In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, San Francisco, CA: Morgan Kaufmann, 2001.

TECHNICAL REPORT
R-266-IJCAI
June 2001

Causes and Explanations: A Structural-Model Approach— Part II: Explanations

Joseph Y. Halpern*
Cornell University
Dept. of Computer Science
Ithaca, NY 14853
halpern@cs.cornell.edu
www.cs.cornell.edu/home/halpern

Judea Pearl†
Dept. of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
judea@cs.ucla.edu
wwww.cs.ucla.edu/∼judea

Abstract

We propose a new definition of *(causal) explanation*, using *structural equations* to model counterfactuals. The definition is based on the notion of *actual cause*, as defined and motivated in a companion paper. Essentially, an explanation is a fact that is not known for certain but, if found to be true, would constitute an actual cause of the fact to be explained, regardless of the agent's initial uncertainty. We show that the definition handles well a number of problematic examples from the literature.

1 Introduction

The automatic generation of adequate explanations is a task essential in planning, diagnosis and natural language processing. A system doing inference must be able to explain its findings and recommendations to evoke a user's confidence. However, getting a good definition of explanation is a notoriously difficult problem, which has been studied for years. (See [Chajewska and Halpern, 1997; Gärdenfors, 1988; Hempel, 1965; Pearl, 1988; Salmon, 1989] and the references therein for an introduction to and discussion of the issues.)

In [Halpern and Pearl, 2001], we give a definition of actual causality using structural equations. Here we show how the ideas behind that definition can be used to give an elegant definition of (causal) explanation that deals well with many of the problematic examples discussed in the literature. The basic idea is that an explanation is a fact that is not known for certain but, if found to be true, would constitute an actual cause of the *explanandum* (the fact to be explained), regardless of the agent's initial uncertainty.

2 Causal Models: A Review

To make this paper self-contained, this section repeats material from [Halpern and Pearl, 2001]; we review the basic definitions of causal models, as defined in terms of structural equations, and the syntax and semantics of a language for reasoning about causality and explanations. See [Galles and Pearl, 1997; Halpern, 2000; Pearl, 2000] for more details, motivation, and intuition.

Causal Models: The basic picture is that the world is described by random variables, some of which may have a causal influence on others. This influence is modeled by a set of *structural equations* Each equation represents a distinct mechanism (or law) in the world, which may be modified (by external actions) without altering the others. In practice, it seems useful to split the random variables into two sets, the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are determined by the endogenous variables. It is these endogenous variables whose values are described by the structural equations.

More formally, a *signature* $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a finite set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y. A *causal* (or *structural*) *model* over signature $\mathcal{S}$ is a tuple $M = (\mathcal{S}, \mathcal{F})$, where $\mathcal{F}$ associates with each variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. F_X tells us the value of X given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.

Example 2.1: Suppose that we want to reason about a forest fire that could be caused by either lightning or a match lit by an arsonist. Then the causal model would have the following endogenous variables (and perhaps others):

- F for fire ($F = 1$ if there is one, $F = 0$ otherwise)
- L for lightning ($L = 1$ if lightning occurred, $L = 0$ otherwise)
- *ML* for match lit (*ML* $= 1$ if the match was lit and $ML = 0$ otherwise).

The set $\mathcal{U}$ of exogenous variables includes conditions that suffice to make all relationships deterministic (such as whether the wood is dry, there is enough oxygen in the air, etc.). Suppose that $\vec{u}$ is a setting of the exogenous variables that makes a forest fire possible (i.e., the wood is sufficiently dry, there is oxygen in the air, and so on). Then, for example, $F_F(\vec{u}, L, ML)$ is such that $F = 1$ if either $L = 1$ or $ML = 1$. ∎

*Supported in part by NSF under grants IRI-96-25901 and IIS-0090145.

†Supported in part by grants from NSF, ONR, AFOSR, and MICRO.

