

# Algorithmic Recourse: from Counterfactual Explanations to Interventions

Amir-Hossein Karimi<sup>1,2</sup> Bernhard Schölkopf<sup>1</sup> Isabel Valera<sup>1</sup>

## 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. Finally, we provide the reader with an extensive discussion on how to realistically achieve recourse beyond structural interventions.

## 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

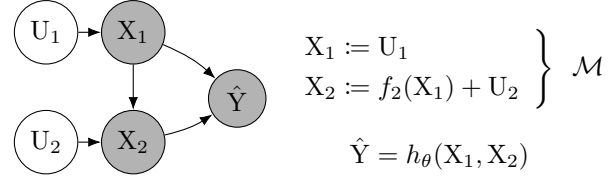


Figure 1: Illustration of an example causal data generative process governing the world, showing both the graphical model,  $\mathcal{G}$ , and the structural causal model,  $\mathcal{M}$ , (Pearl, 2000). In this example,  $X_1$  represents an individual’s annual salary,  $X_2$  is bank balance, and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

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; Murdoch et al., 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).

<sup>1</sup>Max Planck Institute for Intelligent Systems <sup>2</sup>Max Planck ETH Center for Learning Systems. Correspondence to: Amir-Hossein Karimi <amirhkarimi@gmail.com>. Under Review.

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; 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.<sup>1</sup>

**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  $h_\theta = \text{sgn}(X_1 + 5 \cdot X_2 - \$225,000)$ . 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 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 =  $h_\theta$  (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 counterfac-

tual 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  $h_\theta$ .

**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. Finally, we provide a detailed discussion on how to realistically achieve recourse beyond structural interventions, building on existing causal frameworks and posing open questions for the explainability and causality communities.

## 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),<sup>2</sup> 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,  $\text{dist}: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ . This problem has been formulated as the following optimization problem (Wachter et al., 2017):

$$\begin{aligned} \mathbf{x}_*^{\text{CFE}} \in \underset{\mathbf{x}}{\text{argmin}} \quad & \text{dist}(\mathbf{x}, \mathbf{x}^{\text{F}}) \\ \text{s.t.} \quad & h_\theta(\mathbf{x}) \neq h_\theta(\mathbf{x}^{\text{F}}) \\ & \mathbf{x} \in \mathcal{P}_{\text{plausible}} \end{aligned} \quad (1)$$

where  $\mathbf{x}^{\text{F}} \in \mathcal{X}$  is the factual instance;  $\mathbf{x}_*^{\text{CFE}} \in \mathcal{X}$  is a (perhaps not unique) nearest counterfactual instance;  $h_\theta$  is the fixed predictor; and  $\mathcal{P}$  is an optional set of *plausibility* constraints, e.g., the counterfactual instance be from a relatively

<sup>1</sup>See (Barocas et al., 2020) for additional examples.

<sup>2</sup>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.

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 (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  $h_\theta$  (e.g., random forest, multilayer perceptron), and realistic plausibility constraints,  $\mathcal{P}$ . In particular, Wachter et al. (2017) and Mothilal et al. (2019) solve (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 (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  $\mathbf{x}_*^{\text{CFE}}$  from  $\mathbf{x}^{\text{F}}$  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 (1) as:

$$\begin{aligned} \delta^* \in \underset{\delta}{\text{argmin}} \quad & \text{cost}(\delta; \mathbf{x}^{\text{F}}) \\ \text{s.t.} \quad & h_\theta(\mathbf{x}^{\text{CFE}}) \neq h_\theta(\mathbf{x}^{\text{F}}) \\ & \mathbf{x}^{\text{CFE}} = \mathbf{x}^{\text{F}} + \delta \\ & \mathbf{x}^{\text{CFE}} \in \mathcal{P}\text{plausible} \\ & \delta \in \mathcal{F}\text{feasible} \end{aligned} \quad (2)$$

where  $\text{cost}(\cdot; \mathbf{x}^{\text{F}}): \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$  is a user-specified cost that encodes preferences between feasible actions from  $\mathbf{x}^{\text{F}}$ , and  $\mathcal{F}$  and  $\mathcal{P}$  are optional sets of feasibility and plausibility constraints,<sup>3</sup> restricting the actions and the resulting counterfactual explanation, respectively. The feasibility constraints in (2), as introduced by Ustun et al. (2019), aim at restricting

the set of features that the individual may act upon.<sup>4</sup> For instance, recommendations should not ask an individual to reduce their age.

