X}_i := a_i$. Correspondingly, graph surgery is performed on $\mathcal{G}$, severing graph edges incident on an intervened variable, X_i, with a single assignment corresponding to the value of the intervention, i.e., a_i. Thus, performing the actions, $\mathbb{A}$, in a world, $\mathcal{M}$, yields the updated world model $\mathcal{M}_\mathbb{A}$ with structural equations $\mathbb{F}_\mathbb{A} = \{\mathbb{F}_i\}_{\ i \notin I} \cup \{\text{X}_i := a_i\}_{i \in I}$.

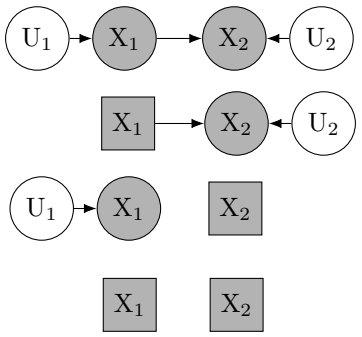
While *structural interventions* are used to predict the effect of actions on the world as a whole (i.e., how $\mathcal{M}$ becomes $\mathcal{M}_\mathbb{A}$), in the context of recourse, we desire to model the effect of actions on one individual's situation (i.e., how $\boldsymbol{x}^{\text{F}}$ becomes $\boldsymbol{x}^{\text{SCF}}$). We compute such effects using *structural counterfactuals* Pearl et al. (2016), as explained below.

Assuming that $\mathcal{M}$ factorizes as a directed acyclic graph (DAG), and full specification of $\mathbb{F}$ (and $\mathbb{F}^{-1}$, 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 background variables that give rise to a particular realization of the endogenous variables, $\{\text{X}_i = x_i^{\text{F}}\}_i \subseteq \mathcal{X}$, as $\mathbb{F}^{-1}(\boldsymbol{x}^{\text{F}})$.[5] As a result, we can compute *any* structural counterfactuals query $\boldsymbol{x}^{\text{SCF}}$, which automatically account for inter-variable causal dependencies, for an individual $\boldsymbol{x}^{\text{F}}$ as $\boldsymbol{x}^{\text{SCF}} = \mathbb{F}_\mathbb{A}(\mathbb{F}^{-1}(\boldsymbol{x}^{\text{F}}))$, that is: "given model $\mathcal{M}$ and having observed $\boldsymbol{x}^{\text{F}}$, what *is* the value of all endogenous variables if the set of actions $\mathbb{A}$ is performed".[6]

[4] The actionability of a feature is determined based on the feature semantic and value in the factual instance.

[5] For simplicity, we slightly abuse notation and use sets and vectors alike, e.g., $\{\text{X}_i = x_i^{\text{F}}\}_i \subseteq \mathcal{X}, \boldsymbol{x}^{\text{F}} \in \mathcal{X}$.

[6] 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 Lagnado et al. (2013); Edgington (2014); Starr (2019), if we assume that the laws governing the world, $\mathbb{F}$, are stationary.



$$\left.\begin{aligned} \mathrm{X}_1 &\coloneqq \mathrm{U}_1 \\ \mathrm{X}_2 &\coloneqq f_2(\mathrm{X}_1) + \mathrm{U}_2 \end{aligned}\right\} \mathcal{M}$$

$$\left.\begin{aligned} \mathrm{X}_1 &\coloneqq a_1 \\ \mathrm{X}_2 &\coloneqq f_2(\mathrm{X}_1) + \mathrm{U}_2 \end{aligned}\right\} \mathcal{M}'_1$$

$$\left.\begin{aligned} \mathrm{X}_1 &\coloneqq \mathrm{U}_1 \\ \mathrm{X}_2 &\coloneqq a_2 \end{aligned}\right\} \mathcal{M}'_2$$

$$\left.\begin{aligned} \mathrm{X}_1 &\coloneqq a_1 \\ \mathrm{X}_2 &\coloneqq a_2 \end{aligned}\right\} \mathcal{M}'_3$$