Given a causal model $M = (\mathcal{S}, \mathcal{F})$, a (possibly empty) vector $\vec{X}$ of variables in $\mathcal{V}$, and vectors $\vec{x}$ and $\vec{u}$ of values for the variables in $\vec{X}$ and $\mathcal{U}$, respectively, we can define a new causal model denoted $M_{\vec{X} \leftarrow \vec{x}}$ over the signature $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{R}|_{\mathcal{V}-\vec{X}})$. Intuitively, this is the causal model that results when the variables in $\vec{X}$ are set to $\vec{x}$ by external action, the cause of which is not modeled explicitly. Formally, $M_{\vec{X} \leftarrow \vec{x}} = (\mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X} \leftarrow \vec{x}})$, where $F_Y^{\vec{X} \leftarrow \vec{x}}$ is obtained from F_Y by setting the values of the variables in $\vec{X}$ to $\vec{x}$.

We can describe (some salient features of) a causal model M using a *causal network*. This is a graph with nodes corresponding to the random variables in $\mathcal{V}$ and an edge from a node labeled X to one labeled Y if F_Y depends on the value of X. Intuitively, variables can have a causal effect only on their descendants in the causal network; if Y is not a descendant of X, then a change in the value of X has no affect on the value of Y.

We restrict attention to what are called *recursive* (or *acyclic*) equations; these are ones that can be described with a causal network that is a dag. It should be clear that if M is a recursive causal model, then there is always a unique solution to the equations in $M_{\vec{X} \leftarrow \vec{x}}$, given a setting $\vec{u}$ for the variables in $\mathcal{U}$. Such a setting is called a *context*. Contexts will play the role of possible worlds when we model uncertainty.

Syntax and Semantics: Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a formula of the form $X = x$, for $X \in V$ and $x \in \mathcal{R}(X)$, is called a *primitive event*. A *basic causal formula* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events; $Y_1, \ldots, Y_k, X$ are variables in $\mathcal{V}$; $Y_1, \ldots, Y_k$ are distinct; $x \in \mathcal{R}(X)$; and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ holds in the counterfactual world that would arise if Y_i is set to y_i, $i = 1, \ldots, k$. A *causal formula* is a Boolean combination of basic causal formulas.

A causal formula φ is true or false in a causal model, given a context. We write $(M, \vec{u}) \models \varphi$ if φ is true in causal model M given context $\vec{u}$. $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$ (that is, the unique vector of values for the exogenous variables that simultaneously satisfies all equations $F_Z^{\vec{Y} \leftarrow \vec{y}}$, $Z \in \mathcal{V} - \vec{Y}$, with the variables in $\mathcal{U}$ set to $\vec{u}$). We extend the definition to arbitrary causal formulas in the obvious way.

Note that the structural equations are deterministic. We later add probability to the picture by putting a probability on the set of contexts (i.e., on the possible worlds).

3 The Definition of Explanation

As we said in the introduction, many definitions of causal explanation have been given in the literature. The "classical" approaches in the philosophy literature, such as Hempel's 1965 *deductive-nomological* model and Salmon's 1989 *statistical relevance* model (as well as many other approaches) have a serious problem: they fail to exhibit the directionality inherent in common explanations. While it seems reasonable to say "the height of the flag pole explains the length of the shadow", it would sound awkward if one were to explain the former with the latter. Despite all the examples in the philosophy literature on the need for taking causality and counterfactuals into account, and the extensive work on causality defined in terms of counterfactuals in the philosophy literature, as Woodward 2001 observes, philosophers have been reluctant to build a theory of explanation on top of a theory of causality. The concern seems to be one of circularity.

In [Halpern and Pearl, 2001], we give a definition of causality that assumes that the causal model and all the relevant facts are given; the problem is to determine which of the given facts are causes. (We discuss this definition in more detail below.) We give a definition of explanation based on this defintion of causality. The role of explanation is to provide the information needed to establish causation. As discussed in the introduction, we view an explanation as a fact that is not known for certain but, if found to be true, would constitute a genuine cause of the explanandum, regardless of the agent's initial uncertainty. Thus, what counts as an explanation depends on what you already know and, naturally, the definition of explanation is relative to the agent's epistemic state (as in Gärdenfors 1988). It is also natural, from this viewpoint, that an explanation includes fragments of the causal model M, or reference to the physical laws which underly the connection between the cause and the effect. To borrow an example from [Gärdenfors, 1988], if we want an explanation of why Mr. Johansson has been taken ill with lung cancer, the information that he worked in asbestos manufacturing for many years is not going to be a satisfactory explanation to someone who does not know anything about the effects of asbestos on people's health. In this case, the causal model (or relevant parts of it) must be part of the explanation. On the other hand, for someone who knows the causal model but does not know that Mr. Johansson worked in asbestos manufacturing, the explanation would involve Mr. Johansson's employment but would not mention the causal model.

