The Knowledge Engineering Review, Vol. 36, e14, 1 of 24. © The Author(s), 2021. Published by Cambridge University Press
doi:10.1017/S0269888921000102

Contrastive explanation: a structural-model approach

TIM MILLER

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
E-mail: tmiller@unimelb.edu.au

Abstract

This paper presents a model of contrastive explanation using structural casual models. The topic of causal explanation in artificial intelligence has gathered interest in recent years as researchers and practitioners aim to increase trust and understanding of intelligent decision-making. While different sub-fields of artificial intelligence have looked into this problem with a sub-field-specific view, there are few models that aim to capture explanation more generally. One general model is based on *structural causal models*. It defines an explanation as a fact that, if found to be true, would constitute an actual cause of a specific event. However, research in philosophy and social sciences shows that explanations are *contrastive*: that is, when people ask for an explanation of an event—the *fact*—they (sometimes implicitly) are asking for an explanation relative to some *contrast case*; that is, 'Why *P* rather than *Q*?'. In this paper, we extend the structural causal model approach to define two complementary notions of *contrastive explanation*, and demonstrate them on two classical problems in artificial intelligence: classification and planning. We believe that this model can help researchers in subfields of artificial intelligence to better understand contrastive explanation.

1 Introduction

> The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case—Hilton (1990, p. 67).

The recent explosion in research and application of artificial intelligence has seen a resurgence of *explainable artificial intelligence* (XAI)—a body of work that dates back over three decades; for example, see Buchanan & Shortliffe (1984); Chandrasekaran *et al*. (1989); Swartout & Moore (1993). This resurgence is driven by lack of trust from users (Stubbs *et al*.,, 2007; Linegang *et al*.,, 2006; Mercado *et al*.,, 2016), and also concerns regarding the ethical and societal implications of decisions made by 'black box' algorithms (Angwin *et al*.,, 2016).

One key mode of XAI is explanation. An explanation is a justification or reason for a belief or action. There has been a recent burst of research on explanation in artificial intelligence, particularly in machine learning. Much of the work in XAI has centered around extracting the causes (or main causes) of a decision or action. While finding causes is an important part of explanation, people do so much more when explaining complex events to each other, and we can learn much from considering how people generate, select, present, and evaluate explanations.

Miller (2018) systematically surveyed over 250 papers in philosophy, psychology, and cognitive science on how people explain to each other, and noted perhaps the most important finding is that explanations are *contrastive*. That is, people do not ask 'Why *P*?'; they ask 'Why *P* rather than *Q*?',

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

although often Q is implicit from the context. Following Lipton (1990), we will refer to P as the *fact* and Q as the *contrast case*.

Researchers in social science argue that contrastive explanation is important for two reasons. First, people ask contrastive *questions* when they are surprised by an event and expected something different. The contrast case identifies what they expected to happen (Hilton, 1990; Van Bouwel & Weber, 2002; Lipton, 1990; Chin-Parker & Cantelon, 2017). This provides a 'window' into the questioner's mental model, identifying what they do not know (Lewis, 1986). Second, giving contrastive *explanations* is simpler, more feasible, and cognitively less demanding to both questioner and explainer (Lewis, 1986; Lipton, 1990; Ylikoski, 2007). Lewis argues that a contrastive question 'requests information about the features that differentiate the actual causal history from its counterfactual alternative.' (Lewis, 1986, p. 231).

Lipton (1990) defines the answer to a contrastive question as the *difference condition*:

> To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q.—Lipton (1990, p. 256).

Following this, the explainer does not need to reason about *or even know about* all causes of the fact—only those relative to the contrast case.

As an example, consider an algorithm that classifies images of animals. Presented with an image of a crow, the algorithm correctly identifies this as a crow. When asked for a reason, a good *attribution* would highlight features corresponding the crow: its beak, feathers, wings, feet, and color—those properties that correspond to the model of a crow. However, if the question is: 'Why did you classify this as a crow instead of a magpie?', the questioner already identifies the image as a bird. The attribution that refers to the beak, feathers, wings, and feet makes a poor explanation, as a magpie also has these features. Instead, a good explanation would point to what is different, such as the magpie's white coloring and larger wingspan.

Importantly, the explanation fits directly within the questioner's 'window' of uncertainty, and is smaller and simpler, even on this trivial example. AI models, though, are typically more complicated and more structured, implying that contrastive explanation can provide much benefit by adhering to Grice's conversational maxims (Grice, 1975) of *quantity*: make your contribution as informative as is required, and do not make it more informative than is required; and *relation*: only provide information that is related to the conversation.

Further to this, Wachter *et al.* (2017) argue that explanations using contrastive explanations can be used as a means for explaining individual decisions, providing sufficient explanatory power for individuals to understand and contest decisions, without imposing requirements of opening up the mechanism for decision-making.

In this paper, we extend Halpern & Pearl's definition of explanation using *structural causal models* (Halpern & Pearl, 2005b) to the case of contrastive explanation, providing a general model of contrastive explanation based on Lipton's difference condition. In particular, we define contrastive explanation for two types of questions: *counterfactual* questions and *bi-factual* questions. A counterfactual question is of the form 'Why P rather than Q?', and asks why some fact P occurred instead of some hypothetical *foil* Q. A bi-factual question is of the form 'Why P but Q?', and asks why some fact P occurred in the current situation while some *surrogate* Q occurred in some *other* factual situation. The difference is that in the former, the foil is hypothetical, while in the latter, the surrogate is actual and we are contrasting two events that happened in different situations. From the perspective of artificial intelligence, the former is asking why a particular algorithm gave an output rather than some other output that the questioner expected, while the latter is asking why an algorithm gave a particular output this time but some (probably different) output another time.

We define what it means to have a cause of these two contrastive questions, and what it means to explain them. Although it is not possible to prove such a model is 'correct', we show that the model is internally consistent, and demonstrate it on two representative examples in artificial intelligence: classification and goal-directed planning.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

2 Related work

2.1 Philosophical foundations

In the social sciences, it is generally accepted that explanations are contrastive (Miller, 2018). The questions that people ask have a contrast case, which is sometimes implicit, and the explanations that people give explain relative to this contrast case. Even when giving an explanation with no question, people explain relative to contrast cases.

Garfinkel (1981) seems to be the first to make a case for contrastive explanation[1]. He provides a story about a well-known bank robber Willie Sutton who purportedly replied to journalist who asked why he robbed banks, with: 'That's where the money is.' Garfinkel argues that Sutton answered why he robs [banks vs. other things], rather than why he [robs vs. does not rob] banks because he answered to a different contrast case: that of banks vs. non-banks, rather than robbing vs. not robbing. Garfinkel notes that these two different contrasts create two different contexts, and that explanations are relative to these contrastive contexts. An object of explanation is not just a state of affairs, but a 'state of affairs together with a definite *space of alternatives* to it' (Garfinkel, 1981, p. 21).

At the same time, Van Fraassen (1980) was also arguing the case of contrastive explanations. He states that the underlying structure of a why-question is: 'Why (is it the case that) *P in contrast to* (other members of) *X*?', in which *P* is the *topic* and *X* is the *contrast class* to *P* (Van Fraassen, 1980, p. 127). An answer to such a question has the structure '*P in contrast to* (the rest of) *X* because *A*' (Van Fraassen, 1980, p. 143). Van Fraassen argues that when a questioner asks such a question, they presuppose that: (1) the topic *P* is true; (2) all other elements of the contrast class *X* are false; and (3) *A* is both true and explanatorily relevant to the topic. He proposes an explicit relation *R* that determines explanatory relevance.

Hesslow (1983); Hesslow (1988) extends this idea of explanatory relevance and seems to be the first to make a case for the idea of contrast cases themselves defining *explanatory relevance*. He argues that there is a distinction between *determining causes* and *explanatory causes*, with the former being the (often large) set of conditions that contribute to causing an event, and the latter being a subset of the determining causes that are selected due to their *explanatory power*. Hesslow's theory of explanation is based on two complementary ideas. The first is that of contrastive explanation. He states that:

> . . . the effect or the explanandum; i.e. the event to be explained, should be construed, not as an object's having a certain property, but as a *difference* between objects with regard to that property.—Hesslow (1988), p. 24

The second is of explanatory relevance. Hesslow argues that by explaining only those causes that are different between the two or more objects, the explanation is more relevant to the questioner as it provides those causes that the questioner does not know. In essence, the contrast case provides a window into the particular causes that the questioner does not understand.

Hesslow presents an example: 'Why did the barn catch on fire?'. The explanation that someone dropped a lit cigarette in the hay has strong explanatory power and would satisfy most people. But what about other causes? The presence of oxygen, the hay being dry, and absence of fire sprinklers are all causes, but the cigarette has particular explanatory power because oxygen is always present in barns, and most barns are dry and have no fire sprinklers. The explanation is *contrasting* to these normal cases.

He formalizes this notion as follows. Given an object a, a property E, and a reference class R (the contrast cases), the cause Ca is an *adequate explanation* of $\langle a, E, R\rangle$ iff:

1. for all x in R, if Cx had been true then Ex would have been true; and
2. if $\neg Ca$ had been true, then $\neg Ea$ would have been true,

in which Cx and Ex refer to the cause C and property E respectively applying to x. This states that Ca is an adequate explanation if and only if (1) the cause C held on all the other objects x in R (e.g. other

[1] Although Van Fraassen (1980), p. 127 attributes the idea of contrastive explanation to Bengt Hannson in an unpublished manuscript circulated in 1974.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

barns), then the property E would also hold (the other barns would have also caught fire); and (2) the cause C did not apply to a, then the property E would not hold. We can see that (1) does not apply to oxygen, because oxygen is present in other barns that do not catch fire, while for the cigarette this is the case; and that (2) applies to the cigarette—if the cigarette had not been dropped, the fire would not have occurred.

At a similar time, Lewis (1986) proposed a short account of contrastive explanation. According to Lewis, to explain why P occurred rather than Q, one should offer an event in the history of P that would not have applied to the history of Q, if Q had occurred. For example, he states: 'Why did I visit Melbourne in 1979, rather than Oxford or Uppsala or Wellington? Because Monash University invited me. That is part of the causal history of my visiting Melbourne; and if I had gone to one of the other places instead, presumably that would not have been part of the causal history of my going there' (Lewis, 1986, p. 229–230). This has parallels with Hesslow's account (Hesslow, 1983; 1988).

Temple (1988) subsequently argued against the case of contrastive explanation. Temple argued that the question 'Why P rather than Q?' presupposes that P is true and Q is not, and that the object of explanation is not to explain why P and Q are mutually exclusive, but instead to ask 'Why [P and not Q]?'. Therefore, contrastive why-questions are just standard propositional why-questions of the form 'Why X?', but with X being [P and not-Q].

However, Lipton (1990) argues that this is a language phenomenon, and semantically, explaining 'Why P rather than Q?' is not the same as explaining 'Why [P and not Q]?'. Building on Lewis's interpretation based on the history of events (Lewis, 1986), Lipton argues that answering 'Why [P and not Q]?' requires an explanation of P and of not-Q. For example, to answer why the barn burned down rather than not burning down would require a complete attribution of why the barn burned down, including the presence of oxygen, as well as why other barns do not typically burn down. Lipton argues that this is not what the explainee wants.

Lipton (1990) proposes that explanation selection is best described using the *difference condition*:

> To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q.—Lipton (1990, p. 256).

This differs from the definition of contrastive explanation from Lewis (1986) in that instead of selecting a cause of P that is not a cause of Q if Q had occurred, we should explain the *actual difference* between P and not-Q; that is, we should cite a cause that is in the actual history of P, and an event that did not occur in the actual history of not-Q.

We can formalize this as the following, in which $\rightsquigarrow$ is the causal relation, and H_P and H_{notQ} are the *history* of P and not-Q respectively, and H_Q is the hypothetical history of Q had it occurred:

Lewis $c \rightsquigarrow P \wedge c \not\rightsquigarrow Q$ where $c \in H_P \wedge c \notin H_Q$
Lipton $c \rightsquigarrow P \wedge c' \rightsquigarrow Q$ where $c \in H_P \wedge c' \notin H_{notQ}$

Thus, Lewis's definition (Lewis, 1986) cites some alternative history of facts in which Q occurred, whereas Lipton's definition (Lipton, 1990) refers to the *actual* history of not-Q. Further, Lewis's definition states that the explanation should be *an* event c (or perhaps set of events), whereas Lipton's states that the explanation is the difference between c and c'.