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

We can now better understand why assumption **A1** fails: whereas $\boldsymbol{\delta}^*$ takes values in $\mathcal{X}$, thus corresponding to a shift in the feature values, the action set $\mathbb{A}^*$ is performed over the structural causal model of the world, $\mathcal{M}$, resulting not only in changes to the value of the endogenous variables, but also to the relations between variables, captured by $\mathbb{F}$. We recall that performing the actions $\mathbb{A}$ in a world model $\mathcal{M}$, yields the updated world model $\mathcal{M}_{\mathbb{A}}$. Thus, the under-specification of $\mathcal{M}$ when generating $\boldsymbol{\delta}^*$, as in equation 2, leads to uncertainty regarding the post-intervention relations. In other words, the optimization problem in equation 2, lacks proper handling of the consequences of actions.[7] From the point of view of actions as a structural interventions, given $\boldsymbol{\delta}^*$, an individual seeking recourse in a world, $\mathcal{M}$, may opt for one of two courses of action:

Option #1: Perform interventions only on the *non-zero* elements of $\boldsymbol{\delta}^*$, i.e., $\mathbb{A}^* = \mathrm{do}(\{\mathrm{X}_i \coloneqq x_i^{\mathrm{F}} + \delta_i^* \mid \delta_i^* \neq 0\}_i)$. This setup is depicted in $\mathcal{M}'_1$ and $\mathcal{M}'_2$ of Figure 2. Due to potential consequences of this set of interventions on those variables that were not intervened upon (e.g., in $\mathcal{M}'_1$, the intervention $\mathrm{do}(\mathrm{X}_1 \coloneqq a_1)$ affects both X_1 and X_2), there is no guarantee that $\boldsymbol{x}_*^{\mathrm{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathrm{F}}))$ will correspond to a counterfactual instance. In other words, due to unverified feature interactions when passing $\boldsymbol{x}^{\mathrm{SCF}}$ through h_θ, it may be that $h_\theta(\boldsymbol{x}^{\mathrm{SCF}}) \neq h_\theta(\boldsymbol{x}_*^{\mathrm{CFE}})$, and therefore $\mathbb{A}^*$ fails to serve the purpose of recourse. Illustratively, $\mathbb{A}^*$ may recommend too much (too little) effort, as in Example #1 (#2) in §1. Therefore, this option is discarded.

Option #2: Perform interventions on *every* dimension of $\boldsymbol{x}^{\mathrm{F}}$, i.e., $\mathbb{A}^* = \mathrm{do}(\{\mathrm{X}_i \coloneqq x_i^{\mathrm{F}} + \delta_i^*\}_i)$ severing all inter-variable edges (i.e., $\mathrm{X}_i \perp\!\!\!\perp \mathrm{X}_j \; \forall \; i \neq j$). This setup is depicted in $\mathcal{M}'_3$ of Figure 2. By the independent world reduction, $\boldsymbol{x}_*^{\mathrm{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathrm{F}}))$ will indeed equal $\boldsymbol{x}_*^{\mathrm{CFE}}$. However, it assumes that non-altering interventions (i.e., $\mathrm{X}_i \coloneqq x_i^{\mathrm{F}} + 0$) are *feasible* and incur *zero cost*, which is an unrealistic assumption. For instance, as in Example #1 (#2) in §1, a variable may change favorably (detrimentally) with respect to the output of the predictor h_θ, as a result of changes to its ancestors — performing a non-altering intervention to prevent [unpredictable] ancestral influence and ensure recourse is itself a costly action which is not captured in existing definitions of cost. For instance, the agriculture team in Example #2 of §1 may need to invest heavily in a greenhouse to ensure temperature remains the same if the paddy were to be situated at a different altitude. Therefore, this option is also discarded.

In summary, for general $\mathcal{M}$, neither option is acceptable: acting on recommendations generated from explanations can be potentially *sub-optimal* (Option #1 may recommend too much/too little effort, and Option #2 does not account for the cost of non-altering interventions), or *infeasible* (Option #2 may recommend non-altering interventions that are not possible). Thus, even when

[7] Consequences relate to the edges that are severed after an intervention, but also perhaps more importantly, the edges that remain in the graph. While intervened variables are no longer affected by changes to their parents, changes to intervened variables continue to consequentially affect their un-altered children.

equipped with knowledge of causal dependencies after-the-fact, generating recommendations from pre-computed counterfactual explanations in the manner of existing approaches is not satisfactory.