Our definition of explanation is motivated by the following intuitions. An individual in a given epistemic state K asks why φ holds. What constitutes a good answer to his question? A good answer must provide information that goes beyond K and be such that the individual can see that it would, if true, be (or be very likely to be) a cause of φ. We may also want to require that φ be true (or at least probable). Although our basic definition does not require this, but it is easy to do so.

To make this precise, we must explain (1) what it means for ψ to be a cause of φ and (2) how to capture the agent's epistemic state. In [Halpern and Pearl, 2001], we dealt with the first question. In the next subsection we review the definitions. The following subsections discuss the second question.

3.1 The Definition of Causality

We want to make sense of statements of the form "event A is an actual cause of event B in context $\vec{u}$ of model M". Note that we assume the context and model are given. Intuitively, they encode the background knowledge. All the relevant facts are known. The only question is picking out which of them are the causes of φ.

The types of events that we allow as actual causes are ones of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$—that is, conjunctions of primitive events; we typically abbreviate this as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary Boolean combinations of primitive events. We argue in [Halpern and Pearl, 2001] that it is reasonable to restrict causes to conjunctions (and, in particular, to disallow disjunctions). This restriction seems less reasonable in the case of explanation; we return to this point below. In any case, the definition of causality we give is restricted to conjunctive causes.

Definition 3.1: (Actual cause) $\vec{X} = \vec{x}$ is an *actual cause of* φ *in* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$. (That is, both $\vec{X} = \vec{x}$ and φ are true in the actual world.)

AC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models (Z = z^*)$ for $Z \in \vec{Z}$, then

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi$. In words, changing $(\vec{X}, \vec{W})$ from $(\vec{x}, \vec{w})$ to $(\vec{x}', \vec{w}')$ changes φ from true to false;

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$. In words, setting $\vec{W}$ to $\vec{w}'$ should have no effect on φ as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2. Minimality ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing φ in AC2(a) are considered part of a cause; inessential elements are pruned. ∎

For future reference, we say that $\vec{X} = \vec{x}$ is a *weak cause* of φ in $(M, \vec{u})$ if AC1 and AC2 hold, but not necessarily AC3.

The core of this definition lies in AC2. Informally, the variables in $\vec{Z}$ should be thought of as describing the "active causal process" from $\vec{X}$ to φ. These are the variables that mediate between $\vec{X}$ and φ. AC2(a) says that there exists a setting $\vec{x}'$ of $\vec{X}$ that changes φ to $\neg\varphi$, as long as the variables not involved in the causal process ($\vec{W}$) take on value $\vec{w}'$. AC2(a) is reminiscent of the traditional counterfactual criterion of Lewis 1986b, according to which φ should be false if it were not for $\vec{X}$ being $\vec{x}$. However, AC2(a) is more permissive than the traditional criterion; it allows the dependence of φ on $\vec{X}$ to be tested under special circumstances.

AC2(b) is an attempt to counteract the "permissiveness" of AC2(a) with regard to structural contingencies. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from φ to $\neg\varphi$; setting $\vec{W}$ to $\vec{w'}$ merely eliminates spurious side effects that tend to mask the action of $\vec{X}$. It captures the fact that setting $\vec{W}$ to $\vec{w}'$ should not affect the causal process, by requiring that changing $\vec{W}$ from $\vec{w}$ to $\vec{w'}$ has no effect on the value of φ.

