Proceedings of Machine Learning Research vol 140:1–15, 2022 1st Conference on Causal Learning and Reasoning

Causal Explanations and XAI

Sander Beckers SREKCEBREDNAS@GMAIL.COM
University of Tübingen

Editors: Bernhard Schölkopf, Caroline Uhler and Kun Zhang

Abstract

Although standard Machine Learning models are optimized for making predictions about observations, more and more they are used for making predictions about the results of actions. An important goal of Explainable Artificial Intelligence (XAI) is to compensate for this mismatch by offering explanations about the predictions of an ML-model which ensure that they are reliably *action-guiding*. As action-guiding explanations are causal explanations, the literature on this topic is starting to embrace insights from the literature on causal models. Here I take a step further down this path by formally defining the causal notions of *sufficient explanations* and *counterfactual explanations*. I show how these notions relate to (and improve upon) existing work, and motivate their adequacy by illustrating how different explanations are action-guiding under different circumstances. Moreover, this work is the first to offer a formal definition of *actual causation* that is founded entirely in action-guiding explanations. Although the definitions are motivated by a focus on XAI, the analysis of causal explanation and actual causation applies in general. I also touch upon the significance of this work for fairness in AI by showing how actual causation can be used to improve the idea of path-specific counterfactual fairness.

Keywords: Explanation; Counterfactual; Actual Causation; Fairness

1. Introduction

Explainable Artificial Intelligence (XAI) is concerned with offering explanations of predictions made by Machine Learning models. Such predictions can serve many different purposes, each of which calls for a different kind of explanation. Part of the XAI literature is beginning to embrace the insight that Pearl's causal hierarchy is an invaluable tool to specify the requirements that different explanations need to fulfil for achieving those purposes (Pearl and Mackenzie, 2018). This hierarchy consists of a ladder that has three rungs: observations, interventions, and counterfactuals. Standard ML-models are optimized solely for making predictions about the first rung of the ladder, observations, and their widespread application is due to the fact that they are very successful in doing so. Yet there is also a growing interest in using ML-models for action-guidance, and that involves predictions about the second rung of the ladder, interventions (Lipton, 2018; Molnar et al., 2021; Schölkopf et al., 2021). Moreover, if the action occurs against the background of an already existing observation, then such predictions take place on the third rung of the ladder, counterfactuals.

Corresponding to these different purposes we have different notions of explanation that are appropriate for each. Simply put, purely associational measures – such as feature attribution (Ribeiro, 2016; Lundberg and Lee, 2017) – are appropriate when one requires explanations of the behavior of the ML-model itself: why did the model produce a particular output $Y = y$ for a particular input $\boldsymbol{I} = \boldsymbol{i}$? Such *model explanations* serve to better understand the model, be it to gain trust in the model or to perform model audit (Lipton, 2018; König et al., 2021).

Actions, on the other hand, take place in the real world, and thus acting on one feature might affect the value of another. Therefore if one desires explanations about the outputs produced as

© 2022 S. Beckers.

the result of performing an action, then one is looking for *action-guiding explanations*. To find those one needs to go beyond the ML-model and consider the target system that it is a model of. Since action-guiding explanations are causal explanations, this requires knowledge of an additional model, namely a causal model of the target system. Fully specified causal models are notoriously hard to come by, but there are promising approaches that focus on combining partial causal models with ML-models (Mahajan et al., 2019; Karimi et al., 2020; Schölkopf et al., 2021). This paper focuses on a conceptual analysis of the relevant notions of causal explanations – as opposed to practical methods for obtaining such explanations – and therefore we assume perfect knowledge of a deterministic causal model of the target system. The results of weakening this assumption are to be investigated in future work.

Action-guiding explanations can occur in two basic forms. If one is interested in explaining under which conditions an action guarantees a particular output, then one needs **Sufficient Explanations**. If one is interested in finding an action that changes an observed input to produce a change in an already observed output, then **Counterfactual Explanations** are called for. Beyond these basic forms, there exist many circumstances in which one is interested more generally in finding out which past actions actually explain an already observed output. Such explanations ask for the **Actual Causes** of an output, and it will turn out that they sit in between sufficient and counterfactual explanations in roughly the following way: an actual cause is a part of a good sufficient explanation for which there exist counterfactual values that would not have made the explanation better.

The next two sections provide the background context for our analysis. Sections 4 and 5 introduce various definitions of sufficient and counterfactual explanations by relating existing accounts of these notions to their causal counterparts. Section 6 offers a definition of actual causation, and concludes by presenting a novel definition of fairness based on actual causation.

2. Structural Equations Modeling

For the purposes of notational consistency across related work, this section reviews the definition of causal models along the lines of Halpern (2016) with little change.

Definition 1 *A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of* exogenous *variables, $\mathcal{V}$ is a set of* endogenous *variables, and $\mathcal{R}$ a function that associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y* ranges*). If $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\mathcal{R}(\boldsymbol{X})$ denotes the crossproduct $\mathcal{R}(X_1) \times \cdots \times \mathcal{R}(X_n)$.*

Exogenous variables represent factors whose causal origins are outside the scope of the causal model, such as background conditions and noise. The values of the endogenous variables, on the other hand, are causally determined by other variables within the model.

Definition 2 *A* causal model *M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a signature and $\mathcal{F}$ defines a function that associates with each endogenous variable X a* structural equation *F_X giving the value of X in terms of the values of other endogenous and exogenous variables. Formally, the equation F_X maps $\mathcal{R}(\mathcal{U} \cup \mathcal{V} - \{X\})$ to $\mathcal{R}(X)$, so F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.*

We call a setting $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ of values of exogenous variables a *context*. The value of X may depend on the values of only a few other variables. *X depends on Y in context $\boldsymbol{u}$* if there is some setting of the endogenous variables other than X and Y such that if the exogenous variables have

value $\boldsymbol{u}$, then varying the value of Y in that context results in a variation in the value of X; that is, there is a setting $\boldsymbol{z}$ of the endogenous variables other than X and Y and values y and y' of Y such that $F_X(y, \boldsymbol{z}, \boldsymbol{u}) \neq F_X(y', \boldsymbol{z}, \boldsymbol{u})$. We then say that Y is a *parent* of X and X is a *child* of Y.