It was generally accepted at the time that Lipton (1990) proposed his ideas, that facts and contrast cases are incompatible events (Temple, 1988; Garfinkel, 1981; Van Fraassen, 1980; Ruben, 1987); for example, a barn cannot both burn down and not burn down, or leaves cannot be blue and yellow at the same time. However, Lipton notes that compatible contrastive cases are also valid. For example, we can ask why one leaf is blue while another leaf is yellow. It is perfectly possible that both leaves could be blue, but we are looking for explanations as to why only one of them is.

Ylikoski (2007) provides a more refined model to explain this, noting that incompatible vs. compatible contrast cases are two different types of question. The first is when we contrast two incompatible contrasts

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

of the same process: one the fact and one the 'imagined' *foil*, such as a leaf being yellow rather than that leaf being blue. The fact and the foil must be inconsistent. The second is when we contrast two facts from two actual and different processes. That is, both facts actually occurred, such as one yellow leaf and one blue leaf. Ylikoski calls the second fact a *surrogate* for a counterfactual claim about the first process. He claims that the surrogate is used to simplify the explanation—as one simply needs to find the difference between the fact and surrogate, which is consistent with the idea from Lipton (1990) that this is cognitively a simpler problem.

Van Bouwel and Weber (2002) divide explanatory questions into four types:

Plain fact: Why does object a have property P?
P-contrast: Why does object a have property P, rather than property Q?
O-contrast: Why does object a have property P, while object b has property Q?
T-contrast: Why does object a have property P at time t, but property Q at time t'?

This defines three types of contrast: within an object (P-contrast), between objects themselves (O-contrast), and within an object over time (T-contrast). P-contrast is the standard 'rather than' interpretation, while O-contrast and T-contrast correspond to Ylikoski's notion of different processes (Ylikoski, 2007).

In Section 4, we will formalize the notion of contrastive questions using the framework of Halpern and Pearl (2005a), and will show that the reasoning of Ylikoski (2007) is natural with respective to structural equations and fits the types of questions we would expect in XAI. The concept of *P-contrast* is captured as *counterfactual* explanations, while *O-contrast* and *T-contrast* are captured as *bi-factual* explanations.

2.2 *Computational approaches*

In artificial intelligence, contrastive questions are not just a matter of academic interest. User studies investigating the types of questions that people have for particular systems identify 'Why not?' questions and contrast classes as important. Lim and Dey (2009) showed that 'Why not?' questions are important in context-aware applications, while Haynes *et al*. (2009) found that users of their virtual aviation pilot system particularly sought information about contrast cases. Given that this is consistent with views from philosophy and psychology, it makes sense to consider the difference condition as key to answering these questions.

The idea of why-not questions in artificial intelligence was around prior to these studies. The explanation module of the MYCIN expert system explicitly allowed users to pose questions such as 'Why didn't you do X?' (Buchanan & Shortliffe, 1984), which is providing a foil for the fact. More recently, there has been a keen interest in answering why-not questions for many different sub-fields of artificial intelligence, including machine learning classification (Dhurandhar *et al*.,, 2018; Mothilal *et al*.,, 2020), belief-desire-intention agents (Winikoff, 2017), reinforcement learning (Madumal *et al*.,, 2020; Waa *et al*.,, 2018), classical planning (Krarup *et al*.,, 2019; Sreedharan *et al*.,, 2018), and image classification (Akula *et al*.,, 2020), to cite just a few papers.

However, while the idea of why-not questions is being addressed, there is not a clear and consistent understanding of contrastive explanation in the explainable AI community. First, while there many of papers what answer why-not questions, most of these are *counterfactual* solutions, not *contrastive*, meaning that they do not make explicit use of the *difference* between the fact and the foil, with some exceptions; for example, Madumal *et al*. (2020). This may seem trivial, and from an algorithmic perspective perhaps it is. But from the perspective of helping people to understand models and decisions, this is a crucial step that requires further research in human factors and human–computer interaction.

Second, there is no general understanding in the community about contrastive and counterfactual explanation, with much of it built on the intuitions of authors. This leads to algorithms that can be useful, but terminology and solutions that are not aligned. For example, Dhurandhar *et al*. (2018) use the term *pertinent negatives/positives* to refer to foils, while Akula *et al*. (2020) use the term *fault lines*, and Krarup *et al*. (2019) use *foil*.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

Third, while there is work on counterfactual explanation that contributes to contrastive explanation, there is no work that the author is aware of that addresses why we call *bi-factual* questions: why P this time but Q last time? As noted by Wang *et al*. (2019) in their study of clinicians using AI-driven diagnosis in intensive care units, some clinicians 'wanted to see past patients which had similar presentations (e.g., complaints, vital signs), but not necessarily similar diagnoses (decision outcomes)'. This is a request for a bi-factual contrastive explanation. Showing similar cases with different outcomes, and importantly, showing the *difference* between the two helps people to understand why the outcomes differ. Currently, this form of why-not question is lacking in explainable AI literature.

The aim of this paper is to mitigate some of the issues above by providing a *general* model of contrastive explanation that can be mapped to other models in artificial intelligence, such as machine learning, planning, reinforcement learning, case-based reasoning, and BDI agents. By presenting a single, coherent general model, we can begin to understand the similarities and differences between proposed solutions and can start to plug holes, such as lack of bi-factual explanation, in explainable AI.

As far as the author is aware, Kean (1998) is the only other author to consider a general computational model of contrastive explanation. Kean's model of contrastive explanation is also built on Lipton's difference condition (Lipton, 1990). Given a knowledge base $\mathcal{K}$ and an observation P, Kean proposes a simple model to calculate why P occurred instead of Q. Kean provides a definition of a *non-preclusive* contrastive explanation for '*Why P rather than Q?*', which refers to the propositions that are required for P to hold but not Q. The definition of a *preclusive* contrastive explanation uses the difference condition, and, as in this paper, identifies that the contrastive explanation must reference both the causes of P and causes of Q that were not true. There are three key differences between Kean's model and the structural approach model approach in this paper. First, Kean's model was published when the understanding of causality in artificial intelligence was in its infancy, and is therefore built on propositional logic, rather than on a logic of causality and counterfactuals, which is more suitable. Second, Kean's model considers only 'rather than' questions, and not contrastive explanations with surrogates rather than foils. Third, and most importantly Kean's model is in fact a model of abductive reasoning, in which assumptions are made about the truth of certain propositions to find the 'best' explanation. As such, this is a model of the cognitive process of contrastive explanation from the perspective of an explainer, and the task is to derive an explanation for an observation. There is no explainee in Kean's model. In contrast, our model is not concerned with abductive reasoning, but instead models an explainer with complete knowledge of the explanation using the difference condition to communicate to an unaware explainee.

3 Structural models

In this paper, we build definitions of contrastive questions and contrastive explanations based on Halpern & Pearl's *structural models* (Halpern & Pearl, 2005a). As opposed to previous models, which use logical implication or statistic relevance, Halpern and Pearl's definition is based on counterfactuals, modeled using *structural equations*.

In Part I (Halpern & Pearl, 2005a) of their paper, Halpern and Pearl provide a formal definition of causality. A *causal model* is defined on two sets of variables: *exogenous* variables, who values are determined by factors external to the model, and *endogenous* variables, who values are determined by relationships with other (exogenous or endogenous) variables.

3.1 Models

Formally, a *signature* $\mathcal{S}$ is a structure $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, in which $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ a set of endogenous variables, and $\mathcal{R}$ is a function that defines the range of values for every variable $Y \in \mathcal{U} \cup \mathcal{V}$; that is, the range of a variable Y is $\mathcal{R}(Y)$.

A *causal model* is a pair, $M = (\mathcal{S}, \mathcal{F})$, in which $\mathcal{F}$ defines a set of functions, one for each endogenous variable $X \in \mathcal{V}$, such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$ determines the value of X based on other variables in the model. A causal model is said to be *recursive* if it is acyclic.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

Table 1. A simple lay model for distinguishing common arthropods

				Compound	
Type	**No. Legs**	**Stinger**	**No. Eyes**	**Eyes**	**Wings**
Spider	8	×	8	×	0
Beetle	6	×	2	✓	2
Bee	6	✓	5	✓	4
Fly	6	×	5	✓	2

A *context*, $\vec{u}$, is a vector that gives a unique value to each exogenous variable $u \in \mathcal{U}$. A model/context pair $(M, \vec{u})$ is called a *situation*.

Halpern & Pearl (2005a) extend this basic structural equation model to support modeling of counterfactuals. To represent counterfactual models, the model $M_{\vec{X}\leftarrow\vec{x}}$ defines the new causal model given a vector $\vec{X}$ of endogenous variables in $\mathcal{V}$ and their values $\vec{x}$ over the new signature $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, R|_{\mathcal{V}-\vec{X}})$. This represents the model M with the values of $\vec{X}$ overridden by $\vec{x}$. Formally, this model is defined as $M_{\vec{X}\leftarrow\vec{x}} = (\mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}\leftarrow\vec{x}})$, in which each $F_Y^{\vec{X}\leftarrow\vec{x}}$ in $\mathcal{F}$ is defined by setting the values of $\vec{X}$ to $\vec{x}$ in function F_Y.

3.2 Language

To reason about these structures, in particular, counterfactuals, Halpern and Pearl (2005a) present a simple but powerful language. Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, variables $X \in \mathcal{V}$ and values $x \in \mathcal{R}(X)$, a formula of the form $X = x$ is called a *primitive event*, and describes the event in which variable X is given the value x. A *basic causal formula* is of the form $[Y_1 \leftarrow y_1, \ldots, Y_n \leftarrow y_n]\phi$, in which ϕ is any Boolean combination of primitive events, each Y_i is a variable in $\mathcal{V}$ (endogenous variable), and $y_i \in \mathcal{R}(Y_i)$. We will follow Halpern & Pearl in abbreviating this formula using $[\vec{Y} \leftarrow \vec{y}]\phi$, in which $\vec{Y}$ and $\vec{y}$ are vectors of variables and values respectively. A *causal formula* is a Boolean combination of basic causal formulas. If $\vec{Y}$ is empty, this is abbreviated as just ϕ.

3.3 Semantics

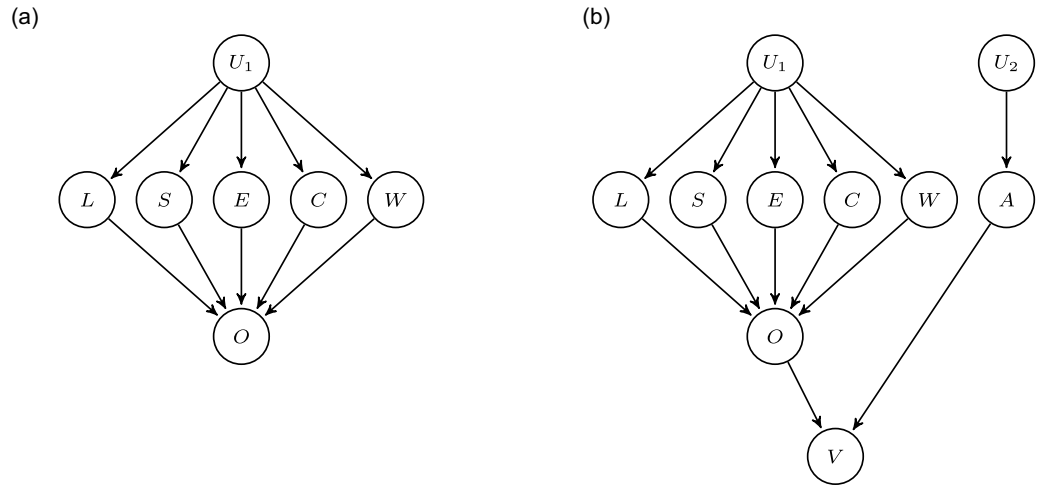
Intuitively, a formula $[\vec{Y} \leftarrow \vec{y}]\phi$ for a situation $(M, \vec{u})$ states that ϕ would hold in the model if the counterfactual case of $Y_i = y_i$ for each $Y_i \in \vec{Y}$ and $y_i \in \vec{y}$ were to occur. More formally, Halpern and Pearl define $(M, \vec{u}) \models \phi$ to mean that ϕ holds in the model and context $(M, \vec{u})$. The $\models$ relation is defined inductively by defining $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ as holding if and only if the unique value of X determined from the model $M_{\vec{Y}\leftarrow\vec{y}}$ is x, and defining Boolean combinations in the standard way.

