Contrastive Explanation: A Structural-Model Approach

Tim Miller

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

Abstract

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 and action. While different sub-fields have looked into this problem with a sub-field-specific view, there are few models that aim to capture explanation in AI 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 AI problems: classification and planning. We believe that this model can be used to define contrastive explanation of other subfield-specific AI models.

Contents

1	Introduction	2				
2	Related Work 2.1 Philosophical Foundations					
	2.2 Computational Approaches	7				
3	Structural Models	8				
	3.1 Models	8				
	3.2 Language	9				
	3.3 Semantics					
4	Contrastive 'Why' Questions	10				
	4.1 Alternative Explananda	11				
	4.2 Congruent Explananda	19				

5	Cor	ntrastive Cause	13
	5.1	Non-contrastive Cause	13
	5.2	Contrastive Causes in Alternative Explananda	14
	5.3	Contrastive Causes in Congruent Explananda	16
	5.4	Presuppositions	19
6	Cor	nstrastive Explanation	21
	6.1	Non-Contrastive Explanation	21
	6.2	Contrastive Alternative Explanation	22
	6.3	Congruent Contrastive Explanation	23
	6.4	Non-Contrastive General Explanation	25
	6.5	General Contrastive Explanation	26
	6.6	Example: Goal-Directed AI Planning	26
7	Cor	nclusion	28

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 [14, 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 [3, 4, 29]. This resurgence is driven by lack of trust from users [27, 18, 20], and also concerns regarding the ethical implications of decisions made by 'black box' algorithms [1].

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 this work has centred around causal attribution: the process of extracting the causes (or main causes) of a decision or action; for example, LIME [22] is a system for extracting simplified, local explanations from black-box classifiers. While causal attribution 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 [21] 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's do not ask "Why P?", but "Why P rather than Q?", although often Q is implicit in the context. Following Lipton [19], 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 [14, 31, 19, 5]. This provides a 'window' into the questioner's mental model, identifying what they do not know [16]. Second, giving contrastive *explanations* is simpler, more feasible, and cognitively less demanding to both questioner and explainer [16, 19, 35].

Lewis argues that a contrastive question "requests information about the features that differentiate the actual causal history from its counterfactual alternative." [16, p. 231].

Lipton [19] defines the answer to a contrastive question as the Difference Condition:

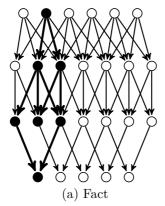
To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q. – Lipton [19, p. 256].

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 colour — 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 colouring 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. Consider the abstract causal graphs in Figure 1, which are larger but still trivial by AI standards. The graph on the left is of the factual case, while the graph on the right is of the contrast case. The dark nodes and thick vertices are the difference between the two graphs — the light nodes and thin vertices are the same. It shows that changing the second input explains a difference value for the first output. A full attribution would explain all 25 causes in both graphs. However, the difference condition merely needs us to describe those darker regions, which is a simpler task. Importantly, note that the differences are symmetric: we need only explain one set of differences, rather than both graphs independently. As Lipton [19] notes, facts and contrast cases are often closely related, meaning that the difference is often small. For example, "Why crow rather than magpie?" is more likely than "Why crow rather than emu?".

In this paper, we extend Halpern and Pearl's definition of explanation using structural causal models [9] 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: alternative questions and congruent questions. An alternative question is of the form "Why P rather than Q?", and asks why some fact P occurred instead of some hypothetical $foil\ Q$. A congruent 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 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.



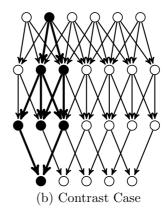


Figure 1: Contrastive Explanation of Casual Graphs Using the Difference Condition.

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: classification and goal-directed planning.

2. Related Work

2.1. Philosophical Foundations

In the social sciences, it is generally accepted that explanations are contrastive [21]. The questions that people ask have a contrast case, which is often 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 [6] seems to be the first to make a case for contrastive explanation¹. 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/other things], rather than why he [robs/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" [6, p. 21].

At the same time, Van Fraassen [33] 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 [33, p. 127]. An answer to such a question has the structure "P in contrast to (the rest of) X because A" [33, p. 143]. Van Fraassen argues that when a questioner asks such a question, they presuppose that: (1) the topic P is true; (2) all

¹Although Van Fraassen [33, p. 127] attributes the idea of contrastive explanation to Bengt Hannson in an unpublished manuscript circulated in 1974.