In this paper we restrict attention to *strongly recursive* models, that is, models where there is a partial order $\preceq$ on variables such that if X depends on Y, then $Y \prec X$. In a strongly recursive model, given a context $\boldsymbol{u}$, the values of all the remaining variables are determined (we can just solve for the value of the variables in the order given by $\preceq$).

An *intervention* has the form $\boldsymbol{X} \leftarrow \boldsymbol{x}$, where $\boldsymbol{X}$ is a set of endogenous variables. Intuitively, this means that the values of the variables in $\boldsymbol{X}$ are set to the values $\boldsymbol{x}$. The structural equations define what happens in the presence of interventions. Setting the value of some variables $\boldsymbol{X}$ to $\boldsymbol{x}$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{\boldsymbol{X} \leftarrow \boldsymbol{x}}$, which is identical to M, except that $\mathcal{F}$ is replaced by $\mathcal{F}^{\boldsymbol{X} \leftarrow \boldsymbol{x}}$: for each variable $Y \notin \boldsymbol{X}$, $F_Y^{\boldsymbol{X} \leftarrow \boldsymbol{x}} = F_Y$ (i.e., the equation for Y is unchanged), while for each X' in $\boldsymbol{X}$, the equation $F_{X'}$ for X' is replaced by $X' = x'$ (where x' is the value in $\boldsymbol{x}$ corresponding to X').

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, an *atomic formula* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where

- ϕ is a Boolean combination of atomic formulas,
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$ for each $1 \leq i \leq k$.

Such a formula is abbreviated as $[\boldsymbol{Y} \leftarrow \boldsymbol{y}]\phi$. The special case where $k = 0$ is abbreviated as ϕ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ says that ϕ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a *causal setting*, which is a causal model given a context. As usual, we write $(M, \boldsymbol{u}) \models \psi$ if the causal formula ψ is true in the causal setting $(M, \boldsymbol{u})$. (If ψ is true for all contexts $\boldsymbol{u}$ we write $M \models \psi$.) The $\models$ relation is defined inductively. $(M, \boldsymbol{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in M in context $\boldsymbol{u}$ (i.e., the unique vector of values that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\boldsymbol{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \boldsymbol{u}) \models [\boldsymbol{Y} \leftarrow \boldsymbol{y}]\phi$ if $(M_{\boldsymbol{Y} \leftarrow \boldsymbol{y}}, \boldsymbol{u}) \models \phi$.

3. Background

Assume that we have an ML-model which implements the overall function $h : \mathcal{R}(\boldsymbol{I}) \rightarrow \mathcal{R}(Y)$, where $\boldsymbol{I}$ consists of all the input variables and Y is the single output variable. As mentioned, we assume knowledge of a causal model M of the target domain (i.e., that part of the world which h is a model of). A first requirement is that such a causal model is consistent with the ML-model, meaning that they at least agree on how to classify all observations. A second requirement that is helpful given the conceptual nature of our analysis is that the endogenous variables $\boldsymbol{I}$ suffice to obtain deterministic causal knowledge of the target domain. Technically this means that there are only two kinds of endogenous variables: those which are determined directly by the unobserved exogenous variables, and those which are determined entirely by members of $\boldsymbol{I}$. In future work both requirements can be loosened by adding probabilities to each. For the first, we can demand that both models are *likely* to agree on observations. For the second, we can add new exogenous

variables and a probability distribution over them, which give us probabilistic causal models. These allow for probabilistic generalizations of all the definitions here presented.

Definition 3 *A causal model* $M = ((\mathcal{V}, \mathcal{U}, \mathcal{R}), \mathcal{F})$ agrees with $h : \mathcal{R}(\boldsymbol{I}) \to \mathcal{R}(Y)$ *if* $\mathcal{V} = \boldsymbol{I} \cup \{Y\}$, *each* $V_i \in \mathcal{V}$ *either has a unique single exogenous parent* $U_i \in \mathcal{U}$ *or no exogenous parents, and for all* $\boldsymbol{i} \in \mathcal{R}(\boldsymbol{I})$, $y \in \mathcal{R}(Y)$*:* $M \models \boldsymbol{I} = \boldsymbol{i} \to Y = y$ *iff* $h(\boldsymbol{i}) = y$.

As knowledge of the causal model is hard to come by many approaches to action-guiding explanations ignore it entirely and assume that the input variables are causally independent (Sharma et al., 2020; Ustun et al., 2019; Poyiadzi et al., 2020; Kommiya Mothilal et al., 2021).

Definition 4 A causal model M that agrees with h satisfies **Independence** *if for each* $I_a, I_b \in \boldsymbol{I}$, I_a *is not a parent of* I_b.

Obviously **Independence** rarely holds, and thus any account of action-guiding explanation that depends on it is limited. Although this limitation is recognized (Mahajan et al., 2019; Karimi et al., 2021b; Molnar et al., 2021), this paper offers several results that formally characterize how extreme this limitation is. Simply put, under **Independence** a variety of very different causal notions become indistinguishable from each other. This holds for notions of sufficient explanation, counterfactual explanations, and various notions of actual causation.