This definition is discussed and defended in much more detail in [Halpern and Pearl, 2001], where it is compared to other definitions of causality. In particular, it is shown to avoid a number of problems that have been identified with Lewis's account (e.g., see [Pearl, 2000, Chapter 10]), such as commitment to transitivity of causes. For the purposes of this paper, we ask that the reader accept the definition. We note that, to some extent, our definition of explanation is modular in its use of causality, in that another definition of causality could be substituted for the one we use in the definition of explanation (provided it was given in the same framework).

The following example will help to clarify the definition of both causality and explanation.

Example 3.2: Suppose that two arsonists drop lit matches in different parts of a dry forest; both cause trees to start burning. Consider two scenarios. In the first, called "disjunctive," either match by itself suffices to burn down the whole forest. That is, even if only one match were lit, the forest would burn down. In the second scenario, called "conjunctive," both matches are necessary to burn down the forest; if only one match were lit, the fire would die down before the forest was consumed. We can describe the essential structure of these two scenarios using a causal model with four variables:

- an exogenous variable U which determines, among other things, the motivation and state of mind of the arsonists. For simplicity, assume that $\mathcal{R}(U) = \{u_{00}, u_{10}, u_{01}, u_{11}\}$; if $U = u_{ij}$, then the first arsonist intends to start a fire iff $i = 1$ and the second arsonist intends to start a fire iff $j = 1$. In both scenarios $U = u_{11}$.
- endogenous variables ML_1 and ML_2, each either 0 or 1, where $ML_i = 0$ if arsonist i doesn't drop the match and $ML_i = 1$ if he does, for $i = 1, 2$.
- an endogenous variable *FB* for forest burns down, with values 0 (it doesn't) and 1 (it does).

Both scenarios have the same causal network (see Figure 1); they differ only in the equation for *FB*. Given $u \in \mathcal{R}(U)$, for the disjunctive scenario we have $F_{FB}(u, 1, 1) = F_{FB}(u, 0, 1) = F_{FB}(u, 1, 0) = 1$ and $F_{FB}(u, 0, 0) = 0$; for the conjunctive scenario we have $F_{FB}(u, 1, 1) = 1$ and $F_{FB}(u, 0, 0) = F_{FB}(u, 1, 0) = F_{FB}(u, 0, 1) = 0$.

In general, the causal model for reasoning about forest fires would involve many other variables; in particular, variables for other potential causes of forest fires such as lightning and unattended campfires. Here we focus on that part of the causal model that involves forest fires started by arsonists. Since for causality we assume that all the relevant facts are given, we can assume here that it is known that there were no unattended campfires and there was no lightning, which makes it safe to ignore that portion of the causal model.

Denote by M_1 and M_2 the (portion of the) causal models associated with the disjunctive and conjunctive scenarios, respectively. The causal network for the relevant portion of M_1 and M_2 is described in Figure 1.

Despite the differences in the underlying models, it is not hard to show that each of $ML_1 = 1$ and $ML_2 = 1$ is a cause of $FB = 1$ in both scenarios. We present the argument for $ML_1 = 1$ here. To show that $ML_1 = 1$ is a cause in M_1 let

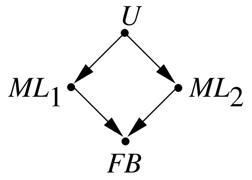

Figure 1: The causal network for M_1 and M_2.

$\vec{Z} = \{ML_1, FB\}$, so $\vec{W} = \{ML_2\}$. It is easy to see that the contingency $ML_2 = 0$ satisfies the two conditions in AC2. AC2(a) is satisfied because, in the absence of the second arsonist ($ML_2 = 0$), the first arsonist is necessary and sufficient for the fire to occur ($FB = 1$). AC2(b) is satisfied because, if the first match is lit ($ML_1 = 1$) the contingency $ML_2 = 0$ does not prevent the fire from burning the forest. Thus, $ML_1 = 1$ is a cause of $FB = 1$ in M_1. (Note that we needed to set ML_2 to 0, contrary to facts, in order to reveal the latent dependence of FB on ML_1. Such a setting constitutes a structural change in the original model, since it involves the removal of some structural equations.) The argument that $ML_1 = 1$ is also a cause of $FB = 1$ in M_2 is similar. (Again, taking $\vec{Z} = \{ML_1, FB\}$ and $\vec{W} = \{ML_2\}$ works.)