Example 3.1. This section presents a simple example of a hypothetical system that classifies images of arthropods into several different types, taken from Miller (2018). The categorization is based on certain physical features of the arthropods, such as number of legs, number of eyes, number of wings, etc. Table 1 outlines a simple model of the features of arthropods for illustrative purposes.

The causal model for this has endogenous variables L (number of legs), S (stinger), E (number of eyes), C (compound eyes), W (number of wings), and O (the output). U_1 is an exogenous variable that determines the actual type of the arthropod, and therefore causes the values of the properties such as legs, wings, etc. The variables L, E, and W range over the natural numbers, while S and C are both Boolean. The output O ranges over the set {*Spider*, *Beetle*, *Bee*, *Fly*, *Unknown*}. A causal graph of this is shown in Figure 1a. The functions are clear from Table 1; for example, $F_O(8, \mathit{false}, 8, \mathit{no}, 0) = \mathit{Spider}$, and $O = \mathit{Unknown}$ for anything not in the table.

It is important to note that SCMs like this can be used to model correlative machine learning algorithms. It is not the physical features of number of legs, eyes, etc., that cause a real spider to be a spider—it

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press



Causal graph for arthropod algorithm defined in Example 3.1

Causal graph for extended arthropod algorithm defined in Example 3.2

Figure 1 (a) Causal graph for arthropod algorithm defined in Example 3.1 (b) (b) Causal graph for extended arthropod algorithm defined in Example 3.2.

is its genetic makeup that causes this. This genetic makeup causes both the spider's physical features and it to be a spider. However, here we model that a correlation is found between the variables represent physical features and the type, and the algorithm uses this to predict the arthropod type. However, the *causal* direction of the prediction model is from physical features to arthropod type because this is the way the prediction is made. The variables representing the physical features cause the decision, even though this does not model the causes in the real world. Of course, a model that represents causes in the real world may offer a better explanatory model, but even in cases where we do not have one, a straightforward mapping from features to outputs fits into our model of contrastive explanation.

Example 3.2. Consider a extension to the arthropod algorithm in Example 3.1 that verifies manual annotations on arthropod images. Images are labels with one of Spider, Beetle, Bee, Fly, or no label (Unknown), and the new algorithm extends the previous one to check whether the manual annotations are correct or not. The same categories exist, but some images are not labeled at all. To model this, we add a new exogenous variable U_2, which determines the new endogenous variable A—the annotation on the image. A second endogenous variable V with domain {*Pass*, *Fail*} determines whether the classifier output O corresponds with A. The causal graph is shown in Figure 1b. The function $F_V(O, A) = Pass$ if either $A = O$ or $A = Unknown$ or $O = Unknown$, to avoid too many false negatives. Otherwise, $F_V(O, A) = Fail$.

4 Contrastive 'why' questions

The basic problem of explanation is to answer a *why-question*. According to Bromberger (1966), a why-question is just a *whether-question*, preceded by the word 'why'. A whether-question is an interrogative question whose correct answer is either 'yes' or 'no'. The *presupposition* within a why-question is the fact referred to in the question that is under explanation, expressed as if it were true (or false if the question is a negative sentence). For example, the question '*why did they do that?*' is a why-question, with the inner whether-question being '*did they do that?*', and the presupposition being '*they did that*'.

However, as discussed already, why-questions are structurally more complicated than this: they are *contrastive*. The question then becomes: *what is a contrastive why-question?*

In this section, we extend Ylikoski's argument for the existence of (at least) two different types of contrastive why-questions (Ylikoski, 2007). In brief, the first asks why some fact happened rather than some other thing, called the *foil*, while the second asks why some fact happened in one situation while another fact, called the *surrogate*, happened in another (presumably similar) situation. The first type we

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

call *counterfactual* questions, because in this case, the foil is a counterfactual possibility to the fact. Intuitively, the fact and the foil are incompatible: it is not possible that both of them could have occurred. This is consistent with Temple's reading that Q offers an 'exclusive alternative in the circumstances' (Temple, 1988). The second type, we call *bi-factual* questions, because both the fact and the surrogate events actually occurred, but just in different contexts. The explainee is using the surrogate as a reference point to contrast against the fact. Using Halpern & Pearl's structural models (Halpern & Pearl, 2005a), we more crisply demonstrate why there is a difference between these two questions based on the relationships between the situations in which the fact and its contrast case (foil or surrogate) did and did not occur respectively.

4.1 Counterfactual questions

Given two events P and Q, Lipton (1990) defines a contrastive why-question as:

$$\text{Why } P \text{ rather than } Q? \tag{1}$$

For a counterfactual question, this means that, in some situation, the fact P occurred and the explainee is asking why foil Q did not occur in that situation instead. To semi-formalize this in structural models, a counterfactual why-question, given a situation $(M, \vec{u})$, is:

$$\text{Why } (M, \vec{u}) \models \phi \text{ rather than } \psi \ ? \tag{2}$$

in which ϕ is the fact and ψ is the foil. This assumes that ϕ is actually true in the situation $(M, \vec{u})$, and that ψ is not. The linguistic reduction to '*Why P and not-Q?*' is:

$$\text{Why } (M, \vec{u}) \models \phi \wedge \neg\psi? \tag{3}$$

To answer the question in Equation (3), one could argue that an explanation of such a case is a proof of ϕ and a counterexample for ψ. However, as argued by Lipton (1990), this is not really what is asked by '*Why ϕ rather than ψ?*'. The 'rather than' is asking for a relationship between the causes of ϕ and the causes (or non-causes) of ψ. As a counterexample to the reductionist argument, Lipton notes that we can answer a 'rather than' question without knowing all causes of the events. For instance, take the arthropod description from Example 3.1, and a question as to why the algorithm classified a particular image as a Bee rather than a Fly. Assume that we only know the value of one variable in the model: W—the number of wings. We cannot give the cause of $O = Bee$ if we do not know the values of the other variables[2]. However, we can still give a perfectively satisfactory answer to the question: it is a Bee rather than a Fly because it has four wings instead of two. As such, 'rather than' questions must be asking something different to just '*Why ϕ and why $\neg\psi$?*', for which we need to know all causes for both ϕ and ψ.

These counterfactual questions make sense as why-questions in artificial intelligence. Given the arthropod classification example, a 'rather than' question represents an observer asking why the output was a particular arthropod rather some other incompatible foil case; which would presumably often be the answer they were expecting.

In this paper, we assume that in counterfactual questions, ϕ and ψ are incompatible. It is clear that questions such as 'Why $x \leq 5$ rather than $x \geq 0$, where $x = 4$ and therefore both fact and foil are true, do not make sense. However, one could argue that it is possible to ask 'rather than' questions with compatible fact and foils over different variables; for example "Why $x = 4$ rather than $y = 5$?'. It is not difficult to find a structural model such that $x = 4$ and $y = 5$. However, the value of y in the actual situation must be something other than 5, otherwise the question does not make sense. So, the question is really 'Why $x = 4 \wedge y = 4$ rather than $x = 4 \wedge y = 5$?', which is incompatible. For this reason, we make the reasonable assumption that ϕ and ψ always refer to the same variables and they are incompatible in the given situation.

[2] Although in this trivial example, technically we could infer them all, but this is a property of the particular example, not of 'rather than' questions and structural models in general.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

4.2 Bi-factual questions

As outlined in Section 2, Ylikoski (2007) argues that some contrastive why-questions can have compatible facts and foils; although he terms a compatible foil as a *surrogate*. To be compatible, he argues that they must occur as part of two different 'processes'.

We model this second type of contrastive question, called a *bi-factual* questions, by modeling the two different processes as two different situations:

$$\text{Why } (M, \vec{u}) \models \phi \text{ but } (M', \vec{u}') \models \psi? \tag{4}$$

in which the $(M, \vec{u})$ and $(M', \vec{u}')$ are two different situations, including two different models M and M', ϕ is the fact, and ψ is the surrogate. Note the absence of 'rather than' in the question. Linguistically, this makes sense because both the fact and the surrogate are actual—there is no hypothetical case.

As a question in explainable AI, this question has a clear interpretation that M and M' refer to two different algorithms and $\vec{u}$ and $\vec{u}'$ define different 'inputs' to the algorithms. For the arthropod example, a valid question is why the algorithm produced the output ϕ for input image J, while some previous execution of the algorithm produced the different output ψ for different image K (note that $M = M'$ in this example). The observer is trying to understand why the outputs were different, when she expected ϕ to be ψ like it was in a previous instance. Another example is the case noted by Wang *et al*. (2019), discussed in Section 2.2, of clinicians wanting to compare similar cases with different outcomes.

In the case where $M \neq M'$, an example is in which model M' is an updated version of M—for example, new data has been feed into a learning approach to produce a more refined model —, and the explainee is asking for why the result has changed between the two models, potentially with $\vec{u} = \vec{u}'$.

Although not naturally worded as a 'rather than' question, it could be argued that the question is actually a 'rather than' question in which the person is asking 'Why ϕ this time and ψ last time *rather than* ϕ (or ψ) both times?':

$$\text{Why } (M, \vec{u}) \models \phi \text{ and } (M', \vec{u}') \models \psi$$
$$\text{rather than}$$
$$(M, \vec{u}) \models \phi \text{ and } (M', \vec{u}') \models \phi \text{ or } (M, \vec{u}) \models \psi \text{ and } (M', \vec{u}') \models \psi \text{ ?}$$

If we reduce this using the template of '*P and not-Q*', and simplify, the result is:

$$\text{Why } (M, \vec{u}) \models \phi \wedge \neg\psi \text{ but } (M', \vec{u}') \models \psi \wedge \neg\phi \text{ ?} \tag{5}$$

This is just the same as the question in Equation (4), however, it assumes that the fact and surrogate are incompatible. This assumption is too strong, because a perfectly valid question is why two different situations are producing the *same* outcome, despite the differences in the situation: the explainee expects the two outcomes to be different and wants an explanation as to why they are the same.

In this section, we have demonstrated a case for two types of contrastive why-question: counterfactual and bi-factual questions. In the remainder of the paper, we use structural causal models to define what answers to these questions look like, starting with how to define *contrastive cause* (Section 5) and then *contrastive explanation* (Section 6).

5 Contrastive cause

Before we turn to contrastive explanation, we define *contrastive cause*. Explanations typically cite only a subset of the actual causes of an event, and research shows that various different criteria are used to select these, such as their abnormality, or epistemic relevance; see Miller (2018) for a discussion of these. In Section 6, we build on the definition of explanation based on epistemic relevance by Halpern and Pearl (2005b). However, to do this, we first need to define what a contrastive cause is.

Informally, a contrastive cause between ϕ and ψ is a pair, in which the first element is a cause of ϕ and the second element is a cause of ψ. Intuitively, a contrastive cause $\langle A, B\rangle$ specifies that A is a cause of ϕ that does not cause ψ, while B is some corresponding event that causes ψ but does not cause ϕ. This is

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

consistent with existing philosophical views; for example, Ruben (1987) defines contrastive explanations as conjunctions between history of the contrasting events. The particular definition depends whether the why-question is counterfactual or bi-factual.

5.1 Non-contrastive cause

Our definition of contrastive cause extends Halpern & Pearl's definition of *actual cause* (Halpern & Pearl, 2005a). In their definition, causes are conjunctions of primitive events, represented as $\vec{X} = \vec{x}$, while the events to be described are Boolean combinations of primitive events.

