Minds and Machines (2023) 33:347–377
https://doi.org/10.1007/s11023-023-09637-x

# Explainable AI and Causal Understanding: Counterfactual Approaches Considered

**Sam Baron**[1]

Received: 25 October 2022 / Accepted: 14 May 2023 / Published online: 9 June 2023


**Abstract**
The counterfactual approach to explainable AI (XAI) seeks to provide understanding of AI systems through the provision of counterfactual explanations. In a recent systematic review, Chou et al. (Inform Fus 81:59–83, 2022) argue that the counterfactual approach does not clearly provide *causal understanding*. They diagnose the problem in terms of the underlying framework within which the counterfactual approach has been developed. To date, the counterfactual approach has not been developed in concert with the approach for specifying causes developed by Pearl (Causality: Models, reasoning, and inference. Cambridge University Press, 2000) and Woodward (Making things happen: A theory of causal explanation. Oxford University Press, 2003). In this paper, I build on Chou et al.'s work by applying the Pearl-Woodward approach. I argue that the standard counterfactual approach to XAI is capable of delivering causal understanding, but that there are limitations on its capacity to do so. I suggest a way to overcome these limitations.

**Keywords** Counterfactuals · Explanation · Causation · Interventions · Understanding · XAI

## 1 Introduction

Artificial intelligence algorithms, especially machine learning models in the form of deep neural networks, are being rolled out as a tool for decision-making in a number of domains. From medical diagnosis, to loan decisions and judgements about criminal recidivism, machine learning models are helping us to make decisions (Kononenko, 2001; McGrath et al., 2018; Tollenaar & van der Heijden, 2013). Many of these models are opaque: no-one understands why the models return the outputs that they do. This gives rise to one of the central challenges for explainable AI (XAI),

✉ Sam Baron
samuel.baron@acu.edu.au

[1] Dianoia Institute of Philosophy, Australian Catholic University, 250 Victoria Parade, East Melbourne, Australia

namely to provide people with understanding of why machine learning models yield specific outputs.

A number of strategies now exist for meeting this challenge (for an overview, see Molnar (2020)). One of these, pioneered by Wachter et al. (2018), appeals to *counterfactual explanations*: explanations concerning how a model's inputs would need to change in order to yield an output of a specific kind. The use of counterfactual explanations is now seen as one of the central ways to meet the demands imposed by XAI. This has led to the development of several strategies for identifying counterfactual explanations, and to a critical literature on the use of counterfactual approaches.[1]

Recent work on the counterfactual approach has reemphasised the importance of *causal understanding*. In a systematic review, Chou et al. (2022) argue that the provision of causal understanding is a necessary condition on XAI and thus that "causal approaches should be emphasised" (Chou et al., 2022, p. 78). However, they also show that "the literature connecting causal relations to explainable AI is scarce." (Chou et al., 2022, p. 66). One upshot of this lacuna, they argue, is that it remains unclear whether the counterfactual approach in fact provides information that supports causal understanding, namely understanding of what caused a given model to yield a given output, in a given case. Verma et al. (2020, p. 8) make a similar point in their review, noting that "although counterfactual explanations have been credited to elicit causal thinking and provide actionable feedback to users, they do not tell which feature(s) was the principal reason for the original decision, and why" (see also (Sokol & Flach, 2019, p. 3)). Chou et al. (2022) diagnose the problem in terms of the underlying framework within which the counterfactual approach to XAI has been developed. To date, the counterfactual approach has not been developed in concert with the framework developed by Pearl (2000) and Woodward (2003) for specifying causes, and so we are not yet in a position to know whether information that supports causal understanding has been produced.

Note that the *formal* aspects of the Pearl-Woodward framework have been implemented in the context of the counterfactual approach to XAI (see, for instance, Karimi et al. (2021); Mahajan et al. (2019)). Chou et al. (2022)'s focus, however, is not on the formal components of the framework but, rather, on the approach to specifying causes embedded within that framework. In addition to providing formal tools for modelling causal systems, the Pearl-Woodward framework tells us what causation is. Chou et al. (2022)'s contention is that counterfactual approaches to XAI have not been analysed in terms of this further aspect of the Pearl-Woodward account. Analysing counterfactual approaches to XAI in this way is important, since doing so is needed to determine whether these approaches identify genuine causal dependence.

[1] For discussion and developments of the counterfactual approach see Barocas et al. (2020); Dandl et al. (2020); de Oliveira and Martens (2021); Dhurandhar et al. (2018); Karimi et al. (2021); Kasirzadeh & Smart (2021); Keane & Smyth (2020); Laugel et al. (2018, 2019); Van Looveren and Klaise (2021); McGrath et al. (2018); Mothilal et al. (2020).

In this paper, I build on Chou et al. (2022)'s work by analysing the counterfactual approach to XAI using the Pearl-Woodward approach to specifying causes. Because the Pearl-Woodward approach is also based on counterfactual dependence, I will henceforth call the counterfactual approach that builds on the work of Wachter et al. (2018) the standard counterfactual approach. I thus argue that the standard counterfactual approach to XAI delivers causal information that supports understanding in the Pearl-Woodward sense of causation, but that it does not reliably deliver full causal information. I go on to propose a way of overcoming the limitations on the standard counterfactual approach by combining it more fully with the Pearl-Woodward framework. The paper thus makes three advances: (i) it develops the connection between causal understanding and standard counterfactual approaches to XAI; (ii) it shows how the standard counterfactual approach to XAI can provide a basis for causal understanding, and where it faces limitations in this respect and (iii) it proposes a new, mixed strategy for XAI.

In recent work, Buijsman (2022) also considers the Pearl-Woodward framework in the context of XAI, including its relationship to standard counterfactual approaches. This important work differs from the study undertaken here in a couple of ways. Whereas Buijsman (2022) focuses on Woodward's notion of *explanation*, I focus on his notion of *causation*.[2] Causation, for Woodward, is necessary for explanation, but not sufficient. What is needed, in addition, is a generalisation connecting causes to effects, and it is the generalisation that interests Buijsman (2022). Moreover, Buijsman (2022) argues that we should replace the standard counterfactual approach with one based on Woodward's notion of explanation, I offer no such argument. Instead, I use Woodward's notion of causation as a tool for assessing the capacity of standard counterfactual approaches to find causes. Ultimately, however, Buijsman and I make similar recommendations, which is to develop an approach to XAI that is based on the Pearl-Woodward framework.

Watson and Floridi (2021) are also interested in using the Pearl-Woodward framework for XAI. Again, while this work significantly advances our understanding of XAI, it differs from the current study in key respects. By contrast to Buijsman (2022), Watson & Floridi (2021) are more focused on the formal aspects of the Pearl-Woodward framework. These formal aspects do not take centre stage here. More important is the notion of causation, and the Pearl-Woodward strategy for specifying causes. Moreover, like Buijsman (2022), Watson & Floridi (2021) are interested in developing a new approach to XAI based mainly on the Pearl-Woodward framework. By contrast, my interest is in combining standard counterfactual approaches with the Pearl-Woodward framework to develop a dual approach to XAI.

Further discussion of the Pearl-Woodward framework can be found in recent work by Asher et al. (2022) and Beckers (2022). Both approaches to XAI yield important insights into the application of the Pearl-Woodward approach. Asher et al. (2022) provide a logical framework for counterfactuals, along with a game-theoretic analysis of the procedure of giving explanations. A core focus of their approach is

[2] For recent, useful discussion of scientific explanation in AI, along with its connection to causation, see Dúran (2021); Nyrup & Robinson (2022).

to provide a semantics for counterfactuals, in part by using causal models of the kind advocated by Pearl and Woodward. As with Watson and Floridi (2021), Asher et al. (2022) focus more on formal matters, rather than the implementation of the Pearl-Woodward notion of causation, which is the focus here. The work of Beckers (2022) is much closer to the present study, insofar as he is interested in applying (and, indeed, refining) notions of causation. However, Beckers (2022) focuses primarily on applying notions of causation to *target systems*: real-world systems that machine learning models represent. By contrast, my focus is on applying causation to machine learning models themselves, not their target systems.[3]

With these preliminaries aside, I can now lay out the plan for the rest of the paper. I begin, in §2, by briefly introducing XAI and the associated goal of producing information that supports causal understanding. In §3, I introduce the standard counterfactual approach to XAI and in §4 I introduce the Pearl-Woodward framework for causation and apply it. In §5 I suggest supplementing the standard counterfactual approach with the Pearl-Woodward framework before summarising the main findings of the paper in §6.

## 2 Causal Understanding

In the simplest terms, a machine learning model is a system that takes certain values for variables as inputs, and yields certain outputs. These models are typically produced via the application of an algorithm to a large data set, which is used to 'train' the way that the model produces outputs based on inputs. Once trained, the model can be fed input data of the kind used to train the model. The model will then process this data and deliver a prediction that can be used for decision-making purposes. As noted, some machine learning models are opaque. The opacity of machine learning models is generally considered to be a problem. When a decision is made by an institution, and that decision has serious implications for an individual, the individual should be able to understand why the decision was reached or prediction was made. This general 'right to explanation' translates directly into a right to understand why machine learning models produce their outputs, when those outputs are used by institutions to make decisions that affect people's lives.

The project, then, of developing XAI is, in part, the project of providing explanations of why machine learning models yield the results that they do. Wachter et al. (2018, p. 843) outline three goals for the provision of such explanations:

1. To inform and help the subject understand why a particular decision was reached.
2. To provide grounds to contest adverse decisions.

[3] In the useful typology of explanation provided by Cabitza et al. (2023), Beckers (2022) focuses on 'causal explanation' which Cabitza et al. (2023) take to be a matter of studying causation in a real-world system, whereas I focus on what Cabitza et al. (2023) call 'mechanistic explanation', which involves the way that the machine learning model works.