This example also illustrates the need for the minimality condition $AC3$. For example, if lighting a match qualifies as the cause of fire then lighting a match and sneezing would also pass the tests of AC1 and AC2, and awkwardly qualify as the cause of fire. Minimality serves here to strip "sneezing" and other irrelevant, over-specific details from the cause.

It might be argued that allowing disjunctive causes would be useful in this case to distinguish M_1 from M_2 as far as causality goes. A purely counterfactual definition of causality would make $ML_1 = 1 \vee ML_2 = 1$ a cause of $FB = 1$ in M_1 (since, if $ML_1 = 1 \vee ML_2 = 1$ were not true, then $FB = 1$ would not be true), but would make neither $ML_1 = 1$ nor $ML_2 = 1$ individually a cause (for example, if $ML_1 = 1$ were not true in M_1, $FB = 1$ would still be true). Clearly, our definition does not enforce this intuition. Purely counterfactual definitions of causality have other well-known problems. We do not have a strong intuition as to the best way to deal with disjunction in the context of causality, and believe that disallowing it is reasonably consistent with intuitions. Interestingly, as we shall see in Section 3.2, our definition of explanation *does* distinguish M_1 from M_2; each of $ML_1 = 1$ and $ML_2 = 1$ is an explanation of $FB = 1$ in M_1 under our definition of explanation, but neither is an explanation of $FB = 1$ in M_2. In M_2, the explanation of $FB = 1$ is $ML_1 = 1 \wedge ML_2 = 1$: both matches being lit are necessary to explain the forest burning down. ∎

3.2 The Basic Definition of Explanation

All that remains to do before giving the definition of explanation is to discuss how to capture the agent's epistemic state in our framework. For ease of exposition, we first consider the case where the causal model is known and the context is uncertain. (The minor modifications required to deal with the general case are described in Section 3.4.) In that case, one way of describing an agent's epistemic state is by simply describing the set of contexts the agent considers possible. This choice is very much in the spirit of the standard "possible worlds" definitions of knowledge and belief.

Definition 3.3: (Explanation) Given a structural model M, $\vec{X} = \vec{x}$ *is an explanation of* φ *relative to a set* $\mathcal{K}$ *of contexts* if the following conditions hold:

EX1. $(M, \vec{u}) \models \varphi$ for each context $\vec{u} \in \mathcal{K}$. (That is, φ must hold in all contexts the agent considers possible—the agent considers what she is trying to explain as an established fact)

EX2. $\vec{X} = \vec{x}$ is a *weak cause* of φ in $(M, \vec{u})$ (that is, AC1 and AC2 hold, but not necessarily AC3) for each $\vec{u} \in \mathcal{K}$ 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 $\vec{u}' \in \mathcal{K}$. (This just says that the agent considers a context possible where the explanation is false, so the explanation is not known to start with, and considers a context possible where the explanation is true, so that it is not vacuous.) ∎

Our requirement EX4 that the explanation is not known may seem incompatible with linguistic usage. Someone discovers some fact A and says "Aha! That explains why B happened." Clearly, A is not an explanation of why B happened relative to the epistemic state *after* A has been discovered, since at that point A is known. However, A can legitimately be considered an explanation of B relative to the epistemic state before A was discovered.

Consider the arsonists in Example 3.2. If the causal model has only arsonists as the cause of the fire, there are two possible explanations in the disjunctive scenario: arsonist 1 did it or arsonist 2 did it (assuming $\mathcal{K}$ consists of three contexts, where either 1, 2, or both set the fire). In the conjunctive scenario, no explanation is necessary, since the agent knows that both arsonists must have lit a match if arson is the only possible cause of the fire (assuming that the agent considers these to be the only possible arsonists).