Because of this limitation, work on action-guiding explanations in XAI has failed to take up the most relevant lesson that the literature on causation has to offer, namely that to give a causal explanation of an outcome is to give actual causes of that outcome (Woodward, 2003; Halpern and Pearl, 2005). An additional reason for this oversight is that the precise relation between explanation and prediction has been the subject of much debate in the history of philosophy of science.

Prediction in its most natural application is a forward-looking notion, meaning one predicts an event before it takes place. Explanation on the other hand is a backward-looking notion, meaning that one explains an event after it has happened. Yet as many papers on XAI clearly illustrate, explanations about past events are often required precisely to inform predictions about future events. Therefore a suitable notion of causal explanation, and thus also of actual causation, needs to specify how it relates to predictions. Given the tumultuous history that this relation has in the philosophy literature, it has been duly neglected in the philosophical work on causation, thereby obscuring the importance of causation for the practical goals that XAI is concerned with. (Hitchcock (2017) forms a notable exception!)[1] This paper corrects this by developing an account of causal explanation that shows both how it is connected to actual causation and how it can lead to action-guiding predictions.

Simply put, the goal of this paper is to upgrade the formalization of Woodward's influential philosophical account of causal explanation described below with the most recent insights from the causation literature, whilst also keeping track of the action-guiding demands that are prevalent in XAI (Woodward, 2003, p. 11).

> Put differently, my idea is that one ought to be able to associate with any successful explanation a hypothetical or counterfactual experiment that shows us that and how manipulation of the factors mentioned in the explanation (the *explanans*, as philosophers call it) would be a way of manipulating or altering the phenomenon explained (the *explanandum*). Put in still another way, an explanation ought to be such that it

1. In fact, his more informal exploration of the connection between actual causation and action-guiding prediction proceeds along somewhat similar lines as ours. One can view this paper as picking up where his left off.

> can be used to answer what I call a *what-if-things-had-been-different question*: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways.

A crucial novel element of my account is the addition that a successful explanation must also be explicit about those factors that *may not be manipulated* for the explanation to hold, i.e., it must state which variables are to be safeguarded from interventions. Importantly, this is distinct from stating which variables must be held fixed at their actual values, for to hold variables fixed in fact means to intervene on them.

The following example (modified from (Karimi et al., 2021b)) is helpful for illustrating the various purposes that explanations can serve.

Example 1 *Consider a system for loan applications that is captured by a causal model such that* $Y = (X_1 + 5 \cdot X_2 - 225,000) > 0$*, where* Y *is a binary variable representing whether the loan is granted,* X_1 *is the applicant's income, and* X_2 *is the applicant's savings. Further, assume that the applicant's savings are determined by their initial deposit* X_3 *and their income in the following manner:* $X_2 = 3/10 \cdot X_1 + X_3$*. It is also the case that people with high savings take out a safety deposit box (*X_4*) at the bank:* $X_4 = X_2 > 1,000,000$*.*

Standard predictions in ML take place on the first rung of Pearl's causal ladder, namely that of observations. For example, an ML-model might pick up on the fact that all *observed* loan applicants who have a safety deposit box ($X_4 = 1$) also get a loan ($Y = 1$), and thus could predict that an applicant who has a safety deposit box will get a loan, i.e., it might learn a function so that $h(x_1, x_2, x_3, 1) = 1$ for all values x_1, x_2, x_3. In terms of the causal model, it is indeed the case that $M \models X_4 = 1 \rightarrow Y = 1$. Here $X_4 = 1$ is what Ribeiro et al. (2018) refer to as an *anchor*, which they interpret as a sufficient explanation of the outcome. Yet clearly such observational explanations are not a good guide towards action, for it would be absurd to recommend to someone to take out a safety deposit box so that their loan application is approved. This point can be brought out by making use of the second rung of the ladder, namely that of interventions: an applicant who takes out a safety deposit box as the result of being advised to do so will not necessarily get a loan: $M \not\models [X_4 \leftarrow 1]Y = 1$. In the same manner, interventions can be used to offer advice that *is* a good guide towards action, for example by telling an applicant that if they manage to obtain savings of $45,001$ then they are guaranteed to get a loan: $M \models [X_2 \leftarrow 45,001]Y = 1$.

4. Sufficient Explanations

Sufficient Explanations generalize the previous point by offering settings $\boldsymbol{X} = \boldsymbol{x}$ that causally suffice for $Y = y$. Existing approaches come in two flavors. Some approaches (Galhotra et al., 2021; Watson et al., 2021) define sufficient explanations using Pearl's notion of "probability of sufficiency", the deterministic version of which can be informally stated as: if we set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}$ and do not intervene on any of the other variables, then $\boldsymbol{Y}$ takes on the values $\boldsymbol{y}$.[2] Elsewhere I have called this interpretation of sufficiency *weak sufficiency* (Beckers, 2021).

Definition 5 $\boldsymbol{X} = \boldsymbol{x}$ is weakly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ in M *if for all* $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ *we have that* $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}]\boldsymbol{Y} = \boldsymbol{y}$.

2. This condition also appears in Halpern's notion of sufficient cause (Halpern, 2016).

Other recent approaches take inspiration from logic and define sufficient explanations of an output as its *prime implicants*, giving so-called PI-explanations (Shih et al., 2018; Darwiche and Hirth, 2020; Koopman and Renooij, 2021). Informally: if we set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}$, then $\boldsymbol{Y}$ takes on the values $\boldsymbol{y}$, regardless of the values of all other variables. In the context of causal models I have coined this *direct sufficiency* (Beckers, 2021).

Definition 6 $\boldsymbol{X} = \boldsymbol{x}$ is directly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ in M *if for all* $\boldsymbol{c} \in \mathcal{R}(\boldsymbol{C})$*, where* $\boldsymbol{C} = \mathcal{V} - (\boldsymbol{X} \cup \boldsymbol{Y})$*, and all* $\boldsymbol{u} \in \mathcal{R}(\mathcal{U})$ *we have that* $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}, \boldsymbol{C} \leftarrow \boldsymbol{c}]\boldsymbol{Y} = \boldsymbol{y}$.