The seemingly innocent reformulation of (1) as (2) is founded on two assumptions:

**A1:** the feature-wise vector difference between factual and nearest counterfactual instances,  $\delta^* = \mathbf{x}_*^{\text{CFE}} - \mathbf{x}^{\text{F}}$ , directly translates to the minimal action set,  $\mathbb{A}^*$ , i.e., performing the actions in  $\mathbb{A}^*$  starting from  $\mathbf{x}^{\text{F}}$  will result in  $\mathbf{x}_*^{\text{CFE}}$ ; 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 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 **A2** does not hold either.

Therefore, except in trivial cases where the model designer actively inputs pair-wise independent features to  $h_\theta$ , 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  $\mathbb{F}: \mathcal{U} \rightarrow \mathcal{X}$ , 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*,  $\mathbb{A}: \Pi \rightarrow \Pi$ , which can be thought of as a transformation between SCMs (Pearl, 1994; 2000). For instance, the set of interventions can be constructed as  $\mathbb{A} = \text{do}(\{X_i := a_i\}_{i \in I})$  where  $I$  contains the indices of the subset of endogenous variables to be

<sup>3</sup>Note the difference in definition: “feasible” means possible to do, whereas “plausible” means possibly true, believable or realistic.

<sup>4</sup>The actionability/mutability of a feature is determined based on the feature semantic and value in the factual instance.

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  $X_i := a_i$ . 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  $\mathbb{F}_{\mathbb{A}} = \{\mathbb{F}_i\}_{i \notin I} \cup \{X_i := a_i\}_{i \in I}$ .

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  $\mathbf{x}^F$  becomes  $\mathbf{x}^{\text{SCF}}$ ). 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  $\mathbb{F}(\mathbb{F}^{-1}(\mathbf{x})) = \mathbf{x}$ ),  $\mathbb{X}$  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,  $\{X_i = x_i^F\}_i \subseteq \mathcal{X}$ , as  $\mathbb{F}^{-1}(\mathbf{x}^F)$ .<sup>5</sup> As a result, we can compute *any* structural counterfactuals query  $\mathbf{x}^{\text{SCF}}$ , which automatically account for inter-variable causal dependencies, for an individual  $\mathbf{x}^F$  as  $\mathbf{x}^{\text{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\mathbf{x}^F))$ , that is: “given model  $\mathcal{M}$  and having observed  $\mathbf{x}^F$ , what is the value of all endogenous variables if the set of actions  $\mathbb{A}$  is performed”.<sup>6</sup>

We can now better understand why assumption **A1** fails: whereas  $\delta^*$  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  $\delta^*$ , as in (2), leads to uncertainty regarding the post-intervention relations. In other words, the optimization problem in (2), lacks proper handling of the consequences of actions.<sup>7</sup> From the point of view of actions as a structural interventions, given  $\delta^*$ , an individual seeking recourse in a world,  $\mathcal{M}$ , may opt for one of two courses of action:

<sup>5</sup>For simplicity, we slightly abuse notation and interchangeably use sets and vectors, e.g.,  $\{X_i = x_i^F\}_i \subseteq \mathcal{X}$  and  $\mathbf{x}^F \in \mathcal{X}$ .

<sup>6</sup>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), as long as we assume that the laws governing the world,  $\mathbb{F}$ , are stationary.

<sup>7</sup>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.