Perhaps more interesting is to consider a causal model with other possible causes, such as lightning and unattended campfires. Since the agent knows that there was a fire, in each of the contexts in $\mathcal{K}$, at least one of the potential causes must have actually occurred. If we assume that there is a context where only arsonist 1 lit the fire (and, say, there was lightning) and another where only arsonist 2 lit the fire then, in the conjunctive scenario, $ML_1 = 1 \wedge ML_2 = 1$ is an explanation of $FB = 1$, but neither $ML_1 = 1$ nor $ML_2 = 1$ by itself is an explanation (since neither by itself is a cause in all contexts in $\mathcal{K}$ that satisfy the formula). On the other hand, in the disjunctive scenario, both $ML_1 = 1$ and $ML_2 = 1$ are explanations.

It is worth noting here that the minimality clause EX3 applies to all contexts. This means that our rough gloss of $\vec{X} = \vec{x}$ being an explanation of φ relative to a set $\mathcal{K}$ of contexts if $\vec{X} = \vec{x}$ is a cause of φ in each context in $\mathcal{K}$ where $\vec{X} = \vec{x}$ holds is not quite correct. For example, although

$ML_1 = 1 \wedge ML_2 = 1$ is an explanation of fire in the conjunctive scenario (if $\mathcal{K}$ includes contexts where there are other possible causes of fire), it is a cause of fire in none of the contexts in which it holds. The minimality condition AC3 would say that each of $ML_1 = 1$ and $ML_2 = 1$ is a cause, but their conjunction is not.

Note that, as for causes, we have disallowed disjunctive explanations. Here the motivation is less clear cut. It does make perfect sense to say that the reason that φ happened is either A or B (but I don't know which). There are some technical difficulties with disjunctive explanations, which suggest philosophical problems. For example, consider the conjunctive scenario of the arsonist example again. Suppose that the structural model is such that the only causes of fire are the arsonists, lightning, and unattended campfires and that $\mathcal{K}$ consists of contexts where each of these possibilities is the actual cause of the fire. Once we allow disjunctive explanations, what is the explanation of fire? One candidate is "either there were two arsonists or there was lightning or there was an unattended campfire (which got out of hand)". But this does not satisfy EX4, since the disjunction is true in every context in $\mathcal{K}$. On the other hand, if we do not allow the disjunction of all possible causes, which disjunction should be allowed as an explanation? As a technical matter, how should the minimality condition EX3 be rewritten? We could not see any reasonable way to allow some disjunctions in this case without allowing the disjunction of all causes (which will not in general satisfy EX4).

We believe that, in cases where disjunctive explanations seem appropriate, it is best to capture this directly in the causal model by having a variable that represents the disjunction. (Essentially the same point is made in [Chajewska and Halpern, 1997].) For example, consider the disjunctive scenario of the arsonist example, where there are other potential causes of the fire. If we want to allow "there was an arsonist" to be an explanation without specifically mentioning who the arsonist is, then it can be easily accomplished by replacing the variables ML_1 and ML_2 in the model by a variable ML which is 1 iff at least one arsonist drops a match. Then $ML = 1$ becomes an explanation, without requiring disjunctive explanations.

Why not just add ML to the model rather than using it to replace ML_1 and ML_2? We have implicitly assumed in our framework that all possible combinations of assignments to the variables are possible (i.e., there is a structural contingency for any setting of the variables). If we add ML and view it as being logically equivalent to $ML_1 \vee ML_2$ (that is, $ML = 1$ *by definition* iff at least one of ML_1 and ML_2 is 1) then, for example, it is logically impossible for there to be a structural contingency where $ML_1 = 0$, $ML_2 = 0$, and $ML = 1$. Thus, in the presence of logical dependences, it seems that we need to restrict the set of contingencies that can be considered to those that respect the dependencies. We have not yet considered the implications of such a change for our framework, so we do not pursue the matter here.