Obviously the second form of sufficient explanations is stronger. Coming back to our example, $X_2 = 45,001$ is a sufficient explanation for $Y = 1$ according to both readings. But $(X_1 = 50,000, X_3 = 25,000)$ is a sufficient explanation of $Y = 1$ only on the first reading, for although it is weakly sufficient for $Y = 1$ it is not directly sufficient. Which one of these notions is correct here? Imagine that a prospective applicant with these values is told that their income and initial deposit suffice for getting a loan. As a result the applicant concludes that there is no need to have such high savings and decides to spend $20,000$, so that their loan application is denied. The applicant would be quite right to be upset about this!

Such misunderstandings cannot occur when using direct sufficiency, as that gives us settings whose explanatory value is immune to the influence of interventions on other variables. Instead of concluding from this that one should always rely on direct sufficiency, I propose a generalization of sufficient explanations that adds an element to inform us explicitly as to which variables are assumed to be safeguarded from interventions. Concretely, in addition to specifying which variables need to be set to particular values, a sufficient explanation should also specify a set of variables $\boldsymbol{N}$ that are *not* to be manipulated for the explanation to be action-guiding. Informally, if we set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}$ and the variables in $\boldsymbol{N}$ are safeguarded from interventions, then Y takes on the value y, regardless of the values of all remaining variables.

I call the relevant notion of causal sufficiency at work *strong sufficiency* (Beckers, 2021), which can be formally defined as follows:

Definition 7 $\boldsymbol{X} = \boldsymbol{x}$ is strongly sufficient for $\boldsymbol{Y} = \boldsymbol{y}$ in M along $\boldsymbol{N}$ *if* $\boldsymbol{Y} \subseteq \boldsymbol{N}$ *and* $\boldsymbol{X} = \boldsymbol{x}$ *is directly sufficient for* $\boldsymbol{N} = \boldsymbol{n}$ *for some values* $\boldsymbol{n} \supseteq \boldsymbol{y}$.

Note that in this definition $\boldsymbol{N}$ cannot just be any set, but rather we require that it is itself entirely determined by $\boldsymbol{X} = \boldsymbol{x}$. This is because the variables in $\boldsymbol{N}$ can be thought of as a *network* that transmits the causal influence of $\boldsymbol{X}$ to $\boldsymbol{Y}$, and the idea of safeguarding this network is that it can continue fulfilling this role even when intervening on $\boldsymbol{X}$. (I refer the reader to (Beckers, 2021) for an elaborate discussion of this definition as well as an equivalent alternative formulation.)

The following straightforward result shows the relative strengths of the above three notions of sufficiency.

Proposition 8 *If* $\boldsymbol{X} = \boldsymbol{x}$ *is directly sufficient for* $\boldsymbol{Y} = \boldsymbol{y}$ *then* $\boldsymbol{X} = \boldsymbol{x}$ *is strongly sufficient for* $\boldsymbol{Y} = \boldsymbol{y}$ *along some* $\boldsymbol{N}$*, and if* $\boldsymbol{X} = \boldsymbol{x}$ *is strongly sufficient for* $\boldsymbol{Y} = \boldsymbol{y}$ *along some* $\boldsymbol{N}$ *then* $\boldsymbol{X} = \boldsymbol{x}$ *is weakly sufficient for* $\boldsymbol{Y} = \boldsymbol{y}$.

Using strong sufficiency we can define a sufficient explanation so that it specifies all the required elements for it to be action-guiding.

Definition 9 *A pair* $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{N})$ is a **sufficient explanation** of $Y = y$ *if* $\boldsymbol{X} = \boldsymbol{x}$ *is strongly sufficient for* $Y = y$ *along* $\boldsymbol{N}$. *Any subset* $\boldsymbol{X}' = \boldsymbol{x}'$ *appearing in* $\boldsymbol{X} = \boldsymbol{x}$ *is called* a part of the explanation.

If $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x}$, *we say that* $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{N})$ is an **actual sufficient explanation** of $Y = y$ in $(M, \boldsymbol{u})$.

If $\boldsymbol{N} = \boldsymbol{Y}$ *we speak of a* direct sufficient explanation.

Of course we do not want to add redundant parts to a sufficient set, as a good explanation should be as concise as possible. Therefore a good sufficient explanation ought to be minimal with respect both to $\boldsymbol{X}$ and to $\boldsymbol{N}$.

Definition 10 *A sufficient explanation* $(\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{N_1})$ dominates an explanation $(\boldsymbol{X_2} = \boldsymbol{x_2}, \boldsymbol{N_2})$ *if both are explanations of the same* $Y = y$, $\boldsymbol{X_1} \subseteq \boldsymbol{X_2}$, *and* $\boldsymbol{N_1} \subseteq \boldsymbol{N_2}$.

$(\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{N_1})$ strictly dominates $(\boldsymbol{X_2} = \boldsymbol{x_2}, \boldsymbol{N_2})$ *if* $(\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{N_1})$ *dominates* $(\boldsymbol{X_2} = \boldsymbol{x_2}, \boldsymbol{N_2})$ *and not vice versa.*

This allows us to define what makes for a good sufficient explanation of an observed output.

Definition 11 *An actual sufficient explanation* $(\boldsymbol{X_1} = \boldsymbol{x_1}, \boldsymbol{N_1})$ *of* $Y = y$ *in* $(M, \boldsymbol{u})$ *is* **good** *if it is not dominated by any other actual sufficient explanation in* $(M, \boldsymbol{u})$.