3. To understand what could be changed to receive a desired result in the future, based on the current decision-making model.

As noted, a number of authors have emphasised the importance of *causal understanding* in the context of XAI (Buijsman (2022); Chou et al. (2022); Holzinger et al. (2019); Miller (2019); Shin (2021); Watson & Floridi (2021)). At the most general level, for a model $M$ with a set of input variables $X$, causal understanding is a matter of correctly identifying those variables in $X$ that caused $M$ to yield a specific output $\psi$ rather than $\phi$, on a particular run. Causal understanding is thus contrastive: it is a matter of understanding what caused $M$ to have $\psi$ rather than $\phi$ as an output. Causal understanding also comes in degrees. When there are many variables that, together, caused the model's output, one can gain information about just one of these variables, or all of them. One has partial causal understanding when one correctly identifies one or more causes of a model's output. One has full causal understanding when one correctly identifies all causes of a machine learning model's output, and does not misidentify any causes. Correct identification is a matter of true belief: one correctly identifies a cause $x$ when one gains the belief that $x$ caused $M$ to output $\psi$ rather than $\phi$ *and one's belief is true*.

It is important to differentiate the kind of causal understanding at issue from three other notions of causal understanding that are sometimes discussed in the context of XAI. First, there is understanding of the causal structure of a physical system, one that is distinct to a machine learning model. So, for instance, there might be a particular physical system being studied within physics. A machine learning model might be employed to try and understand the causal structure of that system (see the discussion in Dúran & Formanek (2018); Räz & Beisbart (2022); Sullivan (2022)). Causal understanding in this sense is a measure of how much the model can reveal about the causal structure of this other system. By contrast, the kind of causal understanding I am interested in concerns the causal structure of the machine learning model itself.

Second, there is understanding of the causal relationships between actual features that are encoded by the input variables of a machine learning model (see, for instance, Karimi et al. (2021)). As with the first kind of causal understanding, this is understanding of the relationship between various facts in the world, such as the way that income and zip code causally interact in a real society. It is not understanding of what caused the machine learning model to yield a certain output which, again, is what interests me here.

Third, there is understanding of what an individual has the capacity to causally control (see, for discussion, Keane et al. (2021); Keane & Smyth (2020)). Understanding of this kind involves an understanding of the causal options one has for action. This is quite different to understanding which inputs into a machine learning model caused it to output a specific value. The two notions can interact, however: understanding what caused a machine learning model to yield a specific outcome is important to understanding what one should change to receive a better outcome from the model which, in turn, requires understanding one's causal options for change.

According to Chou et al. (2022, p. 61) the provision of causal understanding is a necessary condition on satisfying the demand for explanation that drives work on

XAI. The necessity of causal understanding makes sense from a philosophical perspective. Philosophers generally agree that the following is a necessary condition on explanatory understanding:[4]

> **Understanding Why:** *S* understands why *p* only if *S* believes that for some *q*, *q* caused *p* and *S*'s belief is true.

Causal understanding is thus needed to achieve the first of the three goals for XAI outlined by Wachter et al. (2018). As noted, the first of these goals just is the provision of explanatory understanding. Given that causal understanding is necessary for explanatory understanding, it follows that causal understanding is needed for XAI.

In what follows, I will use the Pearl-Woodward framework for specifying causes to assess whether the standard counterfactual approach manages to provide users with causal information. Why relate the standard counterfactual approach to the Pearl-Woodward framework in this manner? Doing so is non-trivial and, one might contend, it is unclear what the benefits might be.

The answer to this challenge lies with the connection between explanatory and causal understanding. It is this link that, in part, motivates the focus on counterfactuals in XAI. As Miller (2019) notes in their seminal review, counterfactual explanations are important precisely because they are supposed to provide information about causes (a point that has been subsequently emphasised in e.g., Asher et al. (2022); Byrne (2019); Buijsman (2022); Chou et al. (2022); Kasirzadeh and Smart (2021)). Thus, providing counterfactual explanations is supposed to be a way of providing information about causation. This information is then supposed to form the basis for user's beliefs about the causal factors that led a machine learning model to have a particular outcome.

It is important, however, to differentiate between two kinds of information about causation: genuine information and spurious information. Genuine information about causation, is information that accurately captures causal factors. Spurious information, by contrast, is information that, while appearing to be about cause and effect, fails to accurately capture causal factors. In the case of XAI, we need to provide individuals with genuine information about causation. If we don't, then we are not providing them with information that can lead to explanatory understanding. For without genuine causal information, users are not in a position to form true beliefs about causal factors, which is needed to satisfy the necessary condition on explanatory understanding stated above.

This suggests the need for a method of *certifying* that the information given to users via a particular approach to XAI is genuine causal information. In the case of the standard counterfactual approach, this means certifying that the information communicated to users through counterfactuals accurately captures causal factors rather than, say, spurious correlations (Chou et al. (2022)). However, the only way

[4] Philosophers disagree about whether understanding of causes is sufficient for explanatory understanding. In addition to understanding of causes, something else may be required, such as an explanatory story linking causes to effects (Pritchard, 2014); an answer to a vertical why-question (Lawler, 2019); or inferential and explanatory abilities related to one's understanding of causes (Hills, 2016).

to certify that the information given to users is genuinely causal is to show that the information being provided satisfies a plausible picture of what causation *is*.

In short, causal certification is necessary for determining whether an approach satisfies the first goal of XAI outlined by Wachter et al. (2018), namely the provision of explanatory understanding. For there is a risk that, when providing an explanation for why a machine learning model produces a certain output, one provides an explanation that, while psychologically compelling, does not convey genuine information about causal factors. Causal certification ensures that the information being provided to users is indicative of genuine causation, and thus provides a sound basis for explanatory understanding. Since the provision of explanatory understanding is increasingly a focus of regulation around AI, causal certification is potentially needed to demonstrate regulatory compliance. For if one cannot show that the information provided is a basis for explanatory understanding, then it is unclear that one has provided enough to satisfy a right to explanation.

One might reply that genuine information about causal factors is not necessary for explanatory understanding. The goal of providing explanatory understanding is achieved so long as one provides an explanation that a user will find plausible or satisfying. What we should do, then, is just focus on providing this kind of information, regardless of whether it is genuinely causal. Here it is important to differentiate between *perceived* understanding and *genuine* understanding. Perceived understanding is a subjective experience that occurs when one feels as though one understands, which can often occur when one is given an explanation that one finds satisfying. Genuine understanding, by contrast, is when one really does understand a phenomenon.

In contrast to perceived understanding, genuine understanding cannot be achieved just by gaining an explanation of some fact that one deems satisfying or plausible. Rather, as philosophers have shown, genuine understanding is a more demanding notion, characterised by having certain true beliefs and possessing specific capacities, rather than having subjective experiences (Hills (2016); Sullivan (2018); Wilkenfeld (2014, 2019)). Importantly, when one genuinely understands, one has true beliefs about causal factors, and one is able to draw a range of accurate inferences about causal factors, including accurate generalisation to other, similar cases (see Hills (2016) for a list of the relevant capacities). Such beliefs and inferential capacities may be gained when given an explanation that is psychologically compelling, but not necessarily. One may find an explanation compelling because it is familiar or, indeed, presents a clear plan of action, rather than because it gives one a true belief about causes or because it gives one the capacity to draw accurate causal inferences about the phenomenon being explained.

It is plausible that the kind of explanatory understanding at issue in XAI is, at least sometimes, genuine explanatory understanding, rather than just perceived understanding. We want to inform users of how something works not just in a manner that they find satisfying, but in a manner that gives them true beliefs about causes and supports their capacity to draw accurate causal inferences. But if that's right then it is not enough to 'scratch' a user's 'explanatory itch' by giving them a perceived sense of understanding only; genuine explanatory understanding is needed. Because genuine understanding is so tightly linked to true beliefs

about causation, providing a basis for genuine explanatory understanding means supplying users with information enough to form accurate beliefs about causal factors, that can then underwrite inferential capacities. This, in turn, requires the provision of accurate causal information which, again, highlights the need to certify that accurate causal information is being provided.

In sum, one good-making feature of explanations provided in XAI is that they provide information that leads users to form true beliefs about causation. This is not the only sense in which those explanations may be good, but it is an important sense because genuine explanatory understanding requires that users form such beliefs. Accordingly, we need a system for certifying that the information provided to users by a given approach to XAI is genuine causal information.

There are, in fact, two ways in which an approach to XAI may be causally certified. First: *basic causal certification*. This is a guarantee that the information provided to users is always genuine causal information. Second: *complete causal certification*. This is a guarantee that the information provided to users is always a complete account of the causal factors that led a model to deliver a particular outcome.

Complete causal certification is needed for the second goal outlined by Wachter et al. (2018) for XAI, namely that of providing grounds for contesting adverse decisions. One important way to contest an adverse decision is by showing that the output of a machine learning model was caused by a particular input feature that should not have played any causal role. For instance, one may wish to contest an adverse loan decision on the grounds that the loan model's output was caused by input features like race or age. In order for a user to properly decide whether to contest a particular decision, they need complete information about the causal factors that determined the outcome in their case. Without this information, the user cannot determine whether, say, race or age are causally important (assuming that their causal importance is sufficient grounds for recourse). Without a guarantee that full causal information has been given one cannot know whether one is in a sound position to contest the decision, and thus whether challenging the decision—which can be costly—is worthwhile.