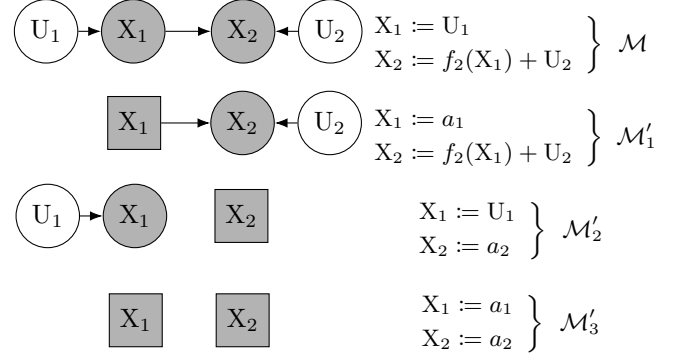


Figure 2: Given world model,  $\mathcal{M}$ , intervening on  $X_1$  and/or on  $X_2$  result in different post-manipulation models:  $\mathcal{M}'_1 = \mathcal{M}_{\mathbb{A}=\{\text{do}(X_1:=a_1)\}}$  corresponds to interventions only on  $X_1$  with consequential effects on  $X_2$ ;  $\mathcal{M}'_2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

**Option #1:** Perform interventions only on the *non-zero* elements of  $\delta^*$ , i.e.,  $\mathbb{A}^* = \text{do}(\{X_i := x_i^F + \delta_i^* \mid \delta_i^* \neq 0\}_i)$ . 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  $\text{do}(X_1 := a_1)$  affects both  $X_1$  and  $X_2$ ), there is no guarantee that  $\mathbf{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\mathbf{x}^F))$  will correspond to a counterfactual instance. In other words, due to unverified feature interactions when passing  $\mathbf{x}^{\text{SCF}}$  through  $h_\theta$ , it may be that  $h_\theta(\mathbf{x}^{\text{SCF}}) \neq h_\theta(\mathbf{x}_*^{\text{CFE}})$ , 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  $\mathbf{x}^F$ , i.e.,  $\mathbb{A}^* = \text{do}(\{X_i := x_i^F + \delta_i^*\}_i)$  severing all inter-variable edges (i.e.,  $X_i \perp\!\!\!\perp X_j \forall i \neq j$ ). This setup is depicted in  $\mathcal{M}'_3$  of Figure 2. By the independent world reduction,  $\mathbf{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\mathbf{x}^F))$  will indeed equal  $\mathbf{x}_*^{\text{CFE}}$ . However, it assumes that non-altering interventions (i.e.,  $X_i := x_i^F + 0$ ) 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  $h_\theta$ , 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.



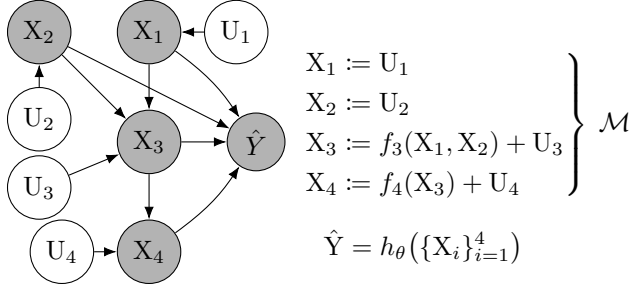


Figure 3: Working example; see §4.1 and §4.3 for details.

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 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

In the previous section, we learned that recommendations only immediately follow from counterfactual explanations in the restricted independent world setting, and otherwise have unpredictable behaviour. Here we present a complete do-over, focusing on interventions rather than explanations.

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  $h_\theta$ . Therefore, we re-formulate (2) as follows:

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

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  $\operatorname{cost}(\cdot; \mathbf{x}^F): \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R}_+$ , and  $\mathbf{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\mathbf{x}^F))$  denotes the resulting structural counterfactual explanation. We remark here that, while  $\mathbf{x}_*^{\text{SCF}}$  is a counterfactual explanation, it does not need to correspond to the nearest counterfactual explanation,  $\mathbf{x}_*^{\text{CFE}}$ , resulting from (2) (see, e.g., Example #1 of §1). Next, we show an example on how to compute  $\mathbf{x}^{\text{SCF}}$ .