4 RECOURSE THROUGH MINIMAL INTERVENTIONS

To achieve algorithmic recourse, we seek a [minimal cost] set of actions, where intervening only on the elements of this set will trigger predictable consequences according to our knowledge of the world, encoded in $\mathcal{M}$, and result in a counterfactual instance giving the favourable output from h_θ. Therefore, we re-formulate equation 2 as follows:

$$\begin{aligned} \mathbb{A}^* \in \arg\min_{\mathbb{A}} \quad & \text{cost}(\mathbb{A}; \boldsymbol{x}^{\text{F}}) \\ \text{s.t.} \quad & h_\theta(\boldsymbol{x}^{\text{SCF}}) \neq h_\theta(\boldsymbol{x}^{\text{F}}) \\ & \boldsymbol{x}^{\text{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\text{F}})) \\ & \boldsymbol{x}^{\text{SCF}} \in \mathcal{P}\text{lausible} \\ & \mathbb{A} \in \mathcal{F}\text{easible} \ , \end{aligned} \tag{3}$$

where $\mathbb{A}^* \in \mathcal{A}$ directly specifies the set of actions (i.e., structural interventions) to be performed to achieve recourse at minimal cost, with $\text{cost}(\cdot; \boldsymbol{x}^{\text{F}}) \colon \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+$, and $\boldsymbol{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\boldsymbol{x}^{\text{F}}))$ denotes the resulting structural counterfactual explanation. We remark here that, while $\boldsymbol{x}_*^{\text{SCF}}$ is a counterfactual explanation, it does not need to correspond to the nearest counterfactual explanation, $\boldsymbol{x}_*^{\text{CFE}}$, resulting from equation 2 (see, e.g., Example #1 of §1). We obtain a structural counterfactual, $\boldsymbol{x}^{\text{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\text{F}}))$, by applying the abduction-action-prediction method of counterfactual reasoning Pearl (2013) The assignment of structural counterfactual values can generally be written as:

$$\begin{aligned} x_i^{\text{SCF}} = \ & [i \in I] \cdot (x_i^F + \delta_i) \\ & + [i \notin I] \cdot \big(x_i^{\text{F}} + f_i(\mathbf{pa}_i^{\text{SCF}}) - f_i(\mathbf{pa}_i^{\text{F}})\big), \end{aligned} \tag{4}$$

where we have made implicit the abduction step in previous section and replaced a_i by $x_i^F + \delta_i$ to make explicit the dependence on the factual instance. Note that equation 4 carries a natural intuition: if variable X_i is intervened on, set it to the intervened value (i.e., a_i), otherwise, offset the original value of the variable (i.e., x_i^{F}) by the difference in value of its structure equation given the factual and counterfactual values of its parent (i.e., $f_i(\mathbf{pa}_i^{\text{SCF}}) - f_i(\mathbf{pa}_i^{\text{F}})$), thus accounting for the consequences of changing other variables on this variable.

5 CONCLUSION

Our work is concerned with algorithmic recourse, i.e., the process by which an individual can change their situation to attain a desired outcome from a machine learning model. We showed that in their current form, counterfactual explanations do not bring about agency for the individual to achieve recourse. In other words, counterfactual explanations do not translate to an *optimal* or *feasible* set of recommendations that would favourably change the prediction of h_θ if acted upon. We attribute this shortcoming primarily to: a lack of consideration of causal relations governing the world and thus, the failure to model the consequences of actions.

To overcome this limitation, we argue for a fundamental reformulation of the recourse problem, by directly minimizing the cost of performing consequential actions in a world governed by a set of laws captured in a structural causal model. Our proposed formulation in equation 3, complemented with several examples and a detailed discussion, allows for *recourse through minimal interventions*, that when performed will result in a *counterfactual explanation*, i.e., an instance that favourably changes the output of the model.

In future work, we will focus on overcoming the two main assumptions of our formulation: the availability of i) the true world model, $\mathcal{M}$; and ii) the predefined cost function. An immediate first step involves learning the true world model (partially or fully) Eberhardt (2017); Malinsky & Danks (2018); Glymour et al. (2019), 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 used broadly used class of SCMs for modeling real-world systems, further investigation

into the effects of confounders (non-independent noise variables), as well as cyclic graphical models for time series data, would extend the reach of recourse to even broader settings. Secondly, future research will involve a thorough study of potential 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.

REFERENCES

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. 2020.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

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

Dorothy Edgington. Indicative conditionals. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2014 edition, 2014.

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

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

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

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

Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. MACE: Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

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

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

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL `http://doi.acm.org/10.1145/3236386.3241340`.

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

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. DiCE: Explaining machine learning classifiers through diverse counterfactual explanations. *arXiv preprint arXiv:1905.07697*, 2019.

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

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

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

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

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. *arXiv preprint arXiv:1909.09369*, 2019.

Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 2018.

Stefan Rüping. *Learning interpretable models*. PhD dissertation, Technical University of Dortmund, 2006.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 20–28. ACM, 2019. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287569. URL http://doi.acm.org/10.1145/3287560.3287569.