The Pearl-Woodward framework is important for causal certification. What it provides is a framework for determining whether an existing approach to XAI always delivers genuine causal information, and, if it does, whether it always delivers complete causal information. The framework can thus be used to test whether an existing approach to XAI passes basic or complete causal certification. If an existing approach fails to pass causal certification, then that provides a potential basis for criticising the approach. Since what it shows is that the approach is not providing the right kind of information to scaffold genuine explanatory understanding. In this way, the Pearl-Woodward approach can be used to select between competing approaches to XAI. If, by contrast, an existing approach passes causal certification, then this information can be provided in the form of a guarantee to users. That is, we can assure users that the explanatory information they've been given is genuine causal information. Any doubts about the causal status of the information being provided to users can thus be assuaged via causal certification. This is important since confidence in causal information

can underwrite trust (see Shin (2021)) and so having a way to offer causal certification of approaches to XAI is of potential value to users.

As we shall see in §5, the Pearl-Woodward framework can also be used as a supplement to existing approaches that fail to achieve causal certification. That is, the Pearl-Woodward framework can be used alongside an existing approach to XAI as a way to ensure that users are provided with complete causal information. Indeed, applying the Pearl-Woodward framework in this way is one of the core recommendations of this paper. Thus, not only can the Pearl-Woodward framework be used to test other approaches to XAI, it can also be used to ensure that complete causal information is always provided to users, thereby placing users in the best position to contest adverse decisions.

The Pearl-Woodward framework and the attendant causal analysis is thus beneficial for three, related reasons: (i) it provides a way to test the quality of the explanations delivered by an approach to XAI, which is important for certifying that causal information is being provided; (ii) it can be used to issue guarantees regarding the quality of information being provided to users and (iii) it provides a framework for ensuring that all relevant causal information is provided. Upshots (i) and (iii) are of practical benefit insofar as they contribute to achieving the goals of XAI, and upshot (ii) is useful as it has the potential to underwrite trust in XAI approaches.

Despite these benefits, one may still question the use of the Pearl-Woodward framework. There are, after all, other ways of specifying causation, and one in particular looms large: the account of causation offered by Lewis (1973) which, is often cited as a motivation for the standard counterfactual approach. In the next section, then, I will introduce the the standard counterfactual approach to XAI and explain why Lewis's approach shouldn't be used for causal certification.

## 3 Counterfactual Explanations

The standard counterfactual approach aims to achieve the goals of XAI through the provision of counterfactuals. Each counterfactual provides information about how the outputs of a machine learning model would have been different, under counterfactual changes to the model's input variables. So, for a specific machine learning model $M$ that yields output $\psi$ based on a set of input variables $X$, each counterfactual has the following form: for a set of variables $\{X_1 ... X_n\} \subseteq X$, with values $x_1 ... x_n$, had the values of variables $X_1 ... X_n$ been $x'_1 ... x'_n$ rather than $x_1 ... x_n$, $M$ would have yielded output $\phi$ rather than $\psi$. Counterfactuals of this form are then used to underwrite explanations for why $M$ yielded the specific output that it did, in the following manner: $M$ yielded output $\psi$ because variables $X_1 ... X_n$ had values $x_1 ... x_n$. Had $x_1 ... x_n$ been $x'_1 ... x_n$, $M$ would have yielded $\phi$ rather than $\psi$.

The goal is usually to find counterfactuals of the above form that satisfy a constraint of *proximity*. Let a data point $p$ for a machine learning model $M$ with a set of input variables $X$ be a particular setting of values for each variable in $X$. Then the set $S$ of data points is the set of all possible combinations of settings for the input variables. Now, for a machine learning model $M$ with output $\psi$ consider all of the counterfactuals of the following form: had $M$ received data point $p'$ rather than $p$, it

**Table 1** A list of counterfactuals

| Data point | Age | Income | Savings | Suburb | Marital | Debt | Output |
|---|---|---|---|---|---|---|---|
| @ | 25 | 60k | 20k | Fitzroy | Single | 200k | $\psi$ |
| $p_1$ | | | +4k | | | | $\phi$ |
| $p_2$ | | +4k | | | | | $\phi$ |
| $p_3$ | -6 | | | | | | $\phi$ |

The first line of Table 1 is the actual data input to the model for Sara

The other lines $p_1 - p_3$ are the smallest changes to the inputs for Sara's data that would yield a different outcome

Blank spaces for $p_1 - p_3$ are values that are unchanged from their actual input values

would have yielded output $\phi$ rather than $\psi$. For each of these counterfactuals, compare the counterfactual data point $p'$ with the actual data point $p$, along one or more dimensions. Proximity tells us to focus on those data points that flip the model's output and that are closest to $p$ with respect to the relevant comparison. Only counterfactuals involving the closest data points are to be used in explanations of why model $M$ yielded a specific output $\psi$.

To develop their approach, Wachter et al. (2018) identify an algorithm along with a metric that provides a measure of the distance between each data point. This algorithm satisfies the proximity constraint by looking for the closest data points that flip the model's output (see (1)–(3)).

$$arg \min_{x'} \max_{\lambda} \ \lambda(f_w(x') - y')^2 + d(x_i, x') \tag{1}$$

$$d(x_i, x') = \sum_{k \in F} \frac{|x_{i,k} - x'_k|}{MAD_k} \tag{2}$$

$$MAD_k = median_{j \in P}(|X_{j.k} - median_{l \in P}(X_l, k)|) \tag{3}$$

To see how the standard counterfactual approach works, it is useful to consider an example. Suppose that Sara applies for a loan and is asked to provide information including: income, age, assets, address, marital status and debt level. The bank takes this information and feeds it into a machine learning model. The model returns a result $\psi$ that is used by the bank to disqualify Sara for the loan. Sara demands to know why the model returned this specific result, rather than a result $\phi$ that would have been more beneficial for her. In order to satisfy this demand, the bank provides her with a list of counterfactuals. These counterfactuals represent the smallest changes to Sara's input data that would lead the machine learning model to produce $\phi$ instead of $\psi$ (see Table 1). This list of counterfactuals is then used to provide Sara with an explanation for why her loan application was rejected. It was rejected because her income and savings were too low, and because she was too young.

Counterfactual approaches vary in terms of the precise method used to retrieve a list of counterfactuals. A number of different algorithms have been suggested, as

well as a number of different metrics (see Chou et al. (2022); Verma et al. (2020); de Oliveira and Martens (2021) for overviews). The identification of new algorithms, metrics and strategies for finding counterfactuals is driven by an appreciation of various constraints on which counterfactuals should be returned. One constraint that is common to almost all versions of the standard counterfactual approach is the proximity constraint described above. However, there are a number of other important constraints that have been identified in the literature to date.

For instance, Keane and Smyth (2020) and Keane et al. (2021) identify counterfactuals that are *plausible* or *actionable*. A similar approach is taken by Kirfel and Liefgreen (2021). Actionable counterfactuals are those where the counterfactual data point $p'$ is closest to the actual data point $p$ compatible with the shift from $p$ to $p'$ being achievable by an individual. In a similar vein, Poyiadzi et al. (2020) include modifications to retrieve the most *feasible* data points, which are ones that are the easiest for the individual to access from the actual data point. Wachter et al. (2018) emphasise *sparsity*, which involves finding data points that involve changing as few input variables as possible while still changing the model's output. Mothilal et al. (2020) emphasise *diversity*, where diversity is a measure of the number of counterfactuals delivered, and is to be balanced with actionability.

I will return to some of these constraints later on. For now, I will simply note them and press on to consider how the standard counterfactual approach might yield causal information. As discussed, the output of the standard counterfactual approach is a list of counterfactuals. This list of counterfactuals will single out specific input variables. So, for instance, in Sara's loan case, the list of counterfactuals singles out income, savings and age. Since this list of counterfactuals is the only output of the standard counterfactual approach, and since it is supposed to underwrite explanation and understanding, it is natural to try and read causation off the list. We can do this by focusing on the variables that are altered. Thus, in the case of Sara's loan, we can infer that it was age, income and savings that caused the model to yield its specific output in her case, since it is those variables that, when changed, flip the model's output.

For information generated in this way to underwrite genuine causal understanding, the information must accurately capture causal factors. That's because, as noted in §1, causal understanding is a matter of correctly identifying causes, which requires not just belief about causes but true belief, and accurate information about causation is needed for true causal beliefs. As Chou et al. (2022) argue, however, it is unclear whether the information generated from the standard counterfactual approach is genuine causal information. The standard counterfactual approach, they maintain, has not been developed within a framework that allows us to determine when we have correctly identified a cause. What is missing, in particular, is a method of specifying what causes what. Without such a method, we don't know whether the information extracted from a list of counterfactuals matches the actual causal facts.

One might take issue with Chou et al. (2022)'s argument, noting that Wachter et al. (2018) do in fact develop their account within a framework that can be used to specify causation. In fact, two such frameworks are mentioned in their paper. The first of these is Lewis's (1973; 1979) and the second is the one outlined by

Pearl (2000) and Woodward (2003). Subsequent work on the standard counterfactual approach has also been developed with reference to these two frameworks. For instance, Mahajan et al. (2019) and Karimi et al. (2021) employ an approach to finding counterfactuals based on the Pearl-Woodward structural equation framework. Similarly, Russell et al. (2020) develop an approach to counterfactuals using Lewis's framework, defending it against objections from Pearl.

Chou et al. (2022)'s criticism cannot be so easily set aside, however. For it is important to distinguish between two aspects of the Lewis and Pearl-Woodward frameworks. On the one hand, there is the formal recommendation about how to work with counterfactuals. Lewis proposes a method that is based on distance between worlds; Pearl and Woodward propose a method based on interventions and structural equation models. On the other hand, there is a way of specifying causes embedded within both frameworks, where a way of specifying causes is just a statement of what it is for one thing to cause another. Within Lewis's framework, causes are specified in terms of counterfactual dependence of a certain kind. For him, counterfactual dependence of the relevant kind is sufficient for causation. Within the Pearl-Woodward picture, by contrast, causes are specified in terms of interventions on variables, which leads them to focus on a different class of counterfactuals.