Halpern and Pearl (2005a) define two types of cause: *sufficient cause* and *actual cause*. Intuitively, a sufficient cause of an event in a situation is a conjunction of primitive events such that changing the values of some variables in that conjunct would cause the event not to occur. An actual cause is simply a minimal sufficient cause; that is, it contains no unnecessary conjuncts.

More formally, the conjunction of primitive events $\vec{X} = \vec{x}$ is an actual cause of event ϕ in a situation $(M, \vec{u})$ if the following three properties hold:

AC1 $(M, \vec{u}) \models \vec{X} = \vec{x} \wedge \phi$—that is, both the event and the cause are true in the actual situation.

AC2 There is a set $\vec{W} \subseteq \mathcal{V}$ and a setting $\vec{x}'$ of variables $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}$ then $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\phi$—that is, if $\vec{X}$ did not have the values $\vec{x}$ and all variables in W remain the same, then event ϕ would not have occurred[3].

AC3 $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies AC1 and AC2—that is, there are no unnecessary primitive events in the conjunction $\vec{X} = \vec{x}$.

A *sufficient cause* is simply the first two items above—that is, a non-minimal actual cause.

Throughout the rest of this paper, we use the term *partial cause* to refer to a subset of conjunctions of an actual cause.

Example 5.1. Consider the arthropod example from Example 3.1. $L = 6$ (6 legs) is an actual cause of $O = Bee$ under the situation u_3 corresponding to line 3 of Table 1. AC1 holds trivially because $L = 6$ is in u_3 and $O = Bee$ is the output. AC2 holds because whenever $L \neq 6$, $O = Bee$ would not hold under u_3. AC3 holds because L is just one variable, so is minimal. Similarly, all other 'input' variables are actual causes in u_3; e.g. $E = 6$.

Example 5.2. For the extended model with annotated images from Example 3.2, consider the situation u_u in which there is no annotation ($A = Unknown$) and we have spider but with 7 legs ($L = 7$). If $L = 7$, then $O = Unknown$ and therefore the verification will pass ($V = Pass$), because this does not indicate an inconsistency.

One actual cause for $V = Pass$ is the pair ($L = 7, A = Unknown$). AC1 holds trivially. For AC2, we need to change both L and A to also change the value of V to $Fail$. If we change L to anything else, V will remain $Pass$ because $A = Unknown$, and similarly if we change A. It requires a mismatch in A and O other than $Unknown$ to produce $V = Fail$. AC3 holds because the pair of L and O is minimal. Similarly, the pair ($O = Unknown, A = Unknown$) is an actual cause. However, the triple ($L = 7, A = Unknown, O = Unknown$) is only a sufficient cause, because it is not minimal (violates AC3): we do not require both $L = 7$ and $O = Unknown$.

5.2 Contrastive causes for counterfactuals

To define contrastive cause, we adopt and formalize Lipton's difference condition (Lipton, 1990), which states that we should find causes that are different in the 'history' of the two events. We define the

[3] Note that this is the later definition from Halpern (2015), which is simplified compared to the original definition of Halpern & Pearl (2005a). Halpern argues this updated definition is more robust.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

'history' as the situation $(M, \vec{u})$ under which the events are evaluated; that is, (M, u) for counterfactual why-questions, and both (M, u) and $(M', \vec{u}')$ for bi-factual why-questions.

The particular explanandum for which we want to define cause is no longer a single event ϕ, but a pair of events $\langle \phi, \psi \rangle$, in which ϕ is the fact and ψ is the foil. Similarly, causes will consist of two events instead of one, consistent with the difference condition.

Informally, a contrastive counterfactual cause of a pair of events $\langle \phi, \psi \rangle$ is a pair of *partial* causes, such that the difference between the two causes is the minimum number of changes required to make ψ become true.

DEFINITION 1 (Contrastive Counterfactual Cause). *A pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is an contrastive counterfactual actual cause (also just a counterfactual cause) of $\langle \phi, \psi \rangle$ in situation $(M, \vec{u})$ if and only if the following conditions holds:*

CC1 *$\vec{X} = \vec{x}$ is a partial cause of ϕ under $(M, \vec{u})$.*

CC2 *$(M, \vec{u}') \models \neg\psi$—the foil ψ is not true.*

CC3 *There is a non-empty set $\vec{W} \subseteq \mathcal{V}$ and a setting $\vec{w}$ of variables in $\vec{W}$ such that $\vec{X} = \vec{y}$ is a partial cause of ψ under situation $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$.*
Informally, this states that there is some hypothetical situation that did not happen, but is feasible in M; and that $\vec{X} = \vec{y}$ is a partial cause of ψ under this hypothetical situation.

CC4 *$(\vec{X} = \vec{x} \cap \vec{X} = \vec{y}) = \emptyset$—that is, there are no common events. This is the difference condition.*

CC5 *$\vec{X}$ is maximal—that is, no superset of $\vec{X}$ satisfies CC1-4.*

Similar to the HP definition, we can define *sufficient* contrastive cause by modifying CC1 and CC3 to refer to partial sufficient causes.

This definition is based on the Halpern (2015) definition of actual cause, as conditions CC1-3 directly access partial causes, which are subsets of actual causes. However, the definition is modular with respect to the underlying definition of actual cause, such that a different definition of actual cause (using structural models), such as the original definition from Halpern and Pearl Halpern & Pearl (2005a), could be substituted, and this would change the semantic interpretation of the above.

The reader may expect to see that CC2 had an additional statement that no part of the hypothetical cause of ψ is true, such as $\bigwedge_{X_i = y_i \in \vec{X} = \vec{y}} X_i \neq y_i$. However, this is implied by CC4, because all elements of $\vec{X} = \vec{x}$ are true, and each element of $\vec{X} = \vec{y}$ is different from its corresponding value in $\vec{X} = \vec{x}$. Also note that condition CC3 implies that the foil ψ is feasible in M. That is, it implies that $M \not\models \neg\psi$. For an infeasible event, there cannot be another situation such $\vec{X} = \vec{y}$ is a cause of ψ, therefore there can be no difference condition. This seems reasonable though: asking why an infeasible foil did not occur should not invoke a difference between the fact and foil, but a description that the foil is infeasible.

Example 5.3. Consider the arthropod example from Example 3.1, asking why an image was categorized as a Bee instead of a Fly. To answer the counterfactual why-question, we take the maximal intersection of two actual causes of *Output* = *Bee* and the hypothetical cause of *Output* = *Fly*. In this case, the following pairs correspond to the possible contrastive causes:

$$\langle S = \checkmark, S = \times \rangle$$
$$\langle W = 4, W = 2 \rangle.$$

The image was classified as a Bee instead of a Fly because the image contains a stinger (S) and four wings (W), while for a Fly, it would have required no stinger and two wings. The other actual causes of ϕ and ψ, such as $L = 6$, are not contrastive causes because they do not satisfy the difference condition in CC4.

It is difficult to argue that a particular definition of contrastive cause is correct. However, we can at least argue that they abide by some commonly accepted properties; specifically, the properties of an

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

adequate explanation defined by Hesslow (1983) (see Section 2.1). This states that, if the counterfactual causes hold in a different situation, so to would the counterfactual events. The following theorem captures this.

THEOREM 1. *If $\mathcal{C}$ comprises all counterfactual contrastive actual causes of $\langle \phi, \psi \rangle$ under situation $(M, \vec{u})$, then for any maximal-consistent subset*[4] $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle \subseteq \mathcal{C}$*:*

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{y}]\psi$*; and*
(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{y}]\neg\phi$.

We need to consider only the maximal-consistent subsets because the set of all contrastive causes could be inconsistent if there are multiple sufficient causes.

Proof. Consider part (a) first. We prove via contradiction. Assume that $(M, \vec{u}) \not\models [\vec{X} \leftarrow \vec{y}]\psi$. From CC3, $\vec{X} = \vec{y}$ contains partial causes of ψ, so there must be a set of additional causes $\vec{Z} = \vec{z}$ such that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{y}, \vec{Z} \leftarrow \vec{z}]\psi$. This implies that there is some (maximal) subset $\vec{Z} = \vec{z}' \subseteq \vec{Z} = \vec{z}$ such that $(M, \vec{u}) \not\models \vec{Z} = \vec{z}'$, and is therefore not in $\vec{X} = \vec{x}$. However, these two implications mean that CC3 and CC4 hold for $\vec{Z} = \vec{z}'$. CC5 also holds because $\vec{Z} = \vec{z}'$ is maximal. Therefore, $\vec{Z} = \vec{z}'$ is (one half of) a contrastive cause for $\langle \phi, \psi \rangle$, and as such, must be part of $\mathcal{C}$. Because $\vec{X} = \vec{y}$ is maximal, $\vec{Z} = \vec{z}'$ must be in $\vec{X} = \vec{y}$, so it is not possible that both $(M, \vec{u}) \not\models [\vec{X} \leftarrow \vec{y}]\psi$ and $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{y}, \vec{Z} \leftarrow \vec{z}]\psi$ are true. This contradiction shows that part (a) holds. Part (b) holds directly because ϕ and ψ are incompatible. □

5.3 Contrastive causes in bi-factuals

For bi-factual questions, the definition of 'history' is different to that of counterfactual questions, citing two different situations. We define the 'history' as the situations $(M, \vec{u})$ of ϕ and $(M', \vec{u}')$ of ψ. For the moment, we simplify this by assuming that the two causal models M and M' are the same; e.g. the same algorithm is executed with different inputs from the environment. We drop this assumption later.

DEFINITION 2 (Contrastive Bi-factual Cause—Simple Case). *A pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a contrastive bi-factual actual cause of $\langle \phi, \psi \rangle$ in their respective situations $(M, \vec{u})$ and $(M, \vec{u}')$ if:*

BC1 *$\vec{X} = \vec{x}$ is a partial cause of ϕ under $(M, \vec{u})$.*
BC2 *$\vec{X} = \vec{y}$ is a partial cause of ψ under $(M, \vec{u}')$.*
BC3 *$(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$—that is, there are no common events. This is the difference condition.*
BC4 *$\vec{X}$ is maximal—that is, no superset of $\vec{X}$ satisfies BC1-3.*

Note that BC1 implies $(M, \vec{u}) \models \vec{X} = \vec{x} \wedge \phi$ (AC1) and similarly for BC2.

A *sufficient* contrastive cause can be obtained by modifying BC1 and BC2 to refer to partial sufficient causes.

This definition is simpler than that of counterfactual explanation (compare CC3 with BC2), because both the fact and surrogate are actual events, whereas in counterfactual questions, the foil is hypothetical.

Example 5.4. Consider again the arthropod example from Example 3.1, and the contrastive why-question for two images B and F, in which B was categorized as a Bee and F a fly. The situations for these two cases are straightforward to extract from Table 1, as are the causes. To answer the contrastive why-question, we take the maximal intersection actual causes of *Output* = *Bee* and *Output* = *Fly* under models $(M, \vec{u}_B)$ and $(M, \vec{u}_F)$ respectively, which is simply the same as in Example 5.3:

$$\langle S = \checkmark, S = \times \rangle$$
$$\langle W = 4, W = 2 \rangle.$$

[4] We abuse notation slightly here: $\vec{X} = \vec{x}$ is the conjunction of the first items of all of the subset; similarly $\vec{X} = \vec{y}$ is the conjunction of the second items.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

Note that the difference condition is the same as is in the counterfactual case, however, in this case, there was no need to find a hypothetical situation for the foil.

THEOREM 2. *If $\mathcal{C}$ comprises all counterfactual contrastive actual causes of $\langle \phi, \psi \rangle$ under respective situations $(M, \vec{u})$ and $(M, \vec{u}')$ then for any maximal-consistent subset $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle \subseteq \mathcal{C}$:*

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{y}]\psi$; *and*
(b) $(M, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\phi$.