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 [12, 13] 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 [13, 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 formalises 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 iff (1) if the cause C held on all the other objects x in R (e.g. other barns), then the property E would also hold (the other barns would have also caught fire); and (2) if 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 [16] 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 [16, p. 229–230]. This has parallels with Hesslow's account [12, 13].

Temple [30] subsequently argued against the case of contrastive explanation. He 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 whyquestions are just standard propositional whyquestions of the form "Why X?", but with X being [P and not- Q].

However, Lipton [19] 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 [P]?". Building on Lewis's interpretation based on the history of events [16], Lipton argues that answering "Why [P] and not [P]?" requires an explanation of [P] and of not-[P]. 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 [19] 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 [19, p. 256].

This differs from the definition of contrastive explanation from Lewis [16] 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 formalise this as the following, in which \leadsto 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:

Thus, Lewis's definition [16] cites some alternative history of facts in which Q occurred, whereas Lipton's definition [19] 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 [19] proposed his ideas, that facts and contrast case are incompatible events [30, 6, 33, 23]; 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 one 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 [35] 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 of the same process; one the fact and one the 'imagined' foil. 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. 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 stated foil, which is consist with the idea from Lipton [19] that this is cognitively a simpler problem.

Van Bouwel and Weber [31] 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 [35].

In Section 4, we will formalise the notion of contrastive questions using the framework of Halpern and Pearl [8], and will show that the reasoning of Ylikoski [35] is natural with respective to structural equations and fits the types of questions we would expect in explainable artificial intelligence. The concept of *P-contrast* is captured as an *alternative* explanations, while *O-contrast* and *T-contrast* are captured as *congruent* 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 [17] showed that "Why not?" questions are important in context-aware applications, while Haynes et al. [10] 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 contrastive 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?" [3], which is providing a foil for the fact. More recently, approaches from various sub-disciplines of artificial intelligence have also defined that contrastive why–questions are important. For example, Winikoff's BDI program debugging system supports questions such as "Why don't you believe ..." and classification tasks that are compared to a contrast case [28], and counterfactual explanations with natural language [11] to reference just a few. However, few such models make explicit use of computing the difference between the fact and the contrast case, with some exceptions; for example, see [24, 32, 26]. Providing two complete explanations does not take advantage of the difference condition, producing larger and less relevant explanations.

The aim of this paper is to provide a general computational model of contrastive explanation that can be mapped to other models in artificial intelligence, such as machine learning, planning, reinforcement learning, case-based reasoning, BDI agents, etc. As far as the author is aware, Kean [15] 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 [19]. Given a knowledge base 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 as well as 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, 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. In contrast, our model assumes the causal graph is known to the explainer, and the task is to find a contrastive explanation to an unaware explainee.

3. Structural Models

In this paper, we build definitions of contrastive questions and contrastive explanations based on Halpern and Pearl's *structural models* [8]. As opposed to previous models, which use logical implication or statistic relevance, Halpern and Pearl's definition is based on counterfactuals, modelled using *structural equations*.

In Part I [8] 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 S is a structure (U, V, R), in which U is a set of exogenous variables, V a set of endogenous variables, and R is a function that defines the range of values for every variable $Y \in U \cup V$; that is, the range of a variable Y is 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 $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \to \mathcal{R}(X)$ determines the value of X based on other variables in the model. A causal model is said to be recursive if it is acyclic.

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 and Pearl [8] extend this basic structural equation model to support modelling 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 [8] 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 and 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 $M_{\vec{Y} \leftarrow \vec{u}}$ is x, and defining Boolean combinations in the standard way.

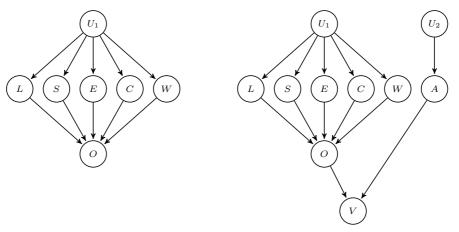
Example 3.1. This section presents a simple example of a hypothetical system that classifies images of arthropods into several different types, taken from Miller [21]. The categorisation is based on certain physical features of the arthropods, such as number of legs, number of eyes, number of wings, etc. Table 1 outlines a simple model of the features of arthropods for illustrative purposes.