Chou et al. (2022)'s point is focused on this second aspect of the Lewis and Pearl-Woodward frameworks. Their point is that despite the use of the first aspect of the Lewis and Pearl-Woodward frameworks within counterfactual approaches to XAI, an approach to specifying causes (such as one of the two approaches just mentioned) has not been systematically applied in order to work out whether standard counterfactual approaches manage to yield causal understanding. To return to the language of causal certification offered above: their point is that causal certification has not been achieved for the standard counterfactual approach to XAI. That's because we have not yet verified that the counterfactuals being provided do indeed yield genuine causal information.

In fact, their point is a bit narrower than that: their concern is that the Pearl-Woodward approach to specifying causes in particular has not been systematically applied to achieve causal certification. One might take issue with their focus on the Pearl-Woodward approach to specifying causes. For if Lewis's approach to specifying causes is assumed instead, then it could be argued that it is relatively easy to achieve causal certification for the standard counterfactual approach. That's because, one might argue, for Lewis, counterfactual dependence is sufficient for causation. Accordingly, *any* true counterfactual reveals genuine causal information, and so the provision of any counterfactual explanation whatsoever provides a basis for inferring true beliefs about how a machine learning model's output causally depends on its input.

Matters are not quite so straightforward, however. For one thing, Lewis does not take counterfactual dependence *in general* to be sufficient for causation. For Lewis, it is only very specific counterfactuals that are sufficient for causation. This restriction is important, because as Lewis recognised, there are clear cases in which there is counterfactual dependence but no causation (Reutlinger (2016)). Consider, for instance, the way that a diamond is constituted by molecules that are in a tight lattice. Given this fact, the following counterfactual seems to be true: if the diamond's

molecules were not in a tight lattice, the diamond would not have been hard. This counterfactual is true, but it is not indicative of causation: the diamond's molecules don't cause it to be hard but, rather constitute it's hardness (constitution is not supposed to be a causal relation, see Baumgartner & Gebharter (2016)). Or, to take another example, when one writes down the word 'party' one has to write the 'a'. So the following counterfactual seems to be true: had one not written 'a' one would not have written 'party'. It is not clear, however, that this is a causal counterfactual: writing 'a' does not seem to cause one to write 'party' in any obvious sense. Finally, philosophers have argued that counterfactuals within pure mathematics are true, such as 'if 13 had not been a prime number, it would have had factors other than one and itself' (Baron et al. (2017)). This counterfactual, while true, does not imply causation, since mathematical objects are not the right kinds of things to be causally related.

Lewis isolates the class of counterfactuals that are sufficient for causation in two main ways. First, he imposes a restriction on the causal relata: he demands that only counterfactuals involving independent events are sufficient for causation. Second, he imposes a complex similarity ordering over counterfactuals, that requires closeness in the laws of nature and in spatiotemporal distribution of matters of fact. Only the counterfactuals that are true by this measure of similarity qualify as causal (Lewis (1979)).

Now, one could try to use Lewis's theory of causation to achieve causal certification for the standard counterfactual approach to XAI. To do this one would need to demonstrate that the counterfactuals delivered by the standard counterfactual approach do in fact qualify as causal counterfactuals, in Lewis's sense. Perhaps this can be done. However, there is good reason to move beyond Lewis's approach and to thus take Chou et al. (2022)'s focus on the Pearl-Woodward approach seriously. For philosophers generally agree that Lewis's theory fails to link counterfactuals to causation, even for the narrow set of counterfactuals that do satisfy his constraints. The consensus being that members of Lewis's favoured class of counterfactuals are not generally sufficient for causation (see, for demonstrations of this fact, Harbecke (2021); Fine (1975); Schaffer (2000); Schulz (2011); Woodward (2003)).

Indeed, it is precisely this fact that partly motivates the Pearl-Woodward approach to specifying causes. As was the case with Lewis's theory, it is only counterfactual dependence of a specific type that is sufficient for causation within the Pearl-Woodward framework. However, the framework is tailor-made to avoid the counterexamples that philosophers have raised against Lewis's theory, and to thus settle on a class of counterfactuals that is more tightly linked to causation. The Pearl-Woodward framework is also the most detailed system for understanding causation based on counterfactuals that has been developed to date. It enjoys widespread use both within philosophy and within science, and is considered to be a leading approach to specifying causes. If we are to move beyond the Lewisian picture, but stay within a broadly counterfactual approach to causation, then there really is no better option than the Pearl-Woodward framework.

In sum, then, Chou et al. (2022) set up the following challenge: show that the standard counterfactual approach to XAI yields genuine causal information by systematically applying the Pearl-Woodward approach to specifying causes. In what

remains, I take up this challenge. In the next section, I briefly outline the Pearl-Woodward approach to specifying causes. After that, I will reconsider the standard counterfactual approach in light of the Pearl-Woodward notion of a cause.

## 4 Causal Discovery

For Woodward (2003, p. 59), causes are specified as follows:[5]

> *Interventionist Causation (IC)* $X$ is a direct cause of $Y$ with respect to a variable set $V$ if and only if there is a possible intervention on $X$ that changes $Y$ or the probability distribution of $Y$ when one holds fixed at some value all other variables $Z_i$ in $V$.

There are a few things to note about this definition. First, **IC** uses variables. It is important to keep these variables distinct from the input variables of machine learning models, for reasons that will become clear as we go. Thus, we can call the variables that appear in **IC** Pearl-Woodward variables, or PW-variables, and the input variables to a machine learning model ML-variables. Second, causation in **IC** is specified with respect to a variable set. A set of variables is a way of representing the features of a system. Generally speaking, we can choose to represent the same system using multiple different variable sets, depending on how features of a system are grouped under variables. Third, causation in **IC** is defined in terms of an intervention, which amounts to a possible way of changing the values of one or more PW-variables to a value other than the actual value.[6] Fourth, holding variables fixed just means setting them at their actual values under an intervention. Fifth, **IC** is a definition of *a* cause. *A* cause is different from *the* cause. When we talk of *the* cause we mean the complete causal antecedents to some event. When we talk of *a* cause, we mean only one antecedent cause, as there may be many. Finally, **IC** is a definition of a *direct* cause. A direct cause $X$ of $Y$ is one that does not proceed via some intermediary $Z$.

**IC** is generally combined with a formal framework of causal models. A causal model is a way of representing the causal structure of a system. Each causal model can be represented as a directed, acyclic graph in which each PW-variable is a node,

[5] Note that I have altered Woodward's definition slightly for readability, but those differences do not matter. A more substantive change is that I have omitted Woodward's reference t:o *types* For Woodward, **IC** is a definition of causation at the type-level. Types are just repeatable events. For a machine learning model we are generally interested in what caused a model to have a specific output. This is, however, compatible with thinking of causation at the type-level. We just need to consider the specific output as a type of event: the event of a model yielding that specific output, which it could do on multiple distinct occasions. We can think of the cause of the model yielding this output as a type too: the model receiving a certain type of input, which it could do on multiple occasions. In this way, causation for a machine learning model can be specified at the type-level: this type of input leads to that type of output. However, since the discussion of types is not important in what follows, I set it aside.

[6] Woodward (2003) specifies interventions in terms of the addition of possible causes to a system. Here I use Pearl's (2000) method of specifying interventions in terms of 'logical surgeries' whereby the PW-variable is changed directly, without the addition of an extra cause.

the causal relations between variables are links and a set of structural equations describes the dependence between PW-variables. Interventions on PW-variables in a causal graph can be represented using Pearl's (2000) do-calculus. The do-calculus is an operation on variables that changes the value of that variable, while leaving other variables unchanged. An intervention will generally break the structural equation that specifies the way in which the intervened-upon variable depends on other variables. In this way, the do-calculus makes the value of the intervened on variable depend only on the do-operation, and not on other variables.

The formal framework of causal models can be used to represent the causal structure of a system, once that causal structure is known. Prior to building a causal model, we generally need to determine what the causal structure of a system is. For this, we can use just **IC** plus a PW-variable set. We start by representing the features of a system using PW-variables. We then intervene on each PW-variable, while holding the others fixed, to check for causation. This typically means checking a range of counterfactuals on the system to see if there are pairs of variables that satisfy **IC**. When there are, we can record the relationships between those variables as causal relationships. Because the application of **IC** reveals causal information, it can also be used to test beliefs about causes, to see if they are true. So, for instance, the belief that $x$ causes $y$ in a system $s$ can be tested by representing $s$ with a PW-variable set $V$ and then applying interventions to a variable representing $x$ to see if there is a change to a variable representing $y$ (holding all other PW-variables in $V$ fixed). If there is, then the belief that $x$ causes $y$ is true. If not, then not.

As discussed in §2, the Pearl-Woodward framework can be used as a test of causal certification for existing approaches to XAI. In order to show how **IC** can be applied to the standard counterfactual approach for this purpose, I will proceed in two stages. First, I will outline an application that is simplified in a key respect. Having done that, I will then lift the simplification to consider a more complex application of **IC**. Note that the simplification is just for expository purposes, it makes the initial application of **IC** a bit easier to follow, but beyond that is entirely dispensable.