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.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. 2000.

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

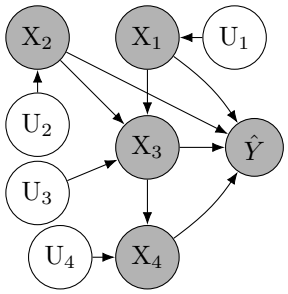
Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. ACM, 2019.

Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. 2020.

Paul Voigt and Axel Von dem Bussche. The EU general data protection regulation (GDPR). 2017.

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.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.



$$\left.\begin{aligned} \mathrm{X}_1 &\coloneqq \mathrm{U}_1 \\ \mathrm{X}_2 &\coloneqq \mathrm{U}_2 \\ \mathrm{X}_3 &\coloneqq f_3(\mathrm{X}_1, \mathrm{X}_2) + \mathrm{U}_3 \\ \mathrm{X}_4 &\coloneqq f_4(\mathrm{X}_3) + \mathrm{U}_4 \end{aligned}\right\} \mathcal{M}$$

$$\hat{\mathrm{Y}} = h_\theta\big(\{\mathrm{X}_i\}_{i=1}^4\big)$$

Figure 3: Working example; see §A.1 for details.

A Structural Counterfactuals

A.1 Working example

Consider the model in Figure 3, and assume that the SCM falls in the class of additive noise models (ANM), where $\{\mathrm{U}_i\}_{i=1}^4$ are mutually independent endogenous variables, and $\{f_i\}_{i=1}^4$ are structural (linear or nonlinear) equations.

Let $\boldsymbol{x}^{\mathrm{F}} = [x_1^{\mathrm{F}}, x_2^{\mathrm{F}}, x_3^{\mathrm{F}}]^T$ be the observed features belonging to an (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 $\mathbb{A}$. Then, we obtain a structural counterfactual, $\boldsymbol{x}^{\mathrm{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\boldsymbol{x}^{\mathrm{F}}))$, by applying the abduction-action-prediction method of counterfactual reasoning Pearl (2013) as:

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

$$\begin{aligned} u_1 &= x_1^{\mathrm{F}}, \\ u_2 &= x_2^{\mathrm{F}}, \\ u_3 &= x_3^{\mathrm{F}} - f_3(x_1^{\mathrm{F}}, x_2^{\mathrm{F}}), \\ u_4 &= x_4^{\mathrm{F}} - f_4(x_3^{\mathrm{F}}). \end{aligned} \tag{5}$$

Step 2. Action modifies the SCM according to the hypothetical interventions, $\mathrm{do}(\{\mathrm{X}_i \coloneqq a_i\}_{i\in I})$, yielding $\mathbb{F}_{\mathbb{A}}$ as:

$$\begin{aligned} \mathrm{X}_1 &\coloneqq [1 \in I] \cdot a_1 + [1 \notin I] \cdot \mathrm{U}_1, \\ \mathrm{X}_2 &\coloneqq [2 \in I] \cdot a_2 + [2 \notin I] \cdot \mathrm{U}_2, \\ \mathrm{X}_3 &\coloneqq [3 \in I] \cdot a_3 + [3 \notin I] \cdot \big(f_3(\mathrm{X}_1, \mathrm{X}_2) + \mathrm{U}_3\big), \\ \mathrm{X}_4 &\coloneqq [4 \in I] \cdot a_4 + [4 \notin I] \cdot \big(f_4(\mathrm{X}_3) + \mathrm{U}_4\big). \end{aligned} \tag{6}$$

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 $\mathbb{F}_{\mathbb{A}}$ from Step 2, as:

$$\begin{aligned} x_1^{\mathrm{SCF}} &\coloneqq [1 \in I] \cdot a_1 + [1 \notin I] \cdot \big(u_1\big), \\ x_2^{\mathrm{SCF}} &\coloneqq [2 \in I] \cdot a_2 + [2 \notin I] \cdot \big(u_2\big), \\ x_3^{\mathrm{SCF}} &\coloneqq [3 \in I] \cdot a_3 + [3 \notin I] \cdot \big(f_3(x_1^{\mathrm{SCF}}, x_2^{\mathrm{SCF}}) + u_3\big), \\ x_4^{\mathrm{SCF}} &\coloneqq [4 \in I] \cdot a_4 + [4 \notin I] \cdot \big(f_4(x_3^{\mathrm{SCF}}) + u_4\big). \end{aligned} \tag{7}$$