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

Table 1: A simple lay model for distinguishing common arthropods.

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 2a. The functions are clear from Table 1; for example, $F_O(8, false, 8, no, 0) = Spider$, and O = Unknown for anything not in the table.

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



(a) Causal graph for arthropod algorithm de- (b) Causal graph for extended arthropod alfined in Example 3.1 gorithm defined in Example 3.2

one to check whether the manual annotations are correct or not. The same categories exist, but some images are not labelled 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 2b. 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 [2], 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 [35]'s argument for the existence of (at least) two different types of contrastive why—questions [35]. 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 call 'rather than' or alternative explananda, because in this case, the foil is an alternative 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" [30]. The second type, we call *congruent* explananda, 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 and Pearl's structural models [8], 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. Alternative Explananda

Given two events P and Q, Lipton [19] defines a contrastive why-question as:

Why
$$P$$
 rather than Q ? (1)

For an alternative explananda, 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-formalise this in structural models: an alternative why–question, given a situation (M, \vec{u}) , is:

Why
$$(M, \vec{u}) \models \phi$$
 rather than ψ ? (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:

Why
$$(M, \vec{u}) \models \phi \land \neg \psi$$
?, (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 counter-example for ψ . However, as argued by Lipton [19], 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 variables². 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 alternative explananda 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.

An assumption of alternative explananda is that 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

²Although 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.

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 because it must be that Y=5 holds. So, the question is really "Why $X=4 \land Y=4$ rather than $X=4 \land Y=5$?", 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.

4.2. Congruent Explananda

As outlined in Section 2, Ylikoski [35] 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 *congruent* explananda, by modelling the two different processes as two different situations:

Why
$$(M, \vec{u}) \models \phi$$
 but $(M', \vec{u}') \models \psi$? (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. 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. 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?":

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

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

Why
$$(M, \vec{u}) \models \phi \land \neg \psi$$
 but $(M', \vec{u}') \models \psi \land \neg \phi$? (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.

In this section, we have demonstrated a case for two types of contrastive whyquestion: alternative and congruent explananda. 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 [21] for a discussion of these. In Section 6, we build on the definition of explanation based on epistemic relevance by Halpern and Pearl [9]. 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 consistent with existing philosophical views; e.g. Ruben [23] defines contrastive explanations as conjunctions between history of the contrasting events. The particular definition depends whether the why–question is alternative or congruent.

5.1. Non-contrastive Cause

Our definition of contrastive cause extends Halpern and Pearl's definition of actual cause [8]. 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 [8] 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** $(M, \vec{u}) \models \vec{X} = \vec{x} \land \phi$ 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 occurred³.

³Note that this is the later definition from Halpern [7], which is simplified compared to the original definition of Halpern and Pearl [9]. Halpern argues this updated definition is more robust.

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 in Alternative Explananda

To define contrastive cause, we adopt and formalise Lipton's Difference Condition [19], which states that we should find causes that are different in the 'history' of the two events. We define the 'history' as the situation (M,\vec{u}) under which the events are evaluated; that is, (M,u) for alternative why–questions, and both (M,u) and (M',\vec{u}') for congruent 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 alternative 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 Alternative Cause). A pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is an contrastive alternative actual cause (also just a alternative cause) of $\langle \phi, \psi \rangle$ in situation (M, \vec{u}) if and only if the following conditions holds:

CAC1 $\vec{X} = \vec{x}$ is a partial cause of ϕ under (M, \vec{u}) .

CAC2 $(M, \vec{u}') \models \neg \psi$ — the foil ψ is not true.

CAC3 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.

CAC4 $(\vec{X} = \vec{x} \cap \vec{X} = \vec{y}) = \emptyset$ — that is, there are no common events. This is the difference condition.

CAC5 \vec{X} is maximal — that is, no superset of \vec{X} satisfies CAC1-4.

Similar to the HP definition, we can define *sufficient* contrastive cause by modifying CAC1 and CAC3 to refer to partial sufficient causes.

This definition is based on the Halpern [7] definition of actual cause, as conditions CAC1-3 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 [8], could be substituted, and this would change the semantic interpretation of the above.