### 4.1 A Simple Model

In order to use **IC** for the purposes of causal certification, some set-up is required. For a machine learning model $M$, we must specify a set of PW-variables $V$ and use them to represent the ML-variables for $M$. Every ML-variable corresponds to some PW-variable and distinct PW-variables correspond to distinct ML-variables.[7] The output of $M$ is represented by exactly one PW-variable which has just two possible values: the actual output $\psi$, and some pre-determined value of interest $\phi$. Each PW-variable that represents an ML-variable is set to the actual values of that

[7] Distinct PW-variables must correspond to distinct ML-variables otherwise, when we intervene on one PW-variable and hold fixed the others, we end up holding fixed the PW-variable we are intervening on. Every ML-variable needs to correspond to some PW-variable so that we can check all ML-variables using **IC.**

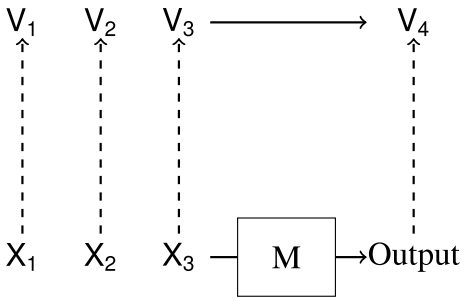


**Fig. 1** A simplified set-up for IC for set of PW-variables $V$, $\{V_1, V_2, V_3, V_4\}$ and set of ML-variables $X$, $\{X_1, X_2, X_3\}$ for machine learning model $M$. Each ML variable corresponds to exactly one PW-variable. There is a PW-variable that corresponds to the model's output. The possible values for the $V_n$ that correspond to the $X_n$ just are the possible values for the $X_n$. The possible values for the PW-variable corresponding to the model's output is just the pair of values $\psi$ (actual value) and $\phi$ (pre-determined value). The values for the $V_n$ are set to the actual values for the ML-variables and the model's output

ML-variable. The PW-variable that represents the output of the model is set to the model's actual output.

As noted, the initial application of **IC** will operate under a simplifying assumption: each PW-variable corresponds to exactly one ML-variable. Under this simplifying assumption, the possible values of each PW-variable correspond to the possible values for the ML-variable that is being represented. The set of possible data points for the machine learning model thus gives us the possible values for each of the PW-variables that corresponds to an ML-variable, since the set of possible data points encodes all of the possible values for each ML-variable. The simplifying assumption allows us to drop talk of PW-variables and talk directly of intervening on ML-variables, which can be understood as counterfactual changes to those ML-variables. That is what I will do while the simplifying assumption is in place.

Note that the simplifying assumption is not essential to the application of **IC** and will be disposed of entirely in §4.2. Disposing of the simplifying assumption is important, for it is extremely restrictive. What it does is effectively force us to consider only counterfactuals in which a *single* ML-variable is altered and the output of the model is flipped. As Keane & Smyth (2020) have argued, however, counterfactuals of this kind tend to be extremely rare. Indeed, many counterfactuals involve altering a large number (20+) of ML-variables to flip a model's output. Thus, under the simplifying assumption, the application of **IC** is bound to miss a great deal of causal information.

At any rate, the initial, simplified set-up for applying **IC** is depicted in Fig. 1.

Having embedded a machine learning model within the Pearl-Woodward framework, we are now in a position to test the standard counterfactual approach using **IC**. Recall that, for the standard counterfactual approach, we can extract information about causation using the list of counterfactuals retrieved. Specifically, for a machine learning model with output $\psi$, we can look at the list of counterfactuals that has been retrieved, and infer that those ML-variables that take values that are different to their actual values caused the model's output to be $\psi$ rather than $\phi$. When an

application of the counterfactual approach retrieves a list of counterfactuals in which certain variables are altered, I will say that these variables have been *highlighted* and the rest have not been highlighted. Thus, the information being extracted is that the highlighted ML-variables are the causal variables. So, for instance, in Table 1, income, savings and age are all highlighted, whereas debt level, address and marital status are not highlighted. We thus infer that income, savings and age caused the model's output in this case, as previously discussed.

For information extracted in this way to underwrite genuine causal understanding, it must be accurate. **IC** can be used to test the information for accuracy: for an ML-variable putatively identified as causal, one can check to see if it is in fact causal, by looking to see whether intervention on that variable satisfies **IC**. Thus, one can ask: is it the case, for a highlighted ML-variable, that there is some possible value of that ML-variable that flips the model's output from $\psi$ to $\phi$, while holding all other ML-variables fixed? If the answer is 'yes', then the application of the counterfactual approach has successfully identified a cause, and has delivered genuine causal information. In this way, the application of **IC** can be used to provide basic causal certification of the causal information generated.

**IC** can also be used as the basis for complete causal certification. For a machine learning model with output $\psi$, we check to see if the highlighted ML-variables are in fact causal using **IC** as before. Thus, we begin with basic causal certification. After that, we use **IC** again to see whether there are any ML-variables that aren't highlighted within the list of counterfactuals retrieved but that nonetheless satisfy **IC** with respect to the model's output. If every highlighted ML-variable is causal and no other ML-variables are causal, then the application of the counterfactual approach has not omitted any factors that cause the model's output. In this situation, we can certify that complete causal information has been delivered to the user.

Note that by applying **IC** in this manner we are considering similar counterfactuals to the ones delivered by the standard counterfactual approach. It is thus worth pausing to consider the difference between the standard counterfactual approach and the use of **IC**. There are two important differences. The first difference concerns the counterfactuals that are under consideration. In the standard counterfactual approach, the counterfactuals that are retrieved are held under the proximity constraint, among others. When applying **IC** to provide causal certification we do not use a proximity constraint. Thus, **IC** does *not* look for causation between $X$ and $Y$ in terms of the closest changes to $X$ that change $Y$. We are simply looking for *any* possible value for $X$ that changes $Y$.

The second, related, difference concerns what we are trying to do when using **IC** versus when applying the standard counterfactual approach. The goal when applying the standard counterfactual approach is to provide counterfactuals that are 'good'. What makes the relevant counterfactuals 'good' is controversial, but goodness appears to include features like plausibility or actionability. The goal of applying **IC**, by contrast, is to identify counterfactuals that are 'good' in a different sense. Rather, than identifying counterfactuals that are plausible, the goal is to identify counterfactuals that yield genuine causal information.

We can see immediately that the standard counterfactual approach has at least the capacity to generate genuine causal information. Whenever an application

of the standard counterfactual approach yields a list of counterfactuals in which there is at least one counterfactual where only a single ML-variable has been altered, a cause has been correctly identified, at least according to the simplified picture we are working with in this section (see §4.2 for the generalisation). That's because at least one highlighted ML-variable satisfies **IC**, since there is a possible value for the highlighted ML-variable that flips the model's output holding the values of all other ML-variables fixed at their actual values. The standard counterfactual approach also has the capacity to generate complete causal information. For it can happen that the list of counterfactuals retrieved highlights all and only the input ML-variables that satisfy **IC**. Nothing precludes this from happening. So, for example, consider again Sara's loan example, and Table 1.

In this case, the standard counterfactual approach has successfully found three causal ML-variables: income, savings and age. That's because each ML-variable satisfies **IC**. Thus, when we infer that income, savings and age are all causal, we have gained partial causal information. We can also suppose, for the sake of argument, that there are no other ML-variables that satisfy **IC** for this case. Thus, when we infer that income, savings and age are causal, we have thereby gained complete causal information. In this way, the standard counterfactual approach has the capacity to pass the test for causal certification.

However, while having the *capacity* to pass the test for causal certification is a good start, what we really want to know is whether the standard counterfactual approach *in fact* passes the test for causal certification. Recall that basic causal certification is the guarantee that an approach to XAI always yields genuine causal information; whereas complete causal certification is a guarantee that an approach to XAI always provides complete causal information. In order for the standard counterfactual approach to pass causal certification, the application of **IC** must deliver both guarantees.

Unfortunately, no guarantee can be provided that the standard counterfactual approach always supplies complete causal information. Thus, the standard counterfactual approach does not pass the test for complete causal certification. Here, surprisingly, the proximity constraint is enough to generate a problem. Under the proximity constraint, the list of counterfactuals that is returned will feature only those counterfactuals that involve the smallest changes to input ML-variables that flip a model's output. Because of this, the proximity constraint always runs the risk of leaving some ML-variable out that in fact satisfies **IC**.

To see this, suppose we have a machine learning model $M$ with input ML-variables $X_1 ... X_n$, output $\psi$ and just three data points: $p_@$, $p_1$ and $p_2$. The point $p_@$ is the actual data point, whereas $p_1$ and $p_2$ are counterfactual data points. Suppose that the data points are ordered with respect to closeness as follows: $p_1$ is closer to $p_@$ than $p_2$. Suppose also that both $p_1$ and $p_2$ flip the output of the model from $\psi$ to some desired outcome $\phi$. In the case of $p_1$, this is due to a change in just one input ML-variable $X_1$, in the case of $p_2$ this is due to a change in just one input ML-variable as well, $X_2$, where $X_1 \neq X_2$. Under the proximity constraint, only one of these data points will be retrieved in a list of counterfactuals. Because $p_1$ is closer than $p_2$, only $p_1$ will end up in the final list. By **IC**, however, both $X_1$ and $X_2$ are causal, since in both cases there is a possible value for each ML-variable

that flips the model's output (holding all other ML-variables fixed). So a causal factor has been left out.

Note the point is not that an application of the standard counterfactual approach under the proximity constraint will always leave out some causal ML-variable. A given application might in fact capture all of the causal ML-variables. The point, rather, is that the proximity constraint introduces an ever-present possibility of mismatch between the highlighted ML-variables in a list of counterfactuals and the ML-variables that in fact satisfy **IC**. That's because, under the proximity constraint, the standard counterfactual approach will only ever retrieve counterfactuals involving the smallest changes to input ML-variables. It is always possible, however, that the ML-variables that are not highlighted in such a list nevertheless flip the model's output at more extreme values and thereby satisfy **IC**. The proximity constraint will generally ignore these extreme values and so there's simply no guarantee that the list of counterfactuals returned will include an exhaustive catalogue of the ML-variables that are causal.

On certain ways of implementing the standard counterfactual approach, the issue is quite stark. Consider the *plausibility* and *feasibility* constraints, introduced briefly in §2. These constraints prevent the counterfactual approach from returning lists of counterfactuals in which protected attributes are changed. Thus, not only is there no guarantee that all causal ML-variables will be captured under these constraints, there is something close to a guarantee in the opposite direction: when ML-variables that correspond to protected attributes in fact satisfy **IC**, they will be excluded.