To conclude the analysis of sufficient explanations, I offer a first result that shows the extreme limitation **Independence** poses. We have seen three very different notions of sufficiency, and yet under **Independence** they all collapse into one.

Theorem 12 *If a causal model* M *that agrees with* h *satisfies* **Independence** *then the following statements are all equivalent:*

- $\boldsymbol{X} = \boldsymbol{x}$ *is weakly sufficient for* $Y = y$ *in* M.
- $\boldsymbol{X} = \boldsymbol{x}$ *is strongly sufficient for* $Y = y$ *in* M.
- $\boldsymbol{X} = \boldsymbol{x}$ *is directly sufficient for* $Y = y$ *in* M.

(Proofs of all Theorems are to be found in the Supplementary Material.)

5. Counterfactual Explanations

Counterfactual Explanations inform us which variables would have had to be different (and different in what way) for the outcome to be different. Although the XAI literature usually only considers one interpretation, it is important to distinguish between three different interpretations. The first interpretation is the standard one as it was introduced into the XAI literature (Wachter et al., 2017), and is in fact appropriate only when one is looking either for model explanations or for a very restricted type of action-guiding explanations. An observation $\boldsymbol{I} = \boldsymbol{i}$ and output $Y = y$ is explained by offering an observation that is identical to the first one except for some variables $\boldsymbol{X}$ that take on values $\boldsymbol{x}'$ rather than $\boldsymbol{x}$. The appropriate causal reading of this interpretation goes: if we had set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}'$ rather than $\boldsymbol{x}$ and held fixed all other variables at their actual values, then Y would have been y' rather than y. This is a deterministic version of what is usually called the direct

effect (Pearl, 2001), which is action-guiding only if one can in fact hold fixed all other variables. Here I will call this relation *direct counterfactual dependence*.

The second interpretation is the usual one found in the causation literature, which takes into account that changing $\boldsymbol{X}$ might change the values of other variables as well: if we had set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}'$ rather than $\boldsymbol{x}$ and had not intervened on any of the other variables, then Y would have been y' rather than y. This relation is standardly referred to as *counterfactual dependence* (Pearl, 2009).

The difference between these first two interpretations is gaining traction in recent work on algorithmic recourse (which is the term used for action-guiding given an already existing observation) (Karimi et al., 2021a). In fact, a slightly simpler version of Example 1 was initially used to illustrate this point precisely (Karimi et al., 2021b). Imagine an unsuccessful applicant with an income of $75,000$ and $25,000$ in savings. On the first interpretation, the "cheapest" counterfactual explanation would be that if their income had been $100,000$ rather than $75,000$ and everything else would have remain fixed, then they would have gotten the loan. The second interpretation takes into account the important fact that increases in income also result in increases in savings, and thus it would offer a cheaper counterfactual explanation: if their income had been $85,000$ rather than $75,000$ then they would have gotten the loan.

Does this mean that the second interpretation is the best one here? As before, an applicant hearing this explanation could be led to believe that they can ignore their savings, and thus spend some of it, which would of course invalidate the explanation given. Similar to the analysis of sufficient explanations, rather than choosing either one I contend instead that we should extend the explanation with information that ensures these misunderstandings cannot arise. This happens in two steps.

As a first step I introduce a third interpretation of counterfactual explanations, which is the explanatory counterpart to the most recent variant of Halpern & Pearl's influential definition of actual causation (the so-called modified definition) (Halpern, 2016). It generalizes the former two interpretations by including a "witness" to the explanation, meaning a set of variables $\boldsymbol{W}$ so that if we had set the variables in $\boldsymbol{X}$ to $\boldsymbol{x}'$ rather than $\boldsymbol{x}$, held fixed the variables in $\boldsymbol{W}$ at their actual values, and had not intervened on any other variables, then Y would have been y' rather than y.[3]

The first interpretation above implicitly assumes that the witness contains all other variables, whereas the second interpretation assumes that the witness is empty. By spelling out the witness explicitly, the third interpretation includes the important information that the explanation remains valid only when holding fixed certain variables at their actual values. The first counterfactual explanation above, for example, depends for its validity on holding fixed X_2 (and is indifferent to the values of X_3 and X_4).

Formally the causal notion that this third interpretation relies on is the latest definition of actual causation by Halpern & Pearl, which is best understood as a generalization of counterfactual dependence in the following way. (Throughout the rest of this paper, values $\boldsymbol{x}'$ are assumed to differ from $\boldsymbol{x}$ for *each* variable in $\boldsymbol{X}$.)

Definition 13 $Y = y$ counterfactually depends on $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ relative to $\boldsymbol{W} = \boldsymbol{w}$ in $(M, \boldsymbol{u})$ *if* $\boldsymbol{X}$ *is a minimal set such that* $(M, \boldsymbol{u}) \models \boldsymbol{X} = \boldsymbol{x} \wedge \boldsymbol{W} = \boldsymbol{w} \wedge Y = y$ *and* $(M, \boldsymbol{u}) \models [\boldsymbol{X} \leftarrow \boldsymbol{x}', \boldsymbol{W} \leftarrow \boldsymbol{w}]Y \neq y$.

3. Note that in many contexts interventions on certain variables are feasible but come at a "cost", and thus one could quantify the "price" of explanations by building on the work of Karimi et al. (2021b).

If $\boldsymbol{W} = \mathcal{V} \setminus (\boldsymbol{X} \cup \{Y\})$*, we speak of* direct counterfactual dependence. *If* $\boldsymbol{W} = \emptyset$*, we speak of* standard counterfactual dependence. *(Note that in both cases we do not need to specify* $\boldsymbol{W}$*.)*

The second step is to apply the insight we gained when analysing sufficient explanations, namely that an explanation should also mention the variables that are assumed to be safeguarded from interventions. This is precisely the information we need to avoid the above misunderstanding: it should be spelled out that the explanation assumes the applicant's savings will continue to be determined as they were. Therefore I define counterfactual explanations as counterfactual analogs of sufficient explanations.