The reader may expect to see that CAC2 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 CAC4, 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 CAC3 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 categorised as a Bee instead of a Fly. To answer the alternative 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:

$$\langle S = \mathbf{V}, S = \mathbf{X} \rangle$$

 $\langle W = 4, W = 2 \rangle.$

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 CAC4.

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 *adequate* explanation defined by Hesslow [12] (see Section 2.1). This states that, if the alternative causes hold in a different situation, so to would the alternative events. The following theorem captures this.

Theorem 1. If C comprises all alternative contrastive actual causes of $\langle \phi, \psi \rangle$ under situation (M, \vec{u}) , then for any maximal-consistent subset⁴ $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle \subseteq C$:

- (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 CAC3, $\vec{X} = \vec{y}$ 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 CAC3 and CAC4 hold for $\vec{Z}' = \vec{z}'$. CAC5 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 Congruent Explananda

For congruent explananda, the definition of 'history' is different to that of alternative explananda, 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 Congruent Cause — Simple Case). A pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a *congruent contrastive actual cause* of $\langle \phi, \psi \rangle$ in their respective situations (M, \vec{u}) and (M, \vec{u}') if:

CCC1 $\vec{X} = \vec{x}$ is a partial cause of ϕ under (M, \vec{u}) .

CCC2 $\vec{X} = \vec{y}$ is a partial cause of ψ under (M, \vec{u}') .

CCC3 $(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$ — that is, there are no common events. This is the difference condition.

CCC4 \vec{X} is maximal — that is, no superset of \vec{X} satisfies CCC1-3.

Note that CCC1 implies $(M, \vec{u}) \models \vec{X} = \vec{x} \land \phi$ (AC1) and similarly for CCC2.

A *sufficient* contrastive cause can be obtained by modifying CCC1 and CCC2 to refer to partial sufficient causes.

This definition is simpler than that of alternative explanation (compare CAC3 with CCC2), because both the fact and surrogate are actual events, whereas in alternative explananda, the foil is hypothetical.

⁴We 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.

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 categorised 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:

$$\langle S = \checkmark, S = \checkmark \rangle$$

 $\langle W = 4, W = 2 \rangle.$

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

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

- (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 alternative causes in Theorem 1, except that we refer to CCC2-4 instead of CAC3-5.

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 $M^{\mathcal{F}'} = (\mathcal{S}, \mathcal{F}')$, 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 Congruent Cause — General Case). A pair $\langle (F^{\phi}, \vec{X} = \vec{x}), (F^{\psi}, \vec{Y} = \vec{y}) \rangle$ is a congruent contrastive 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:

CCC1 $(F^{\phi}, \vec{X} = \vec{x})$ is a partial restricted cause of ϕ under situation (M, \vec{u}) .

CCC2 $(F^{\psi}, \vec{Y} = \vec{y})$ is a partial restricted cause of ψ under situation (M', \vec{u}') .

CCC3 $F^{\phi} \cap F^{\psi} = \emptyset$ 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.

CCC4 $(F^{\phi}, \vec{X} = \vec{x}, F^{\psi}, \vec{Y} = \vec{y}, \vec{X} \cap \vec{Y})$ 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 CCC1, CCC2, and CCC4 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 CCC4 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 $(M, \vec{u}) \models O = Unknown$ and $(M', \vec{u}') \models O = Bee$, in which \vec{u} and \vec{u}' both correspond to features of a Bee but L = 5 (five legs) and A = Bee in \vec{u}' , a contrastive cause would be:

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

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 = Bee as the difference condition.

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 3, 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* alternative 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:

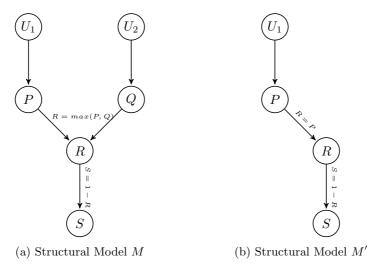


Figure 3: Structural Models for for Example 5.6

(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 its 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 show that our definition is consistent with the the idea of contrastive explanation as *presupposed explanation* [19].

Lipton [19] 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." [19, 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 an alternative explanandum as:

Assuming
$$M \models (\phi \oplus \psi)$$
, why $(M, \vec{u}) \models \phi$? (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 an alternative 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 CAC1-5 hold for some \vec{y} , and vice-versa.