3.3 Partial Explanations and Explanatory Power

Not all explanations are considered equally good. Some explanations are more likely than others. An obvious way to define the "goodness" of an explanation is by bringing probability into the picture. Suppose that the agent has a probability on the set $\mathcal{K}$ of possible contexts. In this case, we can consider the probability of the set of contexts where the explanation $\vec{X} = \vec{x}$ is true. For example, if the agent has reason to believe that the first arsonist is extremely unlikely to have caused the fire (perhaps he had defective matches), then the set of contexts where $ML_2 = 1$ holds would have higher probability than those where $ML_1 = 1$ holds. Thus, $ML_2 = 1$ would be considered a better explanation of the fire in the disjunctive model than $ML_1 = 1$.

But the probability of an explanation is only part of the story; the other part concerns the degree to which an explanation fulfills its role (relative to φ) in the various contexts considered. This becomes clearer when we consider *partial* explanations. The following example, taken from [Gärdenfors, 1988], is one where partial explanations play a role.

Example 3.4: Suppose I see that Victoria is tanned and I seek an explanation. Suppose that the causal model includes variables for "Victoria took a vacation in the Canary Islands", "sunny in the Canary Islands", and "went to a tanning salon". The set $\mathcal{K}$ includes contexts for all settings of these variables compatible with Victoria being tanned. Note that, in particular, there is a context where Victoria both went to the Canaries (and didn't get tanned there, since it wasn't sunny) and to a tanning salon. Gärdenfors points out that we normally accept "Victoria took a vacation in the Canary Islands" as a satisfactory explanation of Victoria being tanned and, indeed, according to his definition, it is an explanation. Victoria taking a vacation is not an explanation (relative to the context $\mathcal{K}$) in our framework, since there is a context $\vec{u}^* \in \mathcal{K}$ where Victoria went to the Canary Islands but it was not sunny, and in $\vec{u}^*$ the actual cause of her tan is the tanning salon, not the vacation.

For us, the explanation would have to be "Victoria went to the Canary Islands *and* it was sunny." In this case we can view "Victoria went to the Canary Islands" as a partial explanation (in a formal sense to be defined below). ■

In Example 3.4 the partial explanation can be extended to a full explanation by adding a conjunct. But not every partial explanation can be extended to a full explanation. Roughly speaking, the full explanation may involve exogenous factors, which are not permitted in explanations. Assume, for example, that going to a tanning salon was not considered an endogenous variable in our model but, rather, the model simply had an exogenous variable U_s that could make Victoria suntanned even in the absence of sun in Canary islands. Likewise, assume that the weather in Canary island was also part of the background context. In this case, we would still consider Victoria's vacation to provide a partial explanation of her sun tan, since the context where it fails to be a cause (no sun in the Canary island) is fairly unlikely, but we cannot add conjuncts to this event to totally exclude that context from the agent's realm of possibilities.

The situation actually is quite common, as the following example shows.

Example 3.5: Suppose that the sound on a television works but there is no picture. Furthermore, the only cause of there being no picture that the agent is aware of is the picture tube being faulty. However, the agent is also aware that there are times when there is no picture even though the picture tube works perfectly well—intuitively, "for inexplicable reasons". This is captured by the causal network described in Figure 2, where T describes whether or not the picture tube is working (1 if it is and 0 if it is not) and P describes whether or not there is a picture (1 if there is and 0 if there is not). The ex-

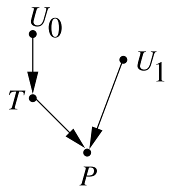

Figure 2: The television with no picture.

ogenous variable U_0 determines the status of the picture tube: $T = U_0$. The exogenous variable U_1 is meant to represent the mysterious "other possible causes". If $U_1 = 0$, then whether or not there is a picture depends solely on the status of the picture tube—that is, $P = T$. On the other hand, if $U_1 = 1$, then there is no picture ($P = 0$) no matter what the status of the picture tube. Thus, in contexts where $U_1 = 1$, $T = 0$ is *not* a cause of $P = 0$. Now suppose that $\mathcal{K}$ includes a context $\vec{u}_{00}$ where $U_0 = U_1 = 0$. Then it is easy to see that there is no explanation of $P = 0$. The only plausible explanation, that the picture tube is not working, is not a cause of $P = 0$ in the context $\vec{u}_{00}$. On the other hand, $T = 0$ is a cause of $P = 0$ in all other contexts in $\mathcal{K}$ satisfying $T = 0$. If the probability of $\vec{u}_{00}$ is small (capturing the intuition that it is unlikely that more than one thing goes wrong with a television at once), then we are entitled to view $T = 0$ as a quite good partial explanation of $P = 0$. ∎

These examples motivate the following definition.