Definition 14 [4] *Given a causal setting* $(M, \boldsymbol{u})$*, if* $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ *is an actual sufficient explanation of* $Y = y$ *and* $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ *is a sufficient explanation of* $Y = y'$ *with* $y' \neq y$ *then we say that* $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is a **counterfactual explanation** of $Y = y$ *relative to* $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$*. We also write this as* $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$.

Just as with sufficient explanations, we want an explanation to be as concise as possible.

Definition 15 *Given a causal setting* $(M, \boldsymbol{u})$*, we say that a counterfactual explanation* $(\boldsymbol{X_1} = (\boldsymbol{x_1}, \boldsymbol{x_1}'), \boldsymbol{W_1} = \boldsymbol{w_1}, \boldsymbol{N_1})$ dominates an explanation $(\boldsymbol{X_2} = (\boldsymbol{x_2}, \boldsymbol{x_2}'), \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{N_2})$ *if both are explanations of the same* $Y = y$*,* $\boldsymbol{X_1} \subseteq \boldsymbol{X_2}$*,* $\boldsymbol{W_1} \subseteq \boldsymbol{W_2}$*, and* $\boldsymbol{N_1} \subseteq \boldsymbol{N_2}$*.*

As before, strict domination *means that there is domination only in one direction. A counterfactual explanation is* **good** *if it is not dominated by any other counterfactual explanation.*

The following result shows that the previous two steps are indeed steps in the same direction:

Theorem 16 *Given a causal setting* $(M, \boldsymbol{u})$*, the following two statements are equivalent:*

- *there exist* $\boldsymbol{W_1}$ *and* $\boldsymbol{w_1} \in \mathcal{R}(\boldsymbol{W_1})$ *so that* $Y = y$ *counterfactually depends on* $\boldsymbol{X} = \boldsymbol{x}$ *rather than* $\boldsymbol{X} = \boldsymbol{x}'$ *relative to* $\boldsymbol{W_1} = \boldsymbol{w_1}$ *in* $(M, \boldsymbol{u})$*.*
- *there exist* $\boldsymbol{W_2}$*,* $\boldsymbol{w_2} \in \mathcal{R}(\boldsymbol{W_2})$*, and* $\boldsymbol{N}$ *so that* $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{N})$ *is a good counterfactual explanation of* $Y = y$*.*

Coming back to our example, we can offer the cheaper of the two explanations and yet avoid any misunderstanding by saying that $X_1 = 75,000$ rather than $X_1 = 85,000$ is a good counterfactual explanation of $Y = 0$ relative to $(\{X_3 = 2,500\}, \{X_2\})$.

6. Actual Causation and Explanation

Actual Causes are those events which explain an output in the most general sense of explanation. There is a lively literature on how to define actual causation using causal models, but for the reasons outlined above, this literature has yet to find its way into the XAI literature. I here aim to set this straight by explicitly connecting actual causation to the notions of explanation that we have come across. A concise counterfactual explanation is good when you can get it, but often you cannot get it, and thus we need a weaker notion of explanation that is more generally applicable.[5] (The causal

4. Note that, unlike sufficient explanations, counterfactual explanations already come with a causal setting $(M, \boldsymbol{u})$ and there is no need to define an *actual* version of counterfactual explanations.
5. Obviously you can always find *some* counterfactual explanation of an output, for changing all of the variables will always allow you to get a different output. But an explanation that involves too many variables is of little use.

counterpart of this message is what initiated the formal causation literature some fifty years ago (Lewis, 1973).)

It is clear from the definitions that sufficient explanations are weaker than counterfactual explanations. But sufficient explanations ignore the counterfactual aspect entirely, which means they are of little value for action-guidance in the presence of an already existing observation. Therefore I define actual causes as parts of explanations that sit in between counterfactual and sufficient explanations: they are parts of good sufficient explanations such that there exist counterfactual values which would not have made the explanation better. This is weaker than demanding that the counterfactual values are part of a sufficient explanation of a *different* output, as we do for counterfactual explanations. Informally, $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an *actual cause* of $Y = y$ if $\boldsymbol{X} = \boldsymbol{x}$ is part of a good sufficient explanation for $Y = y$ that could have not been made better by setting $\boldsymbol{X}$ to $\boldsymbol{x}'$.

First I make precise how changing the values of some variables can turn a sufficient explanation into a better one.

Definition 17 *If $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y$, we say that $\boldsymbol{X} = \boldsymbol{x}'$ can replace $\boldsymbol{X} = \boldsymbol{x}$ if there exists a dominating explanation that includes $(\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{W} = \boldsymbol{w})$.*

The following result ensures that the focus on $\boldsymbol{W}$ (instead of its subsets) in Definition 17 is without loss of generality.

Proposition 18 *If $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is a sufficient explanation of $Y = y$ and there exists a dominating explanation $((\boldsymbol{X} = \boldsymbol{x}', \boldsymbol{A} = \boldsymbol{a}), \boldsymbol{B})$ for some values $\boldsymbol{x}'$ and $\boldsymbol{a} \subseteq \boldsymbol{w}$, then $\boldsymbol{X} = \boldsymbol{x}'$ can replace $\boldsymbol{X} = \boldsymbol{x}$.*

This allows us to formally define actual causation.