The left-to-right case: For CAC1, 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 CAC2, $(M, \vec{u}) \models \neg \psi$ 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 CAC3-5 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 CAC3. In addition, the values in \vec{y} must make ψ true, and therefore must be different from the values in \vec{x} , so CAC4 holds. Finally, we prove CAC5 (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 CAC1: $\vec{X} = \vec{x}$ 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 CAC3: 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 CAC3. 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.

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 an congruent contrastive 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. Constrastive Explanation

Now that we have defined contrastive cause, we can define contrastive explanation. This is a simple extension to the existing definition of Halpern and Pearl [8]'s definition, but using contrastive causes instead of standard actual causes.

6.1. Non-Contrastive Explanation

In Part II [9] of their paper, Halpern and 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 [25].

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 probable⁵.

Halpern and Pearl [9] formalise 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.

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

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.

6.2. Contrastive Alternative Explanation

We extend the above definition to contrastive alternative 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 Alternative Explanation). Given a structural model M, a pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a *contrastive alternative explanation* of $\langle \phi, \psi \rangle$ relative to \mathcal{K} if and only if the following hold:

AEX1 $(M, \vec{u}) \models \phi \land \neg \psi$ for each $\vec{u} \in \mathcal{K}$ — that is, the agent accepts that ϕ and that $\neg \psi$.

AEX2 $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a sufficient alternative cause for $\langle \phi, \psi \rangle$, for each $\vec{u} \in \mathcal{K}$ such that $(M, \vec{u}) \models \vec{X} = \vec{x}$.

AEX3 \vec{X} is minimal — no subset of \vec{X} satisfies AEX2.

AEX4 $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ 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 an alternative contrastive explanation as finding part of an alternative cause that satisfies the conditions AEX1-4. 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 Alternative 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 alternative explanation* of $\langle \phi, \psi \rangle$ relative to \mathcal{K} if and only if the following hold:

AEX1' $\vec{X} = \vec{x}$ is a partial explanation of ϕ in (M, \vec{u}) .

AEX2' 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 $(M_{\vec{W} \leftarrow \vec{v}}, \vec{u})$.

AEX3' $(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$ — the difference condition.

AEX4' \vec{X} is maximal — that is, there is no superset of \vec{X} that satisfies AEX1'-3'.

Theorem 6. AEX1-4 *iff* AEX1-4' — that is, the two definitions of contrastive alternative explanation are equivalent.

Proof. Left-to-right case: (AEX1') This holds from AEX2-4. If $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a sufficient alternative 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. (AEX2') 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 AEX2; and therefore, this hypothetical situation is a witness for AEX2'. (AEX3') The difference condition holds because this is a requirement of AEX2. (AEX4') holds from the maximality condition in AEX2. This establishes the left-to-right case.

Right-to-left case: (AEX1) This holds directly from AEX1', because the acceptance of ϕ in \mathcal{K} is a condition of an explanation under the original Halpern and Pearl definition, and ψ must be false whenever ϕ is true. (AEX2) 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 AEX2 holds. (AEX3) 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 AEX1'-4' and AEX2. However, if this were the case, then AEX4' would not hold: \vec{X} would not be maximal over AEX1'-3', which is a contradiction, so our assumption is false. (AEX4) 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. Congruent Contrastive Explanation

For the congruent 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 Congruent Explanation – Simple Case). Given a structural model M, a pair $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a contrastive congruent explanation of $\langle \phi, \psi \rangle$ relative to two epistemic states \mathcal{K} and \mathcal{K}' if and only if the following hold:

- **CEX1** $(M, \vec{u}) \models \phi$ 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}') .
- **CEX2** 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 an sufficient congruent cause for $\langle \phi, \psi \rangle$ under (M, \vec{u}) and (M', \vec{u}') .
- **CEX3** \vec{X} is minimal that is, no superset of \vec{X} satisfies CEX2.
- **CEX4** $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ 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 AEX, except that the rules refer to an actual situation \vec{u}' , rather than the hypothetical situation implied by AEX2. 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 = Unknown under (M, \vec{u}_7) and O = 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 = 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 alternative explanation.