The problem is by no means isolated to plausibility and feasibility constraints, however. As before, proximity alone is enough to generate a problem. To see this, suppose that Alex applies for a loan. He is asked to provide information including: income, age, assets, address, marital status, debt level and, this time, *race*. The bank takes this information and feeds it into a machine learning model. The model returns a result, which the bank uses to disqualify Alex for the loan. Now, Alex suspects that the race ML-variable caused the result in his case. If he is right, then he has a basis to contest the bank's decision.

But suppose that the bank is of a nefarious bent, and that they intentionally build a distance measure over data points that always places data points that flip the model's output in which the race ML-variable is changed further away than data points that flip the model's output and in which some other ML-variable is changed. Then changes to race will always be considered more extreme than changes to income, savings or debt level. Thus, race will always be excluded from the list of counterfactuals that is retrieved simply because of the proximity constraint and a poor choice of metric. More generally, because, as noted, under the proximity constraint there is no guarantee that full causal information has been provided, the way is always left open for an ML-variable that is causal and thus important for contesting decisions to not be correctly identified.

### 4.2 Lifting the Simplification

So far, the application of **IC** to the standard counterfactual approach has shown that such an approach does not pass complete causal certification. Before considering a

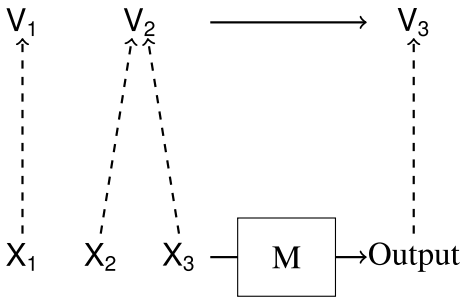


**Fig. 2** A complex set-up for IC for set of PW-variables $V$, $\{V_1, V_2, V_3\}$ and set of ML-variables $X$, $\{X_1, X_2, X_3\}$ for model $M$. An intervention on $V_2$ changes the values of both $X_2$ and $X_3$ at once

way to address this limitation, it is important to first revisit the simplifying assumption that I made when preparing the way for **IC**. I assumed that PW-variables correspond to exactly one ML-variable. By dropping this assumption we allow that PW-variables correspond to multiple ML-variables. This turns PW-variables into vectors of ML-variable information. The possible values for each PW-variable are thus the possible combinations of values for the ML-variables that a single PW-variable represents.

As noted, causation in **IC** is relative to a set of PW-variables. So far I have just been looking at one natural set of PW-variables, namely the set of PW-variables delivered by the simplifying assumption. Once we drop that assumption, we need to consider multiple different sets of PW-variables. In each set, there are PW-variables corresponding to $n$ ML-variables such that every ML-variable corresponds to some PW-variable. In each set there is also a PW-variable that corresponds to the output of the model. Figure 2 depicts this set up for the application of **IC**.

Note that each ML-variable can only correspond to exactly one PW-variable. That's because, if we allow that the same ML-variable corresponds to multiple PW-variables, then we can't apply **IC**. **IC** requires that we intervene on one PW-variable while holding other variables fixed. If the same ML-variable corresponds to two different PW-variables, $V_1$ and $V_2$, then we won't be able to intervene on either variable while holding the other fixed. For instance, suppose that $V_1$ corresponds to ML-variables $X_1$ and $X_2$; whereas $V_2$ corresponds to ML-variables $X_2$ and $X_3$. If we intervene on $V_1$ to change the values of $X_1$ and $X_2$ together, then we can't hold $V_2$ fixed, because $V_1$ corresponds to $X_2$.

Once we move to this more general setting, it becomes clear that the standard counterfactual approach always yields genuine causal information. An application of the standard counterfactual approach will reliably return at least one counterfactual in which there is a possible setting for a group of ML-variables that flips a model's output. With respect to this group of ML-variables, there will always be some PW-variable set in which there is a PW-variable, $V_n$ that represents the group of ML-variables at issue. Moreover, it will always be the case that an intervention on $V_n$ (while holding all other PW-variables fixed) will change the PW-variable that represents the model's output within this variable set. This is so even if the number of ML-variables that we have to change is quite high. Even if we must alter 20+ ML-variables, there will be a set of PW-variables in which those factors are grouped under a single PW-variable. We can thus capture the way in which changes to a large number of

**Table 2** Multiple ML-variables are changed in each counterfactual

| Data point | Age | Income | Savings | Suburb | Marital | Debt | Output |
|---|---|---|---|---|---|---|---|
| @ | 25 | 60k | 20k | Fitzroy | single | 200k | $\psi$ |
| $p_1$ | | | + 4k | | | -20k | $\phi$ |
| $p_2$ | | + 4k | | | Married | | $\phi$ |
| $p_3$ | − 6 | | | Carlton | | | $\phi$ |

variables flip a model's output, by intervening on the PW-variable that corresponds to the group of ML-variables at issue. The group of ML-variables represented by a single PW-variable will thus, jointly, be a cause since the PW-variable at issue satisfies **IC** for a given PW-variable set.

The standard counterfactual approach is thus guaranteed to reveal at least one cause constituted by at least one ML-variable according to at least one set of PW-variables that correspond to ML-variables. As a result, it passes basic causal certification. So, for instance, instead of Table 1, suppose that the standard counterfactual approach delivers Table 2 below:

We can see that the standard counterfactual approach still yields causal information. That's because there is a set of PW-variables that correspond to ML-variables whereby interventions on the PW-variables satisfies **IC**. For instance, consider the PW-variable set containing three PW-variables, $V_1$, $V_2$ and $V_3$ that are specified as follows: $V_1$ corresponds to income and marital status; $V_2$ corresponds to savings and debt level; $V_3$ corresponds to age and address. Then what Table 2 tells us is that if we alter $V_1$, while holding all other PW-variables fixed, the PW-variable corresponding to the model's output will change. Thus, $V_1$ satisfies **IC**. We can thus correctly infer that income and marital status together are causal. One way to capture this idea is by introducing the notion of a part of a cause: $x$ is part of a cause of $y$ when $x$ on its own is not a cause of $y$, but $x$ in conjunction with factors $z_1 ... z_n$ is a cause of $y$. Then we can say that the ML-variables for income and marital status are both parts of a cause of the model's output.

The point quickly generalises. Once we have lifted the simplifying assumption, there will always be some variable set under which at least one counterfactual yielded by the standard counterfactual approach provides information about genuine causal factors. Indeed, this is true even when the only counterfactuals that are returned contain a large number of feature differences. For there will always be a model whereby we can set a single PW-variable to the group of ML-variables that differ, no matter how large that group is. Thus, what the application of **IC** in this more general setting reveals is that the standard counterfactual approach always delivers causal information, and so passes basic causal certification.

Unfortunately, this more general setting does not help the standard counterfactual approach to achieve complete causal certification, and so this limitation remains. Indeed, if anything, the problem is amplified. For there are many more ways for ML-variables to be causal, since they need not be causal individually but in concert with other variables. If, however, there are many more ways for the input ML-variables

to be causal, then there is a much larger body of causal information that a list of counterfactuals might omit. Once again, this matters. Consider again the case of Alex's loan. Suppose that race on its own is not, in fact, causal: it does not satisfy **IC** for a set of PW-variables in which exactly one ML-variable corresponds to each PW-variable. It could still be the case that changing race and income together flips the model's output. Thus, the race ML-variable may be part of a cause, despite not being a cause on its own. Accordingly, not only does Alex need a guarantee that no information about ML-variables being individual causes has been left out, he needs a guarantee that no information about ML-variables being parts of causes has been left out. This is a more demanding requirement and one not easily met.

In a moment, I will consider how to meet this requirement. However, before pressing on it is worth considering two potential difficulties with the application of the Pearl-Woodward framework. First, one might worry that complete causal certification is unattainable. In order to achieve complete causal certification, we must potentially provide users with information about a large number of causal factors, particularly once causation involving multiple ML-variables working in concert is taken into account. But this information, one might argue, is too expansive to be cognitively manageable. As I suggest below, however, we can synthesise this information into a list. In this way, the cognitive burden of comprehending all of the causal factors can be outsourced to a searchable database, one that can be integrated with a toolkit and user interface. Such an interface would allow a user to search for causal factors that may be important to their particular situation, without having to take in the entire catalogue of causal factors. In this way, human understanding of causal factors can be scaffolded.

The second difficulty relates more directly to the Pearl-Woodward framework. Within that framework, all of the variables in the PW-variable set are assumed to be statistically independent of one another. But, one might argue, this is not realistic when we consider machine learning models. That's because, in some cases, it may be that there are interactions or dependencies between multiple ML-variables. The statistical independence of the PW-variables would thus seem to force the independence of the ML-variables they represent. In such a situation, one might argue, the Pearl-Woodward framework distorts the real structure of the machine learning model, which undermines the legitimacy of using **IC** as a test for causal certification.

There are two things to say here. First, the difficulties posed by the independence constraint can be avoided by shifting between different variable sets. For instance, suppose that two ML-variables interact with one another, and so can't be independently manipulated. Rather than using a simple model of the kind discussed in §4.1, whereby each PW-variable corresponds to just one ML-variable, we can employ a model in which PW-variables are allowed to correspond to groups of ML-variables (as in this section). By corresponding to a group of ML-variables, the ML-variables grouped are not incorrectly represented as independent of one another. Indeed, they are represented as in some sense dependent by virtue of falling under a single PW-variable. It is only really in the simplified case in §4.1 that ML-variables are all represented to be statistically independent. In the more complex setting described in this section, the PW-variables don't unrealistically represent all ML-variables to be statistically independent.