Proof. The proofs for both parts are similar to the proof for counterfactual causes in Theorem 1, except that we refer to BC2-4 instead of CC3-5. □

Now we return to the case in which the two models may be different. For this, we define a *restricted cause* of ϕ under situation $(M, \vec{u})$, where $M = (\mathcal{S}, \mathcal{F})$, as a pair $(\mathcal{F}', \vec{X} = \vec{x}')$, in which $\vec{X} = \vec{x}'$ is a sufficient cause of ϕ, and $\mathcal{F}' \subseteq \mathcal{F}$ is the smallest subset of $\mathcal{F}$ required to derive ϕ. That is, for all $\vec{u}'$, $(M, \vec{u}') \models \phi$ iff $(M^{\mathcal{F}'}, \vec{u}') \models \phi$, where $M^{\mathcal{F}'} = (S, \mathcal{F}')$, and therefore all functions $F \setminus F'$ do not influence ϕ in any situation. A *partial restricted cause* is simply $(F^{\phi}, \vec{X} = \vec{x})$ such that $F^{\phi} \subseteq \mathcal{F}'$ and $\vec{X} = \vec{x} \subseteq \vec{X} = \vec{x}'$.

DEFINITION 3 (Contrastive Bi-factual Cause—General Case). *A pair $\langle (F^{\phi}, \vec{X} = \vec{x}), (F^{\psi}, \vec{Y} = \vec{y}) \rangle$ is a contrastive bi-factual actual cause of $\langle \phi, \psi \rangle$ in their respective situations $(M, \vec{u})$ and $(M', \vec{u}')$ if and only if the following conditions hold:*

BC1$_{\mathrm{G}}$ $(F^{\phi}, \vec{X} = \vec{x})$ *is a partial restricted cause of ϕ under situation $(M, \vec{u})$.*
BC2$_{\mathrm{G}}$ $(F^{\psi}, \vec{Y} = \vec{y})$ *is a partial restricted cause of ψ under situation $(M', \vec{u}')$.*
BC3$_{\mathrm{G}}$ $F^{\phi} \cap F^{\psi} = \emptyset$ *and* $(\vec{X} = \vec{x}) \cap (\vec{Y} = \vec{y}) = \emptyset$*—that is, there are no common functions or pairs of events. This is the* difference condition.
BC4$_{\mathrm{G}}$ $(F^{\phi}, \vec{X} = \vec{x}, F^{\psi}, \vec{Y} = \vec{y}, \vec{X} \cap \vec{Y})$ *is maximal.*
That is, there is no tuple $(F^{\phi'}, \vec{X} = \vec{x}', F^{\psi'}, \vec{Y} = \vec{y}', \vec{X}' \cap \vec{Y}') \neq (F^{\phi}, \vec{X} = \vec{x}, F^{\psi}, \vec{Y} = \vec{y}, \vec{X} \cap \vec{Y})$ satisfying BC1$_G$, BC2$_G$, and BC4$_G$ such that $F^{\phi} \subseteq F^{\phi'}$, $F^{\psi} \subseteq F^{\psi'}$, $\vec{X} = \vec{x} \subseteq \vec{X} = \vec{x}'$, $\vec{Y} = \vec{y} \subseteq \vec{Y} = \vec{y}'$, and $\vec{X} \cap \vec{Y} \subseteq \vec{X}' \cap \vec{Y}'$.

Note two differences between this and the less general version. First, the definition refers to differences in the functions of the two models. Second, the sets of variables that are referred to are no longer *shared* between the two models. That is, in the less general definition, the contrastive cause $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ both pointed to $\vec{X}$. However, the sets of variables can be different between the two models M and M', so it is possible that some variable in $\vec{Y}$ does not exist in model M, but is a cause of ψ in $(M', \vec{u}')$. Note that BC4$_{\mathrm{G}}$ also states that $\vec{X} \cap \vec{Y}$ is maximal, meaning that the two parts of a contrastive cause can only cite different variables if at least one of the models does not contain that variable.

In the case where $M = M'$, this definition is the same as Definition 2 because $F^{\phi} = F^{\psi} = \emptyset$ and $\vec{X} = \vec{Y}$.

Example 5.5. Consider now the combination of Examples 3.1 (the simple arthropod classification example) and 3.2 (the extended example in which images may come annotated). Let M be the model without the extension and M′ be the extended model. Asking why $(M, \vec{u}) \models O = \mathit{Unknown}$ and $(M', \vec{u}') \models O = \mathit{Bee}$, in which $\vec{u}$ and $\vec{u}'$ both correspond to features of a *Bee* but $L = 5$ (five legs) and $A = \mathit{Bee}$ in $\vec{u}'$, a contrastive cause would be:

$$\langle (F_O = f, \emptyset), (F_O = f', A = \mathit{Bee}) \rangle,$$

in which f and f' refer to the before and after functions for F_O in M and M' respectively, and are hopefully clear from the description. Here, the contrast cites the change in functions and the additional cause $A = \mathit{Bee}$ as the difference condition.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

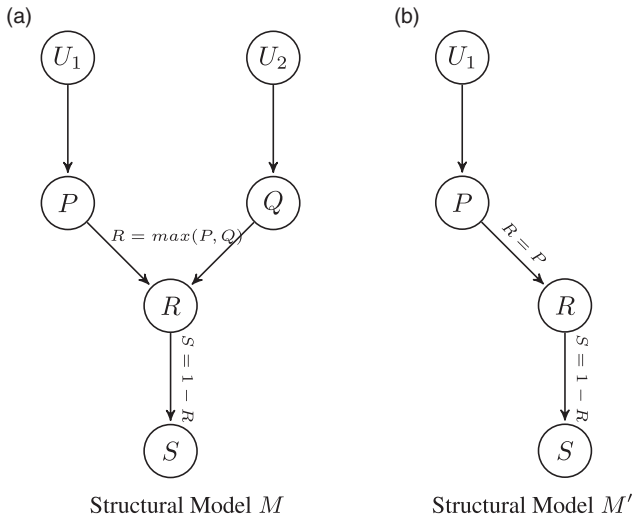


Figure 2 Structural Models for Example 5.6.

Example 5.6. The more general definition is also useful for reasoning about situations in which the fact and surrogate are the same event. That is, 'Why $(M, \vec{u}) \models \phi$ but $(M', \vec{u}') \models \phi'$? This is useful for situations in which an observer wants to understand why the event ϕ still occurs despite the model changing. As an example, consider the two simple structural models in Figure 2, with exogenous variables U_1 and U_2, and endogenous variables P, Q, R, and S. S depends on all four variables in M, but there is no variable Q in model M′.

For the contrast between $(M, \vec{u}) \models S = 0$ and $(M', \vec{u}') \models S = 0$, in which $\vec{u}$ leads to $P = 1, Q = 0, R = 1$, and $S = 0$ and $\vec{u}'$ leads to $P = 1, R = 1, S = 0$, the contrastive cause would be cited as:

$$\langle (F_R = \max(P, Q), \emptyset), (F_R = P, \emptyset) \rangle.$$

That is, the difference is in the function F_R, not in any of the variables nor the output.

To explore some properties of this, we introduce notation that allows us to reason about changes in models at a meta level. Recall that a structural model $M = (\mathcal{S}, \mathcal{F})$ consists of a set of signatures $\mathcal{S}$ and a set of functions $\mathcal{F}$. We define the *override* of a set of functions $\mathcal{F}$ by another set $\mathcal{F}'$, denoted $\mathcal{F} \Leftarrow \mathcal{F}'$ being the same as $\mathcal{F}$, except replacing F_X with F'_X for all variables X such that $F'_X \in \mathcal{F}$. The notation $M \Leftarrow \mathcal{F}'$ represents the overriding of the functions in M with $\mathcal{F}'$.

THEOREM 3. *If $\mathcal{C}$ comprises all counterfactual contrastive actual causes of $\langle \phi, \psi \rangle$ under respective situations $(M, \vec{u})$ and $(M, \vec{u}')$ then for any maximal-consistent subset $\langle (F^{\phi}, \vec{X} = \vec{x}), (F^{\psi}, \vec{Y} = \vec{y}) \rangle \subseteq \mathcal{C}$, the following hold:*

(a) $(M \Leftarrow F^{\psi}, s) \models [\vec{Y} \leftarrow \vec{y}]\psi$
(b) $(M' \Leftarrow F^{\phi}, s) \models [\vec{X} \leftarrow \vec{x}]\phi$.

Proof. The proof for this is an extension of the proof for Theorem 2. The only case that requires attention is when variables are added/removed to/from the model. In this case, the model $M \Leftarrow F^{\psi}$ may contain the function F_X, which is in M but not M'. However, if the variable X is not in M', then ψ cannot refer to it, so it is effectively redundant in $M \Leftarrow F^{\psi}$.

5.4 Presuppositions

As noted previously, it is difficult to argue that a particular definition of contrastive cause is correct, but we can show our definition behaves according to some commonly accepted properties. In this section, we

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

show that our definition is consistent with the idea of contrastive explanation as *presupposed explanation* (Lipton, 1990).

Lipton (1990) notes that to give an explanation for 'Why P rather than Q?' is to give a to 'give a certain type of explanation of P, *given* P or Q, and an explanation that succeeds with the presupposition will not generally succeed without it.' (Lipton, 1990, p. 251) (emphasis original). Thus, this states that if we assume that P and Q are the only two possible outcomes, and are mutually exclusive, then the actual cause of P under this assumption will refer to exactly those variables in the difference condition.

Formally, the assumption is $M \models \phi \oplus \psi$—that is, under all models of M, either ϕ is true or ψ is true, and not both. Note the absence of a situation $\vec{u}$. Thus, we can re-phrase a counterfactual question as:

$$\text{Assuming } M \models (\phi \oplus \psi), \text{ why } (M, \vec{u}) \models \phi \ ? \tag{6}$$

As a shorthand, we use $M^{\phi \oplus \psi}$ to refer to the sub-model of M in which $\phi \oplus \psi$ is always true. That is, the functions in $\mathcal{F}$ are restricted such that assignments to all variables always conform to $\phi \oplus \psi$.

The set of events $\vec{X} = \vec{x}$ is a *presupposed contrastive cause* of ϕ under situation $(M, \vec{u})$ and assumption $M \models \phi \oplus \psi$ if and only if the following condition holds:

PAC $\vec{X} = \vec{x}$ is an actual cause of ϕ under the situation $(M^{\phi \oplus \psi}, \vec{u})$.

That is, if we assume that $\phi \oplus \psi$ always holds in a structural model, then an *actual* cause of ϕ in that model under situation $\vec{u}$ is sufficient to identify the different condition. Note here that the cause is not contrastive because it is not a pair—it just refers to the variables in $\vec{X}$ and their values in $\vec{u}$. However, this is enough for us to propose the following theorem.

THEOREM 4. *$\vec{X} = \vec{x}$ is an actual cause of ϕ under situation $(M, \vec{u})$ assuming $M \models \phi \oplus \psi$ if and only if $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a counterfactual contrastive cause of $\langle \phi, \psi \rangle$ under situation $(M, \vec{u})$ for some $\vec{y}$.*

Proof. This theorem is effectively stating that if AC1-3 hold assuming $\phi \oplus \psi$, then CC1-5 hold for some $\vec{y}$, and vice-versa.

The left-to-right case: For CC1, if $\vec{X} = \vec{x}$ is an actual cause under a restricted model $M^{\phi \oplus \psi}$, then model M must admit $\vec{X} = \vec{x}$ as (at least) a partial cause for ϕ. For CC2, $(M, \vec{u}) \models \neg\psi$ must hold because ϕ holds according to AC1, and ϕ and ψ are mutually exclusive. For the remainder, we need to show that a $\vec{y}$ exists such that CC3-5 hold. From AC2, we know that there exists some counterfactual situation in which ϕ would not have occurred under $M^{\phi \oplus \psi}$. In such a situation, it must be that ψ occurred, so all such situations would be candidate values for $\vec{y}$. This implies CC3. In addition, the values in $\vec{y}$ must make ψ true, and therefore must be different from the values in $\vec{x}$, so CC4 holds. Finally, we prove CC5 (maximality) by contradiction. Assume that $\vec{X}$ is not maximal. This implies there exists some additional variables $\vec{Y}$ not in $\vec{X}$ that must change to make ψ hold under $(M, \vec{u})$. However, this would also require these variables to change under $M^{\phi \oplus \psi}$, which would mean that $\vec{X} = \vec{x}$ is not a complete actual cause of ϕ, contradicting the definition of PAC. Therefore, $\vec{X}$ must be maximal.