Definition 6 defines congruent explanation as finding part of an alternative cause that satisfies the conditions CEX1-4. However, we can think of congruent 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 Congruent Explanation – Simple Case, Alternative Definition). Given a structural model M, a pair of events $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a *contrastive congruent explanation* of $\langle \phi, \psi \rangle$ relative to two epistemic states \mathcal{K} and \mathcal{K}' if and only if the following hold:

CEX1' $\vec{X} = \vec{x}$ is a partial explanation of ϕ in (M, \vec{u}) .

CEX2' $\vec{X} = \vec{y}$ is a partial explanation of ψ in (M, \vec{u}') .

CEX3' $(\vec{X} = \vec{x}) \cap (\vec{X} = \vec{y}) = \emptyset$ — the difference condition.

CEX4' \vec{X} is maximal — that is, there is no superset of \vec{X} that satisfies CEX1'-3'.

Theorem 7. CEX1-4 iff CEX1'-4' — that is, the two definitions of contrastive congruent 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. \Box

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 [9] 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 $M(\mathcal{K}) = \{M \mid (M, \vec{u}) \in \mathcal{K} \text{ for some } \vec{u}\}.$
- **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 characterises 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:

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

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.

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:

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

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 congruent explanation.

Definition 8. General Contrastive Congruent 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 congruent explanation of $\langle \phi, \psi \rangle$ relative to two epistemic states \mathcal{K} and \mathcal{K}' if and only if the following hold:

- **CEX1** $(M, \vec{u}) \models \phi$ for each $(M, \vec{u}) \in \mathcal{K}$ and $(M', \vec{u}') \models \psi$ for each $(M, \vec{u}') \in \mathcal{K}'$.
- **CEX2** 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 $M' \models \beta$, $\langle \vec{X} = \vec{x}, \vec{X} = \vec{y} \rangle$ is a *sufficient congruent cause* of $\langle \phi, \psi \rangle$.
- **CEX3** $(\alpha, \vec{X} = \vec{x}, \beta, \vec{X} = \vec{y})$ 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 CEX2 such that $\{M'' \in M(\mathcal{K}) \mid M'' \models \alpha'\} \supseteq \{M'' \in M(\mathcal{K}) \mid M'' \models \alpha\}$, similarly for $\beta, \vec{X}' = \vec{x}' \subseteq \vec{X} = \vec{x}$, and $\vec{X}' = \vec{y}' \subseteq \vec{X} = \vec{y}$.
- **CEX4** $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ 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 alternative 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 characterise.

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

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

Figure 4 shows the causal graph for this, in which $U_{[1-5]}$ are exogenous variables. Variables are Boolean. The structured 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 , modelling that this is A_3 's intermediate 'goal' – it makes P_2 true, thus enabling A_2 to be selected next.

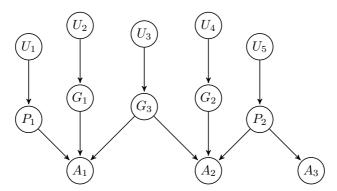


Figure 4: Causal graph for goal-directed planning

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 modelled as "Why $(M, \vec{u}) \models A_1 \land A_2$ rather than $\neg A_1 \land A_2$?". \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 alternative contrastive 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 (CAC3), it would require that the goals G_1 and G_2 are swapped. CAC1-2 hold trivially, and CAC4 (the difference conditions) holds because there are no common events. CAC5 (maximality) holds because changing the values of G_1 or the preconditions P_1 and P_2 do not satisfy the difference condition CAC4.

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 AEX3 (the minimality condition) and AEX4 (the 'meaningful' condition).

Example 6.7. Consider a congruent 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 viceversa 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 congruent 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 .

The effect of AEX4 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, AEX4 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 $P_2 \to A_2$ is missing from the graph in Figure 4. 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 [9] 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: alternative why-questions ('rather than') and and congruent 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 alternative contrastive questions in explainable artificial intelligence, few consider congruent 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 explainable

artificial intelligence to adopt the idea of the difference condition and ultimately give better explanations to people.