Definition 19 *$\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an* **actual cause** *of $Y = y$ in $(M, \boldsymbol{u})$ if it is part of a good sufficient explanation of $Y = y$ in which $\boldsymbol{X} = \boldsymbol{x}$ cannot be replaced by $\boldsymbol{X} = \boldsymbol{x}'$.*

If $((\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}), \boldsymbol{N})$ is the relevant good explanation, we say that $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an actual cause of $Y = y$ relative to $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$.

Except for a minor technical difference regarding the implementation of minimality, Definition 19 is equivalent to the definition of causation I developed in previous work (Beckers, 2021). Although I have here arrived at the same notion, I did so along a very different path, for in the previous work I did not draw any connection to explanations nor to action-guidance, but instead argued for the definition by contrasting it to other proposals for defining causation (including Definition 13).

Contrary to counterfactual explanations, actual causes do not guide you towards actions that, under the same conditions, would ensure the output to be different. But they do guide you towards actions that would not ensure the actual output under the same conditions as the actual action. For example, imagine a (very fortunate) applicant for whom $X_1 = 250,000$, $X_3 = 50,000$, and thus $X_2 = 125,000$. Obviously their application is successful, and their income by itself offers a good explanation of this fact: $X_1 = 250,000$ is a good direct sufficient explanation of $Y = 1$. Also, their income does not counterfactually explain the output, for the application would have been approved regardless of income. Yet we do have that $X_1 = 250,000$ rather than $X_1 = 200,000$ is an actual cause of the output, because we cannot replace $X_1 = 250,000$ by $X_1 = 200,000$ in our sufficient explanation (i.e., $X_1 = 200,000$ is not a sufficient explanation of $Y = 1$). This is helpful for example if the applicant is considering changing jobs and would like to know whether they can use up their savings and still get their application approved.

One might wonder whether we need a separate definition of actual causation, as opposed to simply considering all parts of good sufficient explanations to be causes. The distinction between these two options lies in the existence of alternative actions that would make the actual explanation worse, and it are precisely those actions which make actual causes good guides towards action. Consider for example a situation in which a short-circuit starts a fire ($F = 1$) in an office building. The flames set off the sprinklers ($S = 1$), and those put out the flames, preventing the building from burning down ($B = 0$) – where all variables are binary. Matching equations for this story are: $B = F \wedge \neg S$, $S = F$. In this scenario, the fire offers a good sufficient explanation of the building not burning down relative to the sprinklers functioning as they should. But obviously it would be unwise to conclude from this that we should start fires in order to prevent buildings from burning down! This point can be brought out by noting that the fire is *not* an actual cause of the building not burning down, because *there not being a fire* would have offered a better sufficient explanation of this outcome, as it does not rely on the sprinklers functioning properly. (Concretely: $F = 0$ can replace $F = 1$ in the good sufficient explanation $(F = 1, S)$ of $B = 0$.) Since sprinklers can malfunction, or can be made to malfunction by a malicious actor, the best action is to not set fires in a building, in accordance with actual causation.

The following result specifies the claim that actual causes sit in between counterfactual and sufficient explanations: counterfactual explanations always contain actual causes (and obviously not vice versa).

Theorem 20 *If $\boldsymbol{X_1} = \boldsymbol{x_1}$ rather than $\boldsymbol{X_1} = \boldsymbol{x_1}'$ is a counterfactual explanation of $Y = y$ in $(M, \boldsymbol{u})$ (relative to some $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$) then for some $\boldsymbol{X_2} \subseteq \boldsymbol{X_1}$, $\boldsymbol{X_2} = \boldsymbol{x_2}$ rather than $\boldsymbol{X_2} = \boldsymbol{x_2}'$ is an actual cause of $Y = y$ in $(M, \boldsymbol{u})$ (where $\boldsymbol{x_2}$ and $\boldsymbol{x_2}'$ are the relevant restrictions to $\boldsymbol{X_2}$).*

Finally, replacing strong sufficiency with direct sufficiency offers a notion of *direct causation* .

Definition 21 *$\boldsymbol{X} = \boldsymbol{x}$ is a* **direct cause** *of $Y = y$ in $(M, \boldsymbol{u})$ if it is part of an actual direct good sufficient explanation of $Y = y$ in $(M, \boldsymbol{u})$.*

Proposition 22 *If $\boldsymbol{X} = \boldsymbol{x}$ is a direct cause of $Y = y$ in $(M, \boldsymbol{u})$ then there exist values $\boldsymbol{x}'$ such that $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an actual cause of $Y = y$.*

We can now present two more results regarding the limitation of **Independence**. Similar to the first result, many of the distinctions we have just made disappear.

Theorem 23 *If a causal model M satisfies* **Independence** *then for any $\boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{N_1}, \boldsymbol{N_2}$, it holds that there exist values $\boldsymbol{w_1} \in \mathcal{R}(\boldsymbol{W_1})$ so that $(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_1} = \boldsymbol{w_1}, \boldsymbol{N_1})$ is a good counterfactual explanation of $Y = y$ iff there exist values $\boldsymbol{w_2} \in \mathcal{R}(\boldsymbol{W_2})$ so that$(\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{W_2} = \boldsymbol{w_2}, \boldsymbol{N_2})$ is a good counterfactual explanation of $Y = y$.*

The next result shows that the different notions of causation also reduce to each other.