For the right-to-left case, AC1 is implied trivially by CC1: $\vec{X} = \vec{x}$ and ϕ hold under $M^{\phi \oplus \psi}$, and expanding the model without changing the structural equations themselves will not change the ϕ. AC2 is implied by CC3: if there is an alternative situation $\vec{u}'$ under M such that ψ holds, then that same situation must exist in $M^{\phi \oplus \psi}$ because $M^{\phi \oplus \psi}$ does not exclude situations in which ψ holds, so any such situation gives us the setting for $\vec{x}'$ that is required for the counterfactual situation in AC2.

For AC3, we need to show that the partial cause $\vec{X} = \vec{x}$ under M is minimal under $M^{\phi \oplus \psi}$. We prove this by contradiction. Assume that $\vec{X} = \vec{x}$ is not minimal under $M^{\phi \oplus \psi}$. This means that there is some variable W that has no effect on ϕ under $M^{\phi \oplus \psi}$, but is cited as a contrastive cause. Therefore, some part of the contrastive cause cites the events $(\vec{W} = \vec{w}, \vec{W} = \vec{z})$ for some $\vec{w}, \vec{z}$, and that $\vec{W} = \vec{w}$ is a partial cause of ϕ under $(M, \vec{u})$ and $\vec{W} = \vec{z}$ is a partial cause of ψ under the hypothetical situation in CC3. However, $\vec{W} = \vec{z}$ must then be a counterfactual case for $\vec{W}$ that satisfies AC2 under $M^{\phi \oplus \psi}$, meaning that it affects ϕ. This is a contradiction for our assumption that $\vec{X} = \vec{x}$ is not minimal. □

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

THEOREM 5. *$\vec{X} = \vec{x}$ is an actual cause of ϕ under situation $(M, \vec{u})$ assuming $M \models \phi \oplus \psi$ and $\vec{X} = \vec{y}$ is an actual cause of ψ under situation $(M', \vec{u}')$ assuming $M \models \phi \oplus \psi$ if and only if $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a contrastive bi-factual cause of $\langle \phi, \psi \rangle$ under situations $(M, \vec{u})$ and $(M', \vec{u})$.*

Proof. The proof is a straightforward extension of the proof from Theorem 4. In brief: the $\vec{y}$ referred to in Theorem 4 is from the surrogate. The two cases on the left of the *if and only if* are symmetric, so the proof above extends to this. □

6 Contrastive explanation

Now that we have defined contrastive cause, we can define contrastive explanation. This is a simple extension to the existing definition of Halpern & Pearl (2005a)'s definition, but using contrastive causes instead of standard actual causes.

6.1 Non-contrastive explanation

In Part II (Halpern & Pearl, 2005b) of their paper, Halpern & Pearl build on the definition of causation from Part I to provide a definition of *causal explanation*. They define the difference between causality and explanation as such: causality is the problem of determining which events cause another, whereas explanation is the problem of providing the necessary information in order to establish causation. Thus, an explanation is a fact that, if found to be true, would be a cause for an explanandum, but is initially unknown. As such, they consider that explanation should be relative to an *epistemic state*. This is in fact a definition of contrastive explanation using *epistemic relevance* (Slugoski *et al.*., 1993).

Informally, an explanation is defined in their framework as follows. Consider an agent with an epistemic state $\mathcal{K}$, who seeks an explanation of event ϕ. A good explanation should: (a) provide more information than is contained in $\mathcal{K}$; (b) update $\mathcal{K}$ in such a way that the person can now understand the cause of ϕ; and (c) it *may* be a requirement that ϕ is true or probable[5].

Halpern and Pearl (2005b) formalize this by defining $\mathcal{K}$ as a set of contexts, which represents the set of 'possible worlds' that the questioning agent considers possible. Therefore, an agent believes ϕ if and only if $(M, \vec{u}) \models \phi$ holds for every $\vec{u}$ in its epistemic state $\mathcal{K}$. A complete explanation effectively eliminates possible worlds of the explainee so that they can now determine the cause. Formally, an event $\vec{X} = \vec{x}$ is an *explanation* of event ϕ relative to a set of contexts $\mathcal{K}$ if the following hold:

EX1 $(M, \vec{u}) \models \phi$ for each $\vec{u} \in \mathcal{K}$—that is, the agent believes that ϕ.
EX2 $\vec{X} = \vec{x}$ is a *sufficient cause* of ϕ for all situations $(M, \vec{u})$ where $u \in \mathcal{K}$ such that $(M, \vec{u}) \models \vec{X} = \vec{x}$.
EX3 $\vec{X}$ is minimal—no subset of $\vec{X}$ satisfies EX2.
EX4 $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ for some $\vec{u} \in \mathcal{K}$ and $(M, \vec{u}') \models \vec{X} = \vec{x}$ for some (other) $\vec{u}' \in \mathcal{K}$—that is, before the explanation, the agent is initially uncertain whether the information contained in the explanation is true or not, meaning the explanation meaningfully provides information.

Example 6.1. Consider the basic arthropod example (Example 3.1), in which $O = Unknown$ due to a spider with only 7 legs. The agent knows that the image has 8 eyes and no stinger, but is uncertain of the remaining variables. The explanation for why $O = Unknown$ is just $L = 7$ (7 legs). This is a sufficient cause for $O = Unknown$, is minimal, and the agent does not know it previously.

For the extended arthropod example, consider the same case, but with $V = Pass$ (known to the agent) and $A = Unknown$ (unknown to the agent). An explanation for why $V = Pass$ would cite the pair $(O = Unknown, A = Unknown)$. The agent would need to know both parts of information to determine the cause. Another explanation would be $(L = 7, A = Unknown)$, as knowing $L = 7$ allows the agent to determine $O = Unknown$. If the agent already knows $O = Unknown$, then the explanation is a singleton again; either $L = 7$ or $O = Unknown$ will suffice.

[5] In the case of an explainer and explainee, we may say that it is 'believed' by the explainer.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

6.2 Contrastive counterfactual explanation

We extend the above definition to contrastive counterfactual causes. As with the Halpern and Pearl definition, it is defined relative to an epistemic state and model, however, as it describes a contrastive cause, the explanation is a pair.

DEFINITION 4 (Contrastive counterfactual Explanation). *Given a structural model M, a pair of events* $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a contrastive counterfactual explanation of* $\langle \phi, \psi \rangle$ *relative to* $\mathcal{K}$ *if and only if the following hold:*

CE1 $(M, \vec{u}) \models \phi \wedge \neg\psi$ *for each* $\vec{u} \in \mathcal{K}$*—that is, the agent accepts that* ϕ *and that* $\neg\psi$.

CE2 $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a sufficient counterfactual cause for* $\langle \phi, \psi \rangle$*, for each* $\vec{u} \in \mathcal{K}$ *such that* $(M, \vec{u}) \models \vec{X} = \vec{x}$.

CE3 $\vec{X}$ *is minimal—no subset of* $\vec{X}$ *satisfies CE2.*

CE4 $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ *for some* $\vec{u} \in \mathcal{K}$ *and* $(M, \vec{u}') \models \vec{X} = \vec{x}$ *for some (other)* $\vec{u}' \in \mathcal{K}$*; and for some* $\vec{W} = \vec{w}$ *such that* $\vec{w} \neq \vec{x}$, $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u}) \models \vec{X} = \vec{y}$ *for some* $\vec{u} \in \mathcal{K}$ *and* $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u}') \models \neg(\vec{X} = \vec{y})$ *for some (other)* $\vec{u}' \in \mathcal{K}$*—that is, agent is initially uncertain whether the explanation is true or not, meaning the explanation provides meaningful information.*

Example 6.2. Consider the same two cases from Example 6.1. An explanation for why $O = Unknown$ rather than $O = Spider$ would cite the pair $\langle L = 7, L = 8 \rangle$: the image has 7 legs but requires 8 to be a spider. We can already see that this is more informative than the non-contrastive cause, because we are given the counterfactual case of what should have been to make $O = Spider$.

For the extended case, an explanation for why $V = Pass$ rather than $V = Fail$ (the only possible foil) is the pair of tuples $\langle (O = Unknown, A = Unknown), (O = X, A = X) \rangle$, where X is one of *Spider*, *Beetle*, etc., or the pair of tuple $\langle (L = 7, A = Unknown), (L = 8, A = Spider) \rangle$, and similarly for other types. Again, if the agent already knows A or L, then pairs of singletons suffice.

Definition 4 defines a counterfactual contrastive explanation as finding part of a counterfactual cause that satisfies the conditions CE1-4. However, we can think of this in different way: finding partial explanations for each of ϕ and ψ and taking the difference between these, where we define a *partial explanation* as just a subset of an explanation.

DEFINITION 5 (Contrastive Counterfactual Explanation—Alternative Definition). *Given a structural model M, a pair of events* $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a contrastive counterfactual explanation of* $\langle \phi, \psi \rangle$ *relative to* $\mathcal{K}$ *if and only if the following hold:*

CE1′ $\vec{X} = \vec{x}$ *is a partial explanation of* ϕ *in* $(M, \vec{u})$.

CE2′ *There is a non-empty set* $\vec{W} \subseteq \mathcal{V}$ *and a setting* $\vec{w}$ *of variables in* $\vec{W}$ *such that* $\vec{X} = \vec{y}$ *is a partial explanation of* ψ *under situation* $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$.

CE3′ $(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$*—the difference condition.*

CE4′ $\vec{X}$ *is maximal—that is, there is no superset of* $\vec{X}$ *that satisfies CE1′-3′.*

THEOREM 6. *CE1-4 iff CE1-4′—that is, the two definitions of contrastive counterfactual explanation are equivalent.*

Proof. Left-to-right case: (CE1′) This holds from CE2-4. If $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a sufficient counterfactual cause for $\langle \phi, \psi \rangle$ that is minimal and uncertain, then $\vec{X} = \vec{x}$ must be a partial explanation of ϕ under $(M, \vec{u})$; that is, the agent believes ϕ, some superset of $\vec{X} = \vec{x}$ is an actual cause of ϕ, and the agent is uncertain about some of that superset. (CE2′) The same argument holds, except that $\vec{X} = \vec{y}$ is true under the hypothetical situation $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$ from CE2; and therefore, this hypothetical situation is a witness for CE2′. (CE3′) The difference condition holds because this is a requirement the sufficient counterfactual cause in of CE2. (CE4′) holds from the maximality condition in CE2. This establishes the left-to-right case.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

Right-to-left case: (CE1) This holds directly from CE1′, because the acceptance of ϕ in $\mathcal{K}$ is a condition of an explanation under the original Halpern & Pearl definition, and ψ must be false whenever ϕ is true. (CE2) If $\vec{X} = \vec{x}$ and $\vec{X} = \vec{y}$ are partial explanations of ϕ under $(M, \vec{u})$ and ψ under $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$ respectively, then they must be partial causes too. If their intersection is empty and $\vec{X}$ is maximal, then this defines a sufficient cause, so CE2 holds. (CE3) We prove this via contradiction. Assume $\vec{X}$ is not minimal. This implies that there is some strict superset $\vec{Y} \supset \vec{X}$ that satisfies CE1′-4′ and CE2. However, if this were the case, then CE4′ would not hold: $\vec{X}$ would not be maximal over CE1′-3′, which is a contradiction, so our assumption is false. (CE4) If $\vec{X} = \vec{x}$ and $\vec{X} = \vec{y}$ are partial explanations under $(M, \vec{u})$ and some hypothetical counterfactual case $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$, then the agent must be uncertain about $\vec{X} = \vec{x}$ in $(M, \vec{u})$ and $\vec{X} = \vec{y}$ in $(M_{\vec{W} \leftarrow \vec{w}}, \vec{u})$. This establishes the right-to-left case, and the theorem holds. □

6.3 Bi-factual contrastive explanation

For the bi-factual case, an explanation is similar, however, it refers to two epistemic states, $\mathcal{K}$ and $\mathcal{K}'$, in which $\mathcal{K}$ models the uncertainty of the individual in the situation $(M, \vec{u})$ and $\mathcal{K}'$ models the uncertainty in $(M', \vec{u}')$.