Second, even for a simple model in which each PW-variable corresponds to exactly one ML-variable, it is possible to intervene on multiple PW-variables together, thereby altering the underlying ML-variables in a way that does not violate any dependence. To do this we can employ a 'fat-handed' intervention, which involves manipulating multiple PW-variables in a single intervention (Baumgartner & Gebharter (2016); Scheines (2005)). One potential drawback of using fat-handed interventions, however, is that Woodward's definition of causation does not clearly allow for causation to be identified through such interventions. When applying **IC**, we should thus use the option for handling dependence between ML-variables specified in the previous paragraph.

While the dependencies between ML-variables can be handled within the Pearl-Woodward framework, it is worth emphasising that applying that framework is a sensitive matter. For what one must do is try to apply it only in a way that respects the dependencies between ML-variables. This potentially reduces the number of different ways of grouping ML-variables under PW-variables that should be considered when trying to identify causal factors. What happens if none of the ML-variables are statistically independent? In this situation, we are limited to using a single variable set, in which we have a single PW-variable that represents all of the ML-variables *en masse*. We are then forced to intervene on this one PW-variable. Even in this case, however, we can still apply **IC**. When an intervention on a single PW-variable flips a model's output and changes all of the ML-variables, then all of these variables are revealed as partial causes. Thus, the application of **IC** still works to identify causal factors even in this hypothetical situation, so long as we allow that ML-variables can be parts of causes.

## 5 A Hybrid Approach

The application of **IC** yields two main results. First, the standard counterfactual approach passes the **IC** test for basic causal certification. That's because it reliably generates genuine causal information about causes constituted by at least one ML-variable. One important consequence of this is that users can be given a guarantee that the information provided by the standard counterfactual approach to XAI is genuine causal information. This is potentially important, as it can help to answer user queries about whether the information they've been given is genuinely causal, which has the potential to build trust in the explanatory information being provided. Second, the standard counterfactual approach does not pass the **IC** test for complete causal certification, mainly because of the proximity constraint. This is a problem in those cases where a user needs to know whether or not any causal factors have been left out.

In this section, I propose a way to overcome this limitation by supplementing the standard counterfactual approach with the Pearl-Woodward framework. Note that this is a distinct application of the Pearl-Woodward framework from the use of **IC** discussed so far, and is logically independent. Thus, even if one is not attracted to the proposal I sketch below, one can still use the methodology outlined above as a test for causal certification. At any rate, the idea is to adopt a two-stage approach.

In the first stage, the Pearl-Woodward framework is used to derive a complete list of causal information. In the second stage, the standard counterfactual approach is used to reveal a list of counterfactuals that satisfy other constraints like plausibility, feasibility, diversity, proximity, sparsity and so on. The user is then provided with both lists.

In order to implement the first stage of this approach, we need to apply **IC** to reveal all of the causal information about why a given machine learning model has output $\psi$ rather than $\phi$. We do this using the method already described for applying **IC** to machine learning models. The only difference is that we now do it in a much more comprehensive manner. We first define sets of PW-variables for representing a given machine learning model. Within each such set, either exactly one ML-variable corresponds to a PW-variable, or $n$ ML-variables corresponds to a PW-variable. Moreover, every ML-variable corresponds to some PW-variable, and no ML-variable corresponds to more than one PW-variable. As before, the set of PW-variables also includes a PW-variable representing the machine learning model's output, with actual value of $\psi$ and possible value of $\phi$. Possible values of the PW-variables are given by the possible values of the ML-variables and combinations thereof in the manner already described. Actual values of the PW-variables are set by the actual values of the ML-variables, again in the manner already described.

Using these sets, we apply **IC** to every PW-variable that corresponds to at least one ML-variable. We do this by intervening on each PW-variable in turn for a given PW-variable set, holding all others fixed, to see if the PW-variable that represents the output of the model $M$ changes. We do this for each PW-variable until we either find a possible value at which the PW-variable representing the model output changes, or we exhaust all possible values for a PW-variable. If we find a possible value for a PW-variable that changes the PW-variable corresponding to the model's output, we record the associated ML-variables and then move on to the next PW-variable. We do this until every PW-variable in every PW-variable set has been checked.

In the last stage, we use the intervention information to identify ML-variables that are causal. We start by considering the set of PW-variables in which exactly one ML-variable corresponds to a PW-variable. When interventions on a PW-variable changes the PW-variable corresponding to the model's output (while holding all other PW-variables fixed), we record the associated ML-variables as *individual* causes of the model's output. We then consider the sets of PW-variables that group multiple ML-variables together. When interventions on PW-variables for these PW-variable sets change the PW-variable corresponding to the model's output (while holding all other PW-variables fixed), we record the associated ML-variables as *parts* of causes.[8]

In this way, we reveal the causal sensitivity of the model's actual output to each of its input ML-variables individually, as well as in concert with one another. The output of this process is likely to be a large amount of information. We thus need to synthesise it in a form that can be delivered to a user. We can do this by grouping the

[8] As above, we should respect dependencies between ML-variables. This may mean excluding the case in which each PW-variable corresponds to just one ML-variable. I have included that case for completeness, since it can be relevant.

information as follows. First, we outline the ML-variables that are individual causes (if there are any). Then we outline the ML-variables that are partial causes, along with the complex causes of which they are parts. Finally, we offer a guarantee that no causes have been missed. Below is a very simple example of a synthesised list of this kind for a machine learning model $M$ with six input variables: $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ and $X_6$, and a specific output $\psi$ (where a user wants to know why $\psi$ and not $\phi$ was the output):

1. $X_1$ was a cause on its own.
2. $X_2$ and $X_3$ were each parts of a single cause.
3. $X_4$, $X_5$ and $X_6$ were each parts of a single cause.
4. There were no other causes.

Importantly, such a list can be provided in the form of a searchable database with a user interface. This would then provide a cognitively manageable system for users to work with.

By providing users with two pieces of information, we can better achieve the three goals for XAI outlined by Wachter et al. (2018). The first list—the list of causes—provides a user with complete causal information. This list ensures that the necessary condition on providing explanatory understanding discussed in §2 is met, and so helps to satisfy the first goal. The first list also provides a basis for the second goal, namely to contest decisions, since it will provide a user with a full picture of what caused a model to deliver a certain output in their case. This, in turn, helps them to determine whether and how to contest a decision. By contrast, the second list helps to satisfy the third goal of XAI: that of helping users to work out what they should do differently to receive a better outcome. That's because the counterfactuals produced will satisfy constraints like plausibility or actionability which are geared toward achieving good outcomes for users.

In essence, this two-stage approach is a divide-and-conquer method for addressing the three goals of XAI. One stage is aimed at causal understanding and contesting decisions; the other is aimed at providing practical advice. Why use such an approach? Why not just stop at the first stage, and use the Pearl-Woodward framework on its own? The answer is that this framework is not very useful for providing practical guidance. It excels at identifying those ML-variables that cause models to have specific outputs. However, it may be that for many such ML-variables it is only at *extreme values* that a model's output is affected. So, for instance, it might turn out that in Sara's loan case, there is a possible value for the debt variable that flips the model's output. However, that possible value may be a debt level of 0. This would require of Sara that she pay off 200k of debt, which may not be viable for

her. Similarly, note that in Table 1 there is a possible value for the age variable that flips the model's output (namely age backwards by 6 years). However, changing that value is not possible for Sara.

The reason, then, why we need to use the two approaches in concert is that they have different strengths. The standard counterfactual approach to XAI is not guaranteed to deliver full causal information, but it is built to be action-guiding. Being action-guiding or actionable has been shown to be important in user's judgements of how satisfying an explanation is, which suggests that this kind of information is important to provide (Kirfel & Liefgreen (2021)). The Pearl-Woodward framework is not guaranteed to be action-guiding, but it is built to identify full causal information. Together, the two approaches constitute a more complete approach to XAI than either approach on its own. Thus, what the two-stage strategy essentially recognises is that the second and third goals for XAI come apart to a certain extent. Identifying the proximal, plausible and feasible changes that one can make to receive a better outcome sometimes requires setting aside certain causal factors, since changing those may not be causal options. In this way, satisfying the third goal can sometimes sit in tension with satisfying the second goal, which makes it awkward for a single approach to meet both goals.

In practice, what this means is that we will need to run two sets of computations. We will need to compute causes using **IC**, and we will need to compute counterfactuals that give users practical advice about what to do differently in the future. Evidently, there has been a great deal of work on finding efficient strategies for computing counterfactuals that support practical aims. Less attention has been paid to identifying strategies for efficiently computing full causal information. Clearly, such strategies are needed. For we must compute a large number of possible permutations of input data to find possible values for ML-variables that flip a model's output. Whether this is even computationally viable would need to be shown. The proposed two-stage strategy thus opens up a new line of research, whereby computational methods for finding causes in the case of machine learning models are developed and tested for efficiency.

Before wrapping up, it is worth noting two things. First, as Rawal & Lakkaraju (2020) emphasise, the standard counterfactual approach focuses on providing local explanations: explanations of specific outcomes. It is, however, important to also provide a global analysis of a machine learning model to understand its behaviour across a range of instances. This is important for understanding model behaviour before a system has been deployed. Understanding causal factors is also important to understanding this global behaviour, since it can be helpful to know whether, say, race is a causal factor in any instance within a loan model, not just for a particular case. The Pearl-Woodward framework can be generalised to provide complete causal certification across all instances. This can be achieved by allowing the PW-variable that corresponds to the output to take a range of values. The Pearl-Woodward framework can thus provide a comprehensive check on whether variables like race are causal, prior to the implementation of a machine learning model.