References

- [1] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23.
- [2] S. Bromberger, Why–questions, in: R. G. Colodny (Ed.), Mind and Cosmos: Essays in Contemporary Science and Philosophy, Pittsburgh University Press, Pittsburgh, 68–111, 1966.
- B. Buchanan, E. Shortliffe, Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project, Addison-Wesley, 1984.
- [4] B. Chandrasekaran, M. C. Tanner, J. R. Josephson, Explaining control strategies in problem solving, IEEE Expert 4 (1) (1989) 9–15.
- [5] S. Chin-Parker, J. Cantelon, Contrastive Constraints Guide Explanation-Based Category Learning, Cognitive science 41 (6) (2017) 1645–1655.
- [6] A. Garfinkel, Forms of explanation: Rethinking the questions in social theory, Yale University Press New Haven, 1981.
- [7] J. Y. Halpern, A Modification of the Halpern-Pearl Definition of Causality, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015), 3022–3033, 2015.
- [8] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach. Part I: Causes, The British Journal for the Philosophy of Science 56 (4) (2005) 843–887.
- [9] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach. Part II: Explanations, The British Journal for the Philosophy of Science 56 (4) (2005) 889–911.
- [10] S. R. Haynes, M. A. Cohen, F. E. Ritter, Designs for explaining intelligent agents, International Journal of Human-Computer Studies 67 (1) (2009) 90-110.
- [11] L. A. Hendricks, R. Hu, T. Darrell, Z. Akata, Generating Counterfactual Explanations with Natural Language, in: ICML Workshop on Human Interpretability in Machine Learning, 2018.
- [12] G. Hesslow, Explaining differences and weighting causes, Theoria 49 (2) (1983) 87–111.
- [13] G. Hesslow, The problem of causal selection, Contemporary science and natural explanation: Commonsense conceptions of causality (1988) 11–32.
- [14] D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin 107 (1) (1990) 65–81.
- [15] A. Kean, A characterization of contrastive explanations computation, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 599–610, 1998.
- [16] D. Lewis, Causal explanation, Philosophical Papers 2 (1986) 214–240.
- [17] B. Y. Lim, A. K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th international conference on Ubiquitous computing, ACM, 195–204, 2009.
- [18] M. P. Linegang, H. A. Stoner, M. J. Patterson, B. D. Seppelt, J. D. Hoffman, Z. B. Crittendon, J. D. Lee, Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50 (23) (2006) 2482–2486.
- [19] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27 (1990) 247–266.
- [20] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, K. Procci, Intelligent agent transparency in human-agent teaming for Multi-UxV management, Human Factors 58 (3) (2016) 401–415.
- [21] T. Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, Artificial Intelligence https://arxiv.org/abs/1706.07269.
- [22] M. T. Ribeiro, S. Singh, C. Guestrin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1135–1144, 2016.
- [23] D.-H. Ruben, Explaining contrastive facts, Analysis 47 (1) (1987) 35–37.
- [24] J. Seah, J. Tang, A. Kitchen, J. Seah, Generative Visual Rationales, arXiv e-prints 1804.04539.
- [25] B. R. Slugoski, M. Lalljee, R. Lamb, G. P. Ginsburg, Attribution in conversational context: Effect of mutual knowledge on explanation-giving, European Journal of Social Psychology 23 (3) (1993) 219–238.
- [26] S. Sreedharan, S. Srivastava, S. Kambhampati, Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations., in: IJCAI, 4829–4836, 2018.
- [27] K. Stubbs, P. Hinds, D. Wettergreen, Autonomy and common ground in human-robot interaction: A field study, IEEE Intelligent Systems 22 (2) (2007) 42–50.

- [28] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, arXiv e-prints 1703.01365.
- [29] W. R. Swartout, J. D. Moore, Explanation in second generation expert systems, in: Second Generation Expert Systems, Springer, 543–585, 1993.
- [30] D. Temple, The contrast theory of why–questions, Philosophy of Science 55 (1) (1988) 141–151.
- [31] J. Van Bouwel, E. Weber, Remote causes, bad explanations?, Journal for the Theory of Social Behaviour 32 (4) (2002) 437–449.
- [32] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, M. Neerincx, Contrastive Explanations with Local Foil Trees, ArXiv e-prints 1806.07470.
- [33] B. C. Van Fraassen, The scientific image, Oxford University Press, 1980.
- [34] M. Winikoff, Debugging Agent Programs with Why?: Questions, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS '17, IFAAMAS, 251–259, 2017.
- [35] P. Ylikoski, The idea of contrastive explanandum, in: Rethinking explanation, Springer, 27–42, 2007