DEFINITION 6 (Contrastive Bi-factual Explanation—Simple Case). *Given a structural model M, a pair* $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a contrastive bi-factual explanation of* $\langle \phi, \psi \rangle$ *relative to two epistemic states* $\mathcal{K}$ *and* $\mathcal{K}'$ *if and only if the following hold:*

BE1 $(M, \vec{u}) \models \phi$ *for each* $\vec{u} \in \mathcal{K}$ *and* $(M', \vec{u}') \models \psi$ *for each* $\vec{u}' \in \mathcal{K}'$*—that is, the agent accepts that* ϕ *under* $(M, \vec{u})$ *and that* $\neg\psi$ *under* $(M', \vec{u}')$.

BE2 *for each* $\vec{u} \in \mathcal{K}$ *such that* $(M, \vec{u}) \models \vec{X} = \vec{x}$ *and* $\vec{u}' \in \mathcal{K}'$ *such that* $(M', \vec{u}') \models \vec{X} = \vec{y}$, $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a sufficient bi-factual cause for* $\langle \phi, \psi \rangle$ *under* $(M, \vec{u})$ *and* $(M', \vec{u}')$.

BE3 $\vec{X}$ *is minimal—that is, no superset of* $\vec{X}$ *satisfies BE2.*

BE4 $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ *for some* $\vec{u} \in \mathcal{K}$ *and* $(M, \vec{u}') \models \vec{X} = \vec{x}$ *for some (other)* $\vec{u}' \in \mathcal{K}$*; and* $(M, \vec{u}) \models \vec{X} = \vec{y}$ *for some* $\vec{u} \in \mathcal{K}'$ *and* $(M, \vec{u}') \models \neg(\vec{X} = \vec{y})$ *for some (other)* $\vec{u}' \in \mathcal{K}'$*—that is, the agent is initially uncertain whether the explanation is true or not, meaning the explanation provides meaningful information.*

This is similar to the definition of CE, except that the rules refer to an actual situation $\vec{u}'$, rather than the hypothetical situation implied by CE2. The more general case in which there are differences between the models is straightforward projection of this.

Example 6.3. Consider the case of the 7-legged spider (situation $\vec{u}_7$), and a second case of a 'proper' spider ($\vec{u}_8$). The agent is uncertain of all variables and asks why $O = \textit{Unknown}$ under $(M, \vec{u}_7)$ and $O = \textit{Spider}$ under $(M, \vec{u}_8)$. The explanation is as before: $\langle L = 7, L = 8 \rangle$. Note here that the agent already knows that $L = 8$, because it knows that $O = \textit{Spider}$, so can determine the values of the input variables. However, we still cite this in the explanation because it contrasts $L = 7$. The extended case is similar to the counterfactual explanation.

Definition 6 defines bi-factual explanation as finding part of a counterfactual cause that satisfies the conditions BE1-4. However, we can think of bi-factual explanation in different way: finding partial explanations for each of ϕ and ψ and taking the difference between these, where we define *partial explanation* as just subsets of explanations.

DEFINITION 7 (Contrastive Bi-factual Explanation—Simple Case, Counterfactual Definition). *Given a structural model M, a pair of events* $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ *is a contrastive bi-factual explanation of* $\langle \phi, \psi \rangle$ *relative to two epistemic states* $\mathcal{K}$ *and* $\mathcal{K}'$ *if and only if the following hold:*

BE1′ $\vec{X} = \vec{x}$ *is a partial explanation of* ϕ *in* $(M, \vec{u})$.

BE2′ $\vec{X} = \vec{y}$ *is a partial explanation of* ψ *in* $(M, \vec{u}')$.

BE3′ $(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$*—the difference condition.*

BE4′ $\vec{X}$ *is maximal—that is, there is no superset of* $\vec{X}$ *that satisfies BE1′-3′.*

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

THEOREM 7. *BE1-4 iff BE1′-4′—that is, the two definitions of contrastive bi-factual explanation are equivalent.*

Proof. The proof for this is similar to the proof for Theorem 6, except simpler because we deal only with factual situations and no hypothetical situations.

Left-to-right case: (BE1′ and BE2′) These both hold from BE2-4. If $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a sufficient counterfactual cause for $\langle \phi, \psi \rangle$ that is minimal and uncertain, then $\vec{X} = \vec{x}$ and $\vec{X} = \vec{y}$ must be partial explanations of ϕ and ψ under $(M, \vec{u})$ and $(M, \vec{u}')$ respectively. (BE3′) The difference condition holds because this is a requirement of sufficient bi-factual cause in BE2. (BE4′) holds from the maximality condition in BE2. This establishes the left-to-right case.

Right-to-left case: (BE1) This holds directly from BE1′ and BE2′, because the acceptance of ϕ in $\mathcal{K}$ and $\neg\psi$ in $\mathcal{K}'$ respectively are conditions of explanations under the original Halpern & Pearl definition. (BE2) If $\vec{X} = \vec{x}$ and $\vec{X} = \vec{y}$ are partial explanations of ϕ and ψ under $(M, \vec{u})$ under $(M, \vec{u}')$ respectively, then they must be partial causes too. If their intersection is empty and $\vec{X}$ is maximal, then this defines a sufficient cause, so BE2 holds. (BE3) We prove this via contradiction. Assume $\vec{X}$ is not minimal. This implies that there is some strict superset $\vec{Y} \supset \vec{X}$ that satisfies BE1′-4′ and BE2. However, if this were the case, then BE4′ would not hold: $\vec{X}$ would not be maximal over BE1′-3′, which is a contradiction, so our assumption is false. (BE4) If $\vec{X} = \vec{x}$ and $\vec{X} = \vec{y}$ are partial explanations of ϕ and ψ under $(M, \vec{u})$ and $(M, \vec{u}')$ respectively, then the agent must be uncertain about $\vec{X} = \vec{x}$ in $(M, \vec{u})$ and $\vec{X} = \vec{y}$ in $(M, \vec{u})$. This establishes the right-to-left case, and the theorem holds. □

6.4 *Non-contrastive general explanation*

The definitions provided in the previous section merely allow explanations in which the causal model is known to the explainee agent, but the agent is uncertain which context is the real context. A more general definition allows for explanations in which the agent is also uncertain about the causal model, and thus the explanation is about both the causal model and the context.

Halpern and Pearl (2005b) present an extended definition of explanation based on this idea. In this case, an epistemic state $\mathcal{K}$ is now a set of situations $(M, \vec{u})$ instead of a set of just contexts. A general explanation is of the form $(\alpha, \vec{X} = \vec{x})$, in which α is a causal formula. The first component restricts the set of models, while the second restricts the set of contexts.

A formula-event pair $(\alpha, \vec{X} = \vec{x})$ is an explanation of event ϕ relative to a set of situations $\mathcal{K}$ if:

EX1 $(M, \vec{u}) \models \phi$ for each $(M, \vec{u}) \in \mathcal{K}$ (unchanged).

EX2 for all situations $(M, \vec{u})$ such that $(M, \vec{u}) \models \vec{X} = \vec{x}$ and $M \models \alpha$ (α is valid in all contexts consistent with M), $\vec{X} = \vec{x}$ is a *sufficient cause* of ϕ.

EX3 $(\alpha, \vec{X} = \vec{x})$ is minimal—there is no pair $(\alpha', \vec{X}' = \vec{x}') \neq (\alpha, \vec{X} = \vec{x})$ satisfying EX2 such that $\{M'' \in M(\mathcal{K}) \mid M'' \models \alpha'\} \supseteq \{M'' \in M(\mathcal{K}) \mid M'' \models \alpha\}$ and $\vec{X}' = \vec{x}' \subseteq \vec{X} = \vec{x}$, where $M(\mathcal{K}) = \{M \mid (M, \vec{u}) \in \mathcal{K}$ for some $\vec{u}\}$.

EX4 $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ for some $(M, \vec{u}) \in \mathcal{K}$ and $(M', \vec{u}) \models \vec{X} = \vec{x}$ for some (other) $(M', \vec{u}') \in \mathcal{K}$—that is, the agent is uncertain, as before.

In this definition, the two parts of the explanation play different roles. The formula α characterizes the part of the model that is unknown to the agent to just enough information to understand the causes of ϕ; while ϕ is an explanation in that restricted set of models.

Example 6.4. As a simple example, consider an agent who does not know how the arthropod system works at all, and confronted with $O = Spider$, they ask why. An explanation is the pair:

$$(L = 8 \wedge S = 4 \wedge E = 8 \wedge C = 4 \wedge W = 0 \Rightarrow O = Spider, L = 8)$$

plus one for all other variables other than L. The formula informs the explainee what the properties of a spider are, but does not need to define the entire model nor even the properties of other arthropods.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

However, the α part of the explanation can be arbitrary causal formula. For example, given a 7-legged spider with no annotation, which will cause $V = Pass$, an explanation could refer to formula such as:

$$(O = Unknown \wedge A = Unknown) \Rightarrow [A \leftarrow Spider](V = Pass),$$

which means that when both variables are unknown, adding an annotation will still give a result of *Pass*.

6.5 General contrastive explanation

The more general case of contrastive explanation is straightforward to project from this definition. We give just the definition for bi-factual explanation.

DEFINITION 8. *General Contrastive Bi-factual Explanation Given a structural model M, a pair of formula-event pairs* $\langle(\alpha, \vec{X} = \vec{x}), (\beta, \vec{X} = \vec{y})\rangle$ *is a general contrastive bi-factual explanation of* $\langle \phi, \psi \rangle$ *relative to two epistemic states* $\mathcal{K}$ *and* $\mathcal{K}'$ *if and only if the following hold:*

BE1$_G$ $(M, \vec{u}) \models \phi$ *for each* $(M, \vec{u}) \in \mathcal{K}$ *and* $(M', \vec{u}') \models \psi$ *for each* $(M, \vec{u}') \in \mathcal{K}'$.

BE2$_G$ *for all situations* $(M, \vec{u})$ *such that* $(M, \vec{u}) \models \vec{X} = \vec{x}$ *and* $M \models \alpha$ *and all situations* $(M', \vec{u}')$ *such that* $(M', \vec{u}') \models \vec{X} = \vec{y}$ *and* $M' \models \beta$, $\langle(\alpha, \vec{X} = \vec{x}), (\beta, \vec{X} = \vec{y})\rangle$ *is a sufficient bi-factual cause of* $\langle \phi, \psi \rangle$.

BE3$_G$ $(\alpha, \vec{X} = \vec{x}, \beta, \vec{X} = \vec{y})$ *is minimal—there is no tuple* $(\alpha', \vec{X}' = \vec{x}', \beta', \vec{X} = \vec{y}') \neq (\alpha, \vec{X} = \vec{x}, \beta, \vec{X} = \vec{y})$ *satisfying BE2$_G$ such that* $\{M'' \in M(\mathcal{K}) \mid M'' \models \alpha'\} \supseteq \{M'' \in M(\mathcal{K}) \mid M'' \models \alpha\}$, *similarly for* β, $\vec{X}' = \vec{x}' \subseteq \vec{X} = \vec{x}$, *and* $\vec{X}' = \vec{y}' \subseteq \vec{X} = \vec{y}$.

BE4$_G$ $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ *for some* $(M, \vec{u}) \in \mathcal{K}$ *and* $(M', \vec{u}') \models \vec{X} = \vec{x}$ *for some (other)* $(M', \vec{u}') \in \mathcal{K}$; *and* $(M, \vec{u}) \models \vec{X} = \vec{y}$ *for some* $(M, \vec{u}) \in \mathcal{K}'$ *and* $(M, \vec{u}') \models \neg(\vec{X} = \vec{y})$ *for some (other)* $(M, \vec{u}') \in \mathcal{K}'$*—that is, the agent is initially uncertain whether the explanation is true or not.*

The general counterfactual explanation case is straightforward to extend from Definition 8, however, one important point of difference is that the explanation is not a pair, but a triple, $\langle \alpha, \vec{X} = \vec{y}, \vec{X} = \vec{y}\rangle$. There is no requirement for the second formula β because there is only one model to characterize.