Second, I have framed the approach as one that combines the standard counterfactual approach with the Pearl-Woodward framework. However, there are reasons to doubt the capacity of the standard counterfactual approach to deliver an appropriate

list of counterfactuals. This, again, is largely to do with the proximity constraint. Recent psychological work has shown that people don't necessarily find 'smallest change' counterfactuals of the kind delivered by the proximity constraint to be the most useful. Instead, users tend to focus on prototypical counterfactual cases (Delaney et al. (2022)). More generally, as Keane et al. (2021) argue, there is a dearth of experimental work testing particular modelling choices within the standard counterfactual approach (though this is changing rapidly, see e.g., Celar & Byrne (2023); Ford & Keane (2022); Förster et al. (2020a, 2020b); Kirfel & Liefgreen (2021); van der Waa et al. (2021); Warren et al. (2022)).

Taken together, the apparent failure of the standard counterfactual approach to pass full causal certification, coupled with the potential disconnect with user preferences about explanation, may lead one to reject the standard counterfactual approach altogether. This is one potential moral to be drawn from the discussion here. Note, however, that there remains a need to provide explanations to users that are action-guiding, and so, even if the relevant moral is drawn, there is still a need to replace the standard counterfactual approach with something else (perhaps something that emphasises prototypes, rather than minimal edits (Kim et al. (2016); Li et al. (2018); Van Looveren & Klaise (2021)). Whatever this replacement approach might be, however, the Pearl-Woodward framework remains useful in the ways discussed here. On the one hand, it can be used to test whether any alternative to the standard counterfactual approach passes causal certification. On the other hand, it can be used as a supplement to any approach that fails to achieve either basic or complete causal certification. What the Pearl-Woodward framework provides, then, is a way to ensure that users are being provided with the right information to support genuine explanatory understanding, which is important for realising the goals of XAI.

## 6 Conclusion

It is time to take stock. In this paper, I have situated the standard counterfactual approach within the Pearl-Woodward framework for specifying causes. The main findings of the paper can be summarised as follows:

- The Pearl-Woodward framework can be used for basic and complete causal certification of existing approaches to XAI.
- The standard counterfactual approach to XAI passes the test for basic causal certification, and so a guarantee that the approach supports genuine explanatory understanding can be delivered to users.
- The standard counterfactual approach fails the test for complete causal certification.
- By using the Pearl-Woodward framework in concert with the standard counterfactual approach to XAI, the three goals for XAI can be met.
- Future research should explore the efficient computation of Woodward causes for machine learning models.

**Acknowledgements** Sam Baron would like to thank Kate Lynch and Maureen O'Malley, as well as the audience at the Leeds-Dianoia Rome Seminar on normativity in 2022, for very helpful discussion of the ideas surrounding this paper.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.



## References

Amir-Hossein, K., Schölkopf, B., & Valera, I. (2021), Algorithmic recourse: From counterfactual explanations to interventions. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 353–362.

Asher, N., De Lara, L., Paul, S., & Russell, C. (2022). Counterfactual models for fair and adequate explanations. *Machine Learning and Knowledge Extraction, 4*, 319–349.

Baron, S., Colyvan, M., & Ripley, D. (2017). How mathematics can make a difference. *Philosophers' Imprint, 17*, 1–19.

Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability and fat-handedness. *British Journal for the Philosophy of Science, 67*, 731–756.

Beckers, S. (2022). Causal explanations and xai. *Proceedings of Machine Learning Research, 140*, 1–20.

Been, K., Khanna, R., & Koyejo, O. (2016), Examples are not enough, learn to criticize! criticism for interpretability. Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016).

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines, 32*, 563–584.

Byrne Ruth M. J. (2019), Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 6276–6282.

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., & Holzinger, A. (2023). Quod erat demonstrandum? Towards a typology of the concept of explanation for the design of explainable ai. *Expert Systems with Applications, 213*, 118888.

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms and applications. *Information Fusion, 81*, 59–83.

Chris, R., Mc Grath, R., & Costabello, L. (2020), Learning relevant explanations. 2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020).

Christopher, M. (2020), Interpretable Machine Learning. lulu.com.

Courtney, F., & Keane, M. T. (2022), Explaining classifications to non-experts: An xai user study of post-hoc explanations for a classifier when people lack expertise. https://arxiv.org/abs/2212.09342.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. Parallel problem solving from nature. (pp. 448–469). XVII

de Oliveira, R. M. B., & Martens, D. (2021). A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences, 11*, 7274.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in Neural Information Processing Systems, 31*, 592–603.

Divyat, M., Tan, C., & Sharma, A. (2019), Preserving causal constraints in counterfactual explanations for machine learning classifiers. CausalML: Machine Learning and Causal Inference for Improved Decision Making Workshop, 33rd Conference on Neural PRocessing Systems (NeurIPS2019), https://arxiv.org/abs/1912.03277.

Dúran, J. M. (2021). Dissecting scientific explanation in ai (sxai): A case for medicine and healthcare. *Artificial Intelligence, 297*, 103498.

Dúran, J. M., & Formanek, N. (2018). Grouds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*, 645–666.

Eoin, D., Pakrashi, A., Greene, D., & Keane, M. T. (2022), Counterfactual explanations for misclassified images: How human and machine explanations differ. https://arxiv.org/abs/2212.08733.

Fine, K. (1975). Review of "counterfactuals''. *Mind, 84*, 451–458.

Greta, W., Keane, M. T., & Byrne, R. M. J. (2022), Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai. https://arxiv.org/abs/2204.10152.

Harbecke, J. (2021). Counterfactual theories of causation and the problem of large causes. *Philosophical Studies, 178*, 1647–1668.

Hills, A. (2016). Understanding why. *Noûs, 50*, 661–688.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*, e1312.

Kacper, S., & Flach, P. (2019), Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. SafeAI@ AAAI (2019).

Kaivalya, R., & Lakkaraju, H. (2020), Beyond individualized recorse: Interpretable and interactive summaries of actionable resources. Proceedings of the 34th Conference on Neural Information Processing Systems, 1–12.

Kasirzadeh, A., & Smart, A. (2021), The use and misue of counterfactuals in ethical machine learning. FAccT '21: Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency, 228–236.

Keane Mark T., Kenny, E. M., Delaney, E., & Smyth, B. (2021), If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), https://arxiv.org/abs/2103.01035.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine, 23*, 89–109.

Lara, K., & Liefgreen, A. (2021), What if (and how...)? - actionability shapes people's perceptions of counterfactual explanations in automated decision-making. ICML (International Conference on Machine Learning) Workshop on Algorithmic Recourse, 1–5.

Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M. (2018). Comparison-based inverse classification for interpretability in machine learning. In Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. IPMU 2018. *Communications in Computer and Information Science., 853*, 100–111.

Lawler, I. (2019). Understanding why, knowing why, and cognitive achievements. *Synthese, 196*, 4583–4603.

Lenart, C., & Byrne, RMJ. (2023), How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. Memory & Cognition.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*, 556–567.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*, 455–476.

Mark, K., & Smyth, B. (2020), Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). International Conference on Case-Based Reasoning, 163–178.

Maximilian, F., Klier, M., Kluge, K., & Sigler, I. (2020a), Evaluating explainable artificial intelligence: What users really appreciate. Twenty-Eigth European Conference on Information Systems (ECIS2020).

Maximilian, F., Klier, M., Kluge, K., & Sigler, I. (2020b), Fostering human agency: A process for the design of user-centric xai systems. ICIS 2020 Proceedings.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
Mothilal, R. K., Sharma, A., & Tan, C. (2020), Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617.
Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: a context sensitive framework for explainable medical ai. *Ethics and Information Technology, 24*, 1–15.
Oscar, L., Liu, H., Chen, C., & Rudin, C. (2018), Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 3530–3537.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
Pritchard, D. (2014). Knowledge and understanding. In A. Fairweather (Ed.), *Epistemology naturalized.* Synthese Library.
Rafael, P., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020), Face: Feasible and actionable counterfactual explanations. AIES'20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 344–359.
Räz, T., & Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y
Reutlinger, A. (2016). Is there a monist theory of causal and non-causal explanations? The counterfactual theory of scientific explanation. *Philosophy of Science, 83*, 733–745.
Rory, M., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018), Interpretable credit application predictions with counterfactual explanations. NIPS 2018 Workshop on challenges and opportunities for AI in financial services: The impact of fairness, explainability, accuracy, and privacy.
Sahil, V., Dickerson, J., & Hines, K. (2020), Counterfactual explanations for machine learning: A review. https://arxiv.org/abs/2010.10596.
Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy, 9*, 165–181.
Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science, 72*, 927–940.
Schulz, K. (2011). If you'd wiggled a, then b would've changed: causality and counterfactual conditionals. *Synthese, 179*, 239–251.
Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies, 146*, 1–10.
Solon, B., Selbst, A. D., Raghavan, M (2020), The hidden assumptions behind counterfactual explanations and principal reasons. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 80–89.
Sullivan, E. (2018). Understanding: Not know-how. *Philosophical Studies, 175*, 221–240.
Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science, 73*, 109–133.
Thibault, L., Lesot, M-J., Marsala, C., Renard, X., & Detyniecki, M. (2019), The dangers of post-hoc interpretability: Unjustified counterfactual explanations. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, ICJAI-19, 2801–2807.
Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? a comparison of statistical, machine learning and data mining predictive models. *Statistics in Society A, 176*, 565–584.
van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404.
van Looveren, A., & Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. Lecture Notes in Computer Science. In: Oliver, N, Pérez-Cruz, F., Kramer, S., Read, J., & Lozano, J. A. (Eds.), *Machine learning and knowledge discovery in databases. Research track. EXCML PKDD 2021*. Springer, Cham (pp. 650–665)
Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology, 31*, 841–887.
Watson, D. S., & Floridi, L. (2021). The explanation game: a formal framework for interpretble machine learning. *Synthese, 198*, 9211–9242.
Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese, 191*, 3367–3391.
Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies, 176*, 2807–2831.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.