### 4.1. Working example

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

Let  $\mathbf{x}^F = [x_1^F, x_2^F, x_3^F]^T$  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,  $\mathbf{x}^{\text{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\mathbf{x}^F))$ , 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,  $\{X_i = x_i^F\}_{i=1}^4$ :

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

**Step 2. Action** modifies the SCM according to the hypothetical interventions,  $\operatorname{do}(\{X_i := a_i\}_{i \in I})$ , yielding  $\mathbb{F}_{\mathbb{A}}$  as:

$$\begin{aligned}
 X_1 &:= [1 \in I] \cdot a_1 + [1 \notin I] \cdot U_1, \\
 X_2 &:= [2 \in I] \cdot a_2 + [2 \notin I] \cdot U_2, \\
 X_3 &:= [3 \in I] \cdot a_3 + [3 \notin I] \cdot (f_3(X_1, X_2) + U_3), \\
 X_4 &:= [4 \in I] \cdot a_4 + [4 \notin I] \cdot (f_4(X_3) + U_4),
 \end{aligned} \tag{5}$$

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

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

### 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 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 gen-

erally be written as:

$$x_i^{\text{SCF}} = [i \in I] \cdot (x_i^F + \delta_i) + [i \notin I] \cdot (x_i^F + f_i(\mathbf{pa}_i^{\text{SCF}}) - f_i(\mathbf{pa}_i^F)), \quad (7)$$

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 (7) 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.,  $f_i(\mathbf{pa}_i^{\text{SCF}}) - f_i(\mathbf{pa}_i^F)$ ), thus accounting for the consequences of changing other variables on this variable. Importantly, the relations in (7) can simply replace the second constraint in (3), i.e.,  $\mathbf{x}^{\text{SCF}} = \mathbb{F}_{\mathbb{A}}(\mathbb{F}^{-1}(\mathbf{x}^F))$ . We also remark that the above 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  $\mathbf{x} = \mathbb{F}(\mathbb{F}^{-1}(\mathbf{x}))$ .

**Remark #1:** We assume the time horizon is such that the effects of all interventions have played out through the entire model, before the result is again evaluated on  $h_\theta$ . E.g., the bank asks the individual to wait a year before re-applying for a loan, to allow for income-based savings to accumulate.

**Remark #2:** One criticism of our work is the requirement for knowledge of causal relations. While this is a valid concern of any approach suggesting actions to be performed in the real world, we argue that existing approaches already implicitly make causal assumptions (i.e., that of independence). Therefore, at worst case we replace an imperfect assumption about the data generative process with another. Causal discovery of the true world model is out of scope, and we refer the interested reader to the works of Eberhardt (2017); Malinsky & Danks (2018), and Glymour et al. (2019).

**Remark #3:** While solving the optimization problem in (7) is out of scope of this paper, solutions may be inspired and built on existing frameworks for generating nearest counterfactual explanations, including gradient-based, evolutionary-based, heuristics-based, or verification-based approaches as referenced in §2. For the illustrations below, we extended the open-source code of MACE (Karimi et al., 2019); we will submit a pull-request to the respective repository.

### 4.3. Experiments

We showcase our proposed method 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 consider two settings: i) Example #1 in §1, where  $\mathcal{M}$  follows Figure 1; and ii) a real-world setting based on the `german`

credit dataset (Bache & Lichman, 2013), where  $\mathcal{M}$  follows Figure 3. We computed the cost of actions as the  $\ell_1$  norm over normalized feature changes to make effort comparable across features, i.e.,  $\text{cost}(\cdot; \mathbf{x}^F) = \sum_{i \in I} |\delta_i|/R_i$ , 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  $X_1 := U_1$ ,  $X_2 := 3/10 \cdot X_1 + U_2$ , with  $U_1 \sim \$10000 \cdot \text{Poisson}(10)$  and  $U_2 \sim \$2500 \cdot \mathcal{N}(0, 1)$ ; and the predictive model  $h_\theta = \text{sgn}(X_1 + 5 \cdot X_2 - \$225000)$ . Given  $\mathbf{x}^F = [\$75000, \$25000]^T$ , solving our formulation, (3), identifies the optimal action set  $\mathbb{A}^* = \text{do}(X_1 := x_1^F + \$10000)$  which results in  $\mathbf{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\mathbf{x}^F)) = [\$85000, \$28000]^T$ , whereas solving previous formulations, (2), yields  $\delta^* = [\$0, +\$5000]^T$  resulting in  $\mathbf{x}_*^{\text{CFE}} = \mathbf{x}^F + \delta^* = [\$75000, \$30000]^T$ . Importantly, while  $\mathbf{x}_*^{\text{SCF}}$  appears to be at a further distance from  $\mathbf{x}^F$  compared to  $\mathbf{x}_*^{\text{CFE}}$ , achieving the former is less costly than the latter, specifically,  $\text{cost}(\delta^*; \mathbf{x}^F) \approx 2 \text{cost}(\mathbb{A}^*; \mathbf{x}^F)$ .