Theorem 24 *If a causal model M satisfies* **Independence** *then the following statements are all equivalent:*

- *$\boldsymbol{X} = \boldsymbol{x}$ is a direct cause of $Y = y$ in $(M, \boldsymbol{u})$.*
- *there exist values $\boldsymbol{x}'$ so that $\boldsymbol{X} = \boldsymbol{x}$ rather than $\boldsymbol{X} = \boldsymbol{x}'$ is an actual cause of $Y = y$ in $(M, \boldsymbol{u})$.*
- *$\boldsymbol{X} = \boldsymbol{x}$ is part of a good sufficient explanation of $Y = y$ in $(M, \boldsymbol{u})$.*

6.1. Actual Causation and Fairness

So far this paper has focused on explanations as they relate to forward-looking action-guiding, but explanations (and actual causes in particular) are also of fundamental importance in backward-looking contexts, where the primary aim is to evaluate what has already happened. One such area that is of particular concern to XAI is that of fairness, which aims to evaluate whether a protected variable "contributed" in some way or other to produce an output. Here as well do we find a growing influence of causal models, most prominently in the work on counterfactual fairness, which in its basic form can be characterized as the condition that an output should not standardly counterfactually depend – see Def. 13 – on a protected variable (Kusner et al., 2017; Loftus et al., 2018). As we have seen, actual causes offer better explanations than counterfactual explanations, and thus I propose that fairness should be evaluated using actual causation instead.

Interestingly though, there already is a growing consensus that counterfactual dependence is too strong a condition and should be replaced with counterfactual dependence along *unfair paths* (again relying for its foundations on Pearl's work (Pearl, 2001)) (Loftus et al., 2018; Nabi and Shpitser, 2018; Chiappa, 2019; Wu et al., 2019). My definition of actual causation naturally accommodates this replacement as well, for actual causation occurs along a network $\boldsymbol{N}$. Therefore we can define fairness by demanding that protected variables do not cause the outcome along an unfair network, i.e., a network that consists entirely of unfair paths.

Concretely, I propose the following deterministic definition of fairness, which can easily be extended into a probabilistic definition by moving to probabilistic causal models.

Definition 25 *Given a causal model M, a protected variable A, and a set of unfair paths Pa_M, we say that* the model is **fair** for A relative to Pa_M *if there does not exist a context $\boldsymbol{u}$ and values y, a, a' so that $A = a$ rather than $A = a'$ is an actual cause of $Y = y$ in $(M, \boldsymbol{u})$ relative to some $(\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{N})$ such that $Pa_{\boldsymbol{N}} \subseteq Pa_M$, where $Pa_{\boldsymbol{N}}$ are the paths in $\boldsymbol{N}$.*

We consider a very simple example as an illustration. The equations are $Y = \neg B \wedge \neg C$, $B = A$, $C = \neg A$, all variables are binary, and all paths from A to Y are unfair (meaning path-specific counterfactual fairness reduces to counterfactual fairness). Y never standardly counterfactually depends on A, and thus existing notions of path-specific counterfactual fairness would always consider this model fair for A. In contrast, in a context where $A = 1$, $A = 1$ rather than $A = 0$ is an actual cause of $Y = 0$ (along $\{B\}$), and thus Definition 25 would consider this model unfair. Imagine Y represents the outcome of a hiring process such that a candidate is hired if neither Billy (B) nor Cindy (C) rejects the applicant. Billy doesn't like religious people ($A = 1$), and therefore rejects their applications. Cindy doesn't like areligious people ($A = 0$), and therefore rejects their applications. So we get that a candidate's religiosity caused their application to be denied, and this matches the intuition that the outcome is unfair in a context where religiosity is a protected variable.

7. Conclusion

This paper has integrated work on causation into the field of action-guiding explainable AI by formally defining several causal notions of explanation as well a novel account of causal fairness. I explored the connections between these notions as well as the consequences that come with ignoring the causal structure. Moreover, the proposed integration of actual causation with explanation extends beyond the practical needs of XAI to philosophy and science in general.

Acknowledgments

I would like to thank the reviewers, as well as Timo Freiesleben, Konstantin Genin, Amir-Hossein Karimi, and Gunnar König for helpful comments on a preliminary version of this paper. This research was initially funded by the NIAS-Lorentz Theme-Group Fellowship on "Accountability in Medical Autonomous Expert Systems: Ethical and Epistemological Challenges for Explainable AI", and later by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645

References

Sander Beckers. Causal sufficiency and actual causation. *Journal of Philosophical Logic*, 50:1341–1374, 2021.

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

Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *Proceeding of the 24th European Conference on Artificial Intelligence*, pages 712–720, 2020.

Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

Joseph Y. Halpern. *Actual Causality*. MIT Press, 2016.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

Christopher Hitchcock. Actual causation: What's the use? In Christopher Hitchcock Helen Beebee and Huw Price, editors, *Making a Difference: Essays on the Philosophy of Causation*, pages 116–131. Oxford University Press, 2017.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 265–277, 2020.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint*, https://arxiv.org/pdf/2010.04050.pdf, 2021a.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021b.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, 2021.

Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. A causal perspective on meaningful and robust algorithmic recourse. In *International Conference on Machine Learning: Workshop on Algorithmic Recourse*, 2021.

Tara Koopman and Silja Renooij. Persuasive contrastive explanations for bayesian networks. In *Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 229–242, 2021.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4069–4079, 2017.

David Lewis. Causation. *Journal of Philosophy*, 70:113–126, 1973.

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

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint*, https://arxiv.org/pdf/1805.05859.pdf, 2018.

S.M. Lundberg and S.I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint*, http://arxiv.org/abs/1912.03277, 2019.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. *arXiv preprint*, https://arxiv.org/pdf/2007.04131.pdf, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 32, pages 1931–1940, 2018.

Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference; 2nd edition*. Cambridge University Press, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., 2018.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

M.T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of The 32nd AAAI Conference on Artificial Intelligence*, volume 32, pages 1527–1535, 2018.

Singh S. Guestrin C. Ribeiro, M.T. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. In *Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks*, volume 109, pages 612–634, 2021.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172, 2020.

Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5103–5111, 2018.

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

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harvard Journal of Law & Technology*, 31 (841), 2017.

David S. Watson, Limor Gultchin, Ankur Taly, and Lucianio Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, 2021.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.