Example 6.5. Consider the extended arthropod system in the situation in Example 6.1, where there is an image of a spider with 7 legs. In this case, the verification passes because $O = Unknown$. The agent knows all variables but is unaware of F_V, so does not know the verification procedure, and asks 'Why $V = Pass$ instead of $V = Fail$?'

In this case, the explanation is a formula expressing the semantics of F_V, and no variables:

$$\langle(O = Unknown \vee A = Unknown \Rightarrow V = Pass), L = 7, L = 8\rangle$$

6.6 Example: goal-directed AI planning

Throughout the paper, we have used the two examples of the arthropod system to illustrate ideas. In this section, we consider a different type of AI system: goal-directed planning.

Example 6.6. Consider an abstract example of a goal-directed planning system that needs to choose which actions A_1, A_2, and A_3 to apply. Using a simple action language, we define these actions, their preconditions, and their effects as:

Action	**Pre**		**Effect**
A_1	P_1	$\rightarrow$	$G_1 \wedge G_3$
A_2	P_2	$\rightarrow$	$G_2 \wedge G_3$
A_3	*true*	$\rightarrow$	P_2

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

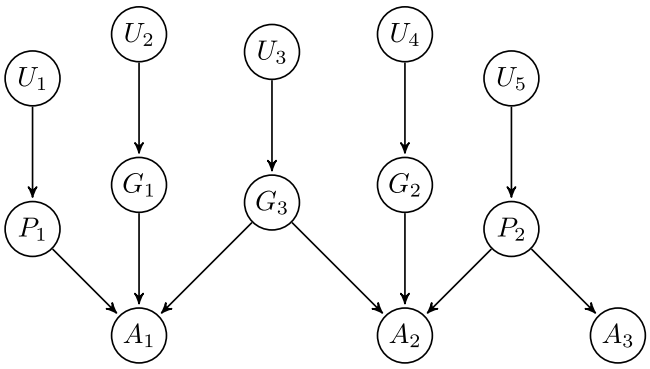


Figure 3 Causal graph for goal-directed planning.

in which $A_{[1-3]}$ are names of the actions, $G_{[1-3]}$ are propositions modeling goals, and $P_{[1-2]}$ are propositions modeling action preconditions. The planner can apply none, one, or many actions.

Figure 3 shows the causal graph for this, in which $U_{[1-5]}$ are exogenous variables. All variables are Boolean. Variables $P_{[1-2]}$ and $G_{[1-3]}$ represent preconditions and goals being true. Variables $A_{[1-3]}$ represent whether an action could be selected for execution given the goals and its precondition. The structural equations are such that action A_1 is selected if G_1 or G_3 is the goal, and its precondition P_1 holds; A_2 is selected if G_2 or G_3 is the goal, and its precondition P_2 holds; and A_3 is selected if precondition P_2 needs to be made true. Note that this does not model the cause of the preconditions goals becoming true/false, but the *cause* of action selection, which makes the graph appear somewhat inverted. The parent node for each action has both the variables it requires to be true to execute the action as well as the variables the action will change; e.g. A_1 will be 'fired' if P_1 is true and G_1 is true, which counter-intuitively models that in the actual planning problem, the goal is currently false and should *become* true. We could also add a node which states whether the goal is true/false and only execute the action if the goal is false, but we omit this for simplicity. Note that P_2 is the parent of A_3, modeling that this is A_3's intermediate 'goal'—it makes P_2 true, thus enabling A_2 to be selected next.

Now consider the case in which G_1 and G_3 are the goals (while G_2 is false) and P_1 and P_2 are both true, implying that A_1 is true and A_2 is false. A contrastive question could be: 'Why A_1 rather than A_2?', which would be modeled as 'Why $(M, \vec{u}) \models A_1 \wedge A_2$ rather than $\neg A_1 \wedge A_2$?'. $\mathcal{K}$ (the epistemic state of the explainee) is such that G_1 is known to be true, but the agent is unsure of the other goals and the preconditions.

The contrastive counterfactual cause for this is the pair $\langle(G_1, \neg G_1), (\neg G_2, G_2)\rangle$. That is, for the A_2 to be true instead of A_1 (CC3), it would require that the goals G_1 and G_2 are swapped. CC1-2 hold trivially, and CC4 (the difference conditions) holds because there are no common events. CC5 (maximality) holds because changing the values of G_1 or the preconditions P_1 and P_2 do not satisfy the difference condition CC4.

The contrastive *explanation*, however, consists only of $\langle(\neg G_2, G_2)\rangle$—the agent already knows that G_1 is true so including G_1 would not satisfy both CE3 (the minimality condition) and CE4 (the 'meaningful' condition).

Example 6.7. Consider a bi-factual setting with two situations $\vec{u}_1$ and $\vec{u}_2$. In both situations, G_3 is the only goal. In $\vec{u}_1$, precondition P_1 is true while P_2 is false, and vice-versa for $\vec{u}_2$. The explainee agent knows only that action A_1 was selected under $\vec{u}_1$ and A_2 was selected under $\vec{u}_2$. The bi-factual explanation for this is $\langle(P_1, \neg P_2), (\neg P_1, P_2)\rangle$. The goals are not included even though the agent does not know their values, because they are the same between the two situations, so do not satisfy the difference condition.

Example 6.8. Finally, consider the example of A_3 being selected in order to make P_2 true and allow A_2 to be selected in the next time step. The goal is G_2 and the explainee knows the values of all goal variables and action variables, does not know the values of the preconditions, and asks why A_3 rather than A_2.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

The effect of CE4 is that this has no explanation! Intuitively, one may expect that $(\neg P_2, P_2)$ to be offered, however for this to be an explanation, CE4 requires that there is some situation $\vec{u} \in \mathcal{K}$ in which P_2 could be true. But this is not possible because the agent knows that A_2 is false and G_2 is true, which cannot be the case if P_2 is true, so no such situation exists. According to the model M, there can be only situation in $\mathcal{K}$ where the goals are all known as $\neg G_1$, G_2, and $\neg G_3$, and in that situation $\neg P_1$ and P_2 hold. This offers the agent a complete explanation already. This makes sense: the agent does not require an explanation because it can infer the values of P_1 and P_2 itself.

However, consider the case of a general contrastive explanation in which the agent's knowledge is missing part of the structure of the causal graph; specifically, that P_2 is the precondition of A_2, meaning that the edge $P_2 \rightarrow A_2$ is missing from the graph in Figure 3. Now we have an explanation! In this case, the explanation is $\langle (F_{A_2} = f, \emptyset), (F_{A_2} = f', \neg P_2) \rangle$, in which f is the definition of F_{A_2} without the precondition, and f' includes the precondition.

7 Conclusion

Using structural causal models, Halpern and Pearl (2005b) define explanation as a fact that, if found to be true, would constitute an actual cause of a specific event. In this paper, we extend this definition of explanation to consider *contrastive explanations*. Founded on existing research in philosophy and cognitive science, we define two types of contrastive why-questions: counterfactual why-questions ('rather than') and bi-factual why-questions ('but'). We define 'contrastive cause' for these two questions and from this, build a model of contrastive explanation. We show that this model is consistent with well-accepted properties of contrastive explanation, and with alternative definitions.

The aim of this work is to provide a general model of contrastive explanation. While there are many examples of researchers considering counterfactual contrastive questions in XAI, few consider bi-factual questions. Even fewer exploit the power of the difference condition, instead providing two full explanations: one of the fact and one of the foil. In essence, they consider contrastive questions but not contrastive explanations. The difference condition is what brings power and relevance to contrastive explanations, and as such, giving two complete explanations does not correctly answer the question. We hope that this article serves as a basis for researchers in XAI to adopt the idea of the difference condition and ultimately give better explanations to people.

References

Akula, A. R., Wang, S. & Zhu, S.-C. 2020. CoCoX: Generating conceptual and counterfactual explanations via fault-lines. In *AAAI*, 2594–2601.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. 2016. Machine bias. *ProPublica, May*, 23.

Bromberger, S. 1966. Why–questions. In *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, Colodny, R. G. (ed.). Pittsburgh University Press, 68–111.

Buchanan, B. & Shortliffe, E. 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

Chandrasekaran, B., Tanner, M. C. & Josephson, J. R. 1989. Explaining control strategies in problem solving. *IEEE Expert* **4** (1), 9–15.

Chin-Parker, S. & Cantelon, J. 2017. Contrastive constraints guide explanation-based category learning. *Cognitive Science* **41** (6), 1645–1655.

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. *In Advances in Neural Information Processing Systems*, 592–603.

Garfinkel, A. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory*. Yale University Press.

Grice, H. P. 1975. Logic and conversation. In *Syntax and Semantics 3: Speech Arts*. Academic Press, 41–58.

Halpern, J. Y. 2015. A modification of the Halpern-Pearl definition of causality. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 3022–3033.

Halpern, J. Y. & Pearl, J. 2005a. Causes and explanations: A structural-model approach. *part i: Causes. The British Journal for the Philosophy of Science* **56**(4), 843–887.

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

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press

Haynes, S. R., Cohen, M. A. & Ritter, F. E. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* **67** (1), 90–110.
Hesslow, G. 1983. Explaining differences and weighting causes. *Theoria* **49** (2), 87–111.
Hesslow, G. 1988. The problem of causal selection. *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, 11–32.
Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin* **107** (1), 65–81.
Kean, A. 1998. A characterization of contrastive explanations computation. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 599–610.
Krarup, B., Cashmore, M., Magazzeni, D. & Miller, T. 2019. Model-based contrastive explanations for explainable planning. In *2nd ICAPS Workshop on Explainable Planning (XAIP-2019)*. AAAI Press.
Lewis, D. 1986. Causal explanation. *Philosophical Papers* **2**, 214–240.
Lim, B. Y. & Dey, A. K. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. ACM, 195–204.
Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B. & Lee, J. D. 2006. Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 500 (23), 0 2482–2486.
Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* **27**, 247–266.
Madumal, P., Miller, T., Sonenberg, L. & Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2493–2500.
Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D. & Procci, K. 2016. Intelligent agent transparency in human–agent teaming for multi-UxV management. *Human Factors* **580** (3), 401–415.
Miller, T. 2018. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. https://arxiv.org/abs/1706.07269.
Mothilal, R. K., Sharma, A. & Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.
Ruben, D.-H. 1987. Explaining contrastive facts. *Analysis* **470** (1), 35–37.
Slugoski, B. R., Lalljee, M., Lamb, R. & Ginsburg, G. P. 1993. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology* **230** (3), 219–238.
Sreedharan, S., Srivastava, S. & Kambhampati, S. 2018. Hierarchical expertise level modeling for user specific contrastive explanations. *In IJCAI*, 4829–4836.
Stubbs, K., Hinds, P. & Wettergreen, D. 2007. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intell. Syst.* **22** (2), 42–50.
Swartout, W. R. & Moore, J. D. 1993. Explanation in second generation expert systems. In *Second Generation Expert Systems*. Springer, 543–585.
Temple, D. 1988. The contrast theory of why–questions. *Philosophy of Science* **55**(1), 141–151.
Van Bouwel, J. & Weber, E. 2002. Remote causes, bad explanations? *Journal for the Theory of Social Behaviour* **320** (4), 437–449.
Van Fraassen, B. C. 1980. *The Scientific Image*. Oxford University Press.
Waa, J., van Diggelen, J., Bosch, K. & Neerincx, M. 2018. Contrastive explanations for reinforcement learning in terms of expected consequences. In *Proceedings of the Workshop on Explainable AI at IJCAI.*
Wachter, S., Mittelstadt, B. & Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* **31**, 841.
Wang, D., Yang, Q., Abdul, A. & Lim, B. Y. 2019. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.
Winikoff, M. 2017. Debugging agent programs with Why?: Questions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, *AAMAS 2017*. IFAAMAS, 251–259.
Ylikoski, P. 2007. The idea of contrastive explanandum. In *Rethinking Explanation*. Springer, 27–42.

https://doi.org/10.1017/S0269888921000102 Published online by Cambridge University Press