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  $\hat{Y}$  is the predicted customer risk, according to  $h_\theta$  (logistic 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  $\mathbf{x}^F = [\text{Male}, 32, \$1938, 24]^T$  identified as a risky customer, solving our formulation, (3), yields the optimal action set  $\mathbb{A}^* = \text{do}(\{X_2 := x_2^F + 1, X_3 := x_3^F - \$800\})$  which results in  $\mathbf{x}_*^{\text{SCF}} = \mathbb{F}_{\mathbb{A}^*}(\mathbb{F}^{-1}(\mathbf{x}^F)) = [\text{Male}, 33, \$1138, 22]^T$ , whereas solving (2) yields  $\delta^* = [\text{N/A}, +6, 0, 0]^T$  resulting in  $\mathbf{x}_*^{\text{CFE}} = \mathbf{x}^F + \delta^* = [\text{Male}, 38, \$1938, 24]^T$ . 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 next year for a lower ( $-\$800$ ) credit amount. More generally, in a population of 50 negatively affected test individuals, previous approaches suggest actions that are on average  $\%39 \pm \%24$  and  $\%65 \pm \%8$  more costly than our approach when considering, respectively, a logistic regression and a decision tree as the predictive model  $h_\theta$ .

## 5. Towards Realistic Interventions

In §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 car-

ried 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 realistically” (Korb et al., 2004; Eberhardt, 2007). 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

Thus far, our framework assumes that actions are performed as *structural* (a.k.a., *hard*) interventions 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 instances (e.g., 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, however, also determine the form of an intervention, e.g., specifying whether an intervention cancels out or complements existing causal effect relations.

For instance, consider Example #1 in §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 *additive* (a.k.a., *soft*) intervention (Eberhardt & Scheines, 2007). 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.

The previous example illustrates a scenario where an individual actually has the agency to perform a structural intervention, but prefers an additive intervention instead. 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 (Pearl et al., 2016). This action cannot disable pre-existing mechanisms regulating blood insulin levels and therefore, the action can only be performed additively.

Additive interventions are easily handled in our framework, where the general assignment formulation (7) is updated for variables that can be intervened upon in an additive manner:

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

The choice of whether interventions should be applied in a additive/soft or structural/hard manner will depend on the variable semantic, and should be decided prior to solving (3). Finally, in a world where only additive interventions are allowed, the work of Mahajan et al. (2019), introduced independently of ours, may assist in finding causally consistent counterfactual explanations. However, it is still unclear, as pointed out in §3 and in contrast to our formulation in (3), how to obtain recommendations from explanation.

### 5.2. On the Feasibility of Interventions

We saw in §3 that earlier works motivated the addition of *feasibility* constraints as a means to provide more actionable recommendations for the individual seeking recourse (Ustun et al., 2019). 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  $\mathbb{A} \in \mathcal{F}$  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 (Kusner et al., 2017). 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 variable (i.e.,  $x_i^F$ ); (b) the pre-intervention value of other variables (i.e.,  $\{x_j^F\}_{j \in [d] \setminus i}$ ); (c) the post-intervention value of the intervened variable (i.e.,  $x_i^{\text{SCF}}$ ); and (d) the post-intervention value of other variables (i.e.,  $\{x_j^{\text{SCF}}\}_{j \in [d] \setminus i}$ ). Such feasibility conditions can easily be encoded into  $\mathcal{F}$ ; consider the following scenarios:

- (a) an individual’s age can only increase, i.e.,  $[x_{\text{age}}^{\text{SCF}} \geq x_{\text{age}}^F]$ ;
- (b) an individual cannot apply for credit on a temporary visa, i.e.,  $[x_{\text{visa}}^F = \text{PERMANENT}] \geq [x_{\text{credit}}^{\text{SCF}} = \text{TRUE}]$ ;
- (c) an individual may undergo heart surgery (an additive intervention) only if they won’t remiss due to sustained smoking habits, i.e.,  $[x_{\text{heart}}^{\text{SCF}} \neq \text{REMISSION}]$ ; and
- (d) an individual may undergo heart surgery only *after* their blood pressure is regularized due to medicinal intervention, i.e.,  $[x_{\text{bp}}^{\text{SCF}} = 0.K.] \geq [x_{\text{heart}}^{\text{SCF}} = \text{SURGERY}]$ .

In summary, while previous works on algorithmic recourse distinguished between actionable, conditionally actionable,<sup>8</sup> and immutable variables (Ustun et al., 2019), 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.

### 5.3. On the Scope of Interventions

One final assumption has been made throughout our discussion of actions as interventions is the one-to-one mapping between an action in the real world and an intervention on an endogenous variable in the structural causal model, which in turn are also input features to the predictive model.

<sup>8</sup>Ustun et al. (2019) 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).

As exemplified in (Barocas et al., 2020), it is possible for some actions (e.g., finding a higher-paying job) to simultaneously intervene multiple variables in the model (e.g., income and length of employment). Alternatively, for Example #2 in §1, choosing a new paddy location is equivalent to intervening jointly on several input features of the predictive model. Such confounded/correlated interventions, referred to as *fat-hand/non-atomic* interventions (Eberhardt & Scheines, 2007), will be explored further in follow-up work, by modelling the world at different causally consistent levels (Rubenstein et al., 2017; Beckers & Halpern, 2019).

## 6. 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  $h_\theta$  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 (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

- Bache, K. and Lichman, M. UCI machine learning repository, 2013.
- Barocas, S., Selbst, A. D., and Raghavan, M. The hidden assumptions behind counterfactual explanations and principal reasons. 2020.
- Beckers, S. and Halpern, J. Y. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2678–2685, 2019.
- Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- Eberhardt, F. *Causation and intervention*. PhD dissertation, California Institute of Technology, 2007.
- Eberhardt, F. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.
- Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.
- Edgington, D. Indicative conditionals. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2014 edition, 2014.
- Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.
- Gunning, D. Darpa’s explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. ii–ii. ACM, 2019.
- Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.
- Karimi, A.-H., Barthe, G., Belle, B., and Valera, I. MACE: Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.
- Kodratoff, Y. The comprehensibility manifesto. *KDD Nugget Newsletter*, 94(9), 1994.
- Korb, K. B., Hope, L. R., Nicholson, A. E., and Axnick, K. Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 322–331. Springer, 2004.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.
- Lagnado, D. A., Gerstenberg, T., and Zultan, R. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.
- Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detynecki, M. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.
- Lipton, Z. C. 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>.
- Mahajan, D., Tan, C., and Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.
- Malinsky, D. and Danks, D. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.
- Mothilal, R. K., Sharma, A., and Tan, C. DiCE: Explaining machine learning classifiers through diverse counterfactual explanations. *arXiv preprint arXiv:1905.07697*, 2019.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.
- Pearl, J. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*, pp. 454–462. Elsevier, 1994.
- Pearl, J. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.
- Pearl, J. Structural counterfactuals: A brief introduction. *Cognitive Science*, 37(6):977–985, 2013.
- Pearl, J., Glymour, M., and Jewell, N. P. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. FACE: Feasible and actionable counterfactual explanations. *arXiv preprint arXiv:1909.09369*, 2019.

- Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.
- Rudin, C. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 2018.
- Rüping, S. *Learning interpretable models*. PhD dissertation, Technical University of Dortmund, 2006.
- Russell, C. 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>.
- Sharma, S., Henderson, J., and Ghosh, J. CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.
- Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. 2000.
- Starr, W. Counterfactuals. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.
- Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. ACM, 2019.
- Venkatasubramanian, S. and Alfano, M. The philosophical basis of algorithmic recourse. 2020.
- Voigt, P. and Von dem Bussche, A. The EU general data protection regulation (GDPR). 2017.
- Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2017.
- Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.