Definition 3.6: Let $\mathcal{K}_{\vec{X}=\vec{x},\varphi}$ be the largest subset $\mathcal{K}'$ of $\mathcal{K}$ such that $\vec{X} = \vec{x}$ is an explanation of φ relative to $\mathcal{K}_{\vec{X}=\vec{x},\varphi}$. (It is easy to see that there is a largest such set.) Then $\vec{X} = \vec{x}$ is a *partial explanation of* φ *with goodness* $\Pr(\mathcal{K}_{\vec{X}=\vec{x},\varphi}|\vec{X} = \vec{x})$. Thus, the goodness of a partial explanation measures the extent to which it provides an explanation of φ.[1] ∎

In Example 3.4, if the agent believes that it is sunny in the Canary Islands with probability .9 (that is, the probability that it was sunny given that Victoria is suntanned and that she went to the Canaries is .9), then Victoria going to the Canaries is a partial explanation of her being tanned with goodness .9. The relevant set $\mathcal{K}'$ consists of those contexts where it is sunny in the Canaries. Similarly, in Example 3.5, if the agent believes that the probability of both the picture tube being faulty and the other mysterious causes being operative is .1, then $T = 0$ is a partial explanation of $P = 0$ with goodness .9 (with $\mathcal{K}'$ consisting of all the contexts where $U_1 = 1$).

A full explanation is clearly a partial explanation with goodness 1, but we are often satisfied with partial explanations $\vec{X} = \vec{x}$ that are not as good, especially if they have high probability (i.e., if $\Pr(\vec{X} = \vec{x})$ is high). In general, there is a tension between the goodness of an explanation and its probability.

These ideas also lead to a definition of explanatory power. Consider Example 3.2 yet again, and suppose that there is an endogenous random variable O corresponding to the presence of oxygen. Now if $O = 1$ holds in all the contexts that the agent considers possible, then $O = 1$ is excluded as an explanation by EX4. But suppose that $O = 0$ holds in one context that the agent considers possible (for example, there may be another combustible gas), albeit a very unlikely one. In that case, $O = 1$ becomes a very good partial explanation of the fire. Nevertheless, it is an explanation with, intuitively, very little explanatory power. How can we make this precise?

Suppose that there is a probability distribution $\Pr^-$ on a set $\mathcal{K}^-$ of contexts larger than $\mathcal{K}$ that intuitively represents the agent's prior probability before the explanandum φ is discovered. That is, $\Pr$ is the result of conditioning $\Pr^-$ on φ and $\mathcal{K}$ consists of the subset of $\mathcal{K}^-$ that satisfies φ. Gärdenfors identifies the explanatory power of the (partial) explanation $\vec{X} = \vec{x}$ of φ with $\Pr^-(\varphi|\vec{X} = \vec{x})$ (see [Chajewska and Halpern, 1997; Gärdenfors, 1988]). If this probability is higher than $\Pr^-(\varphi)$, then the explanation makes φ more likely. While this explanatory power, we would argue that a better measure of the explanatory power of $\vec{X} = \vec{x}$ is $\Pr^-(\mathcal{K}_{\vec{X}=\vec{x},\varphi}|\vec{X} = \vec{x})$. According to either definition, under reasonable assumptions about $\Pr^-$, $O = 1$ has much lower explanatory power than, say $ML = 1$. Moreover, the two definitions agree in the case that $\vec{X} = \vec{x}$ is a full explanation (since then $\mathcal{K}_{\vec{X}=\vec{x},\varphi}$ is just $\mathcal{K}$, the set of contexts in $\mathcal{K}^-$ where φ is true). The difference between the two definitions arises if there are contexts where φ and $\vec{X} = \vec{x}$ both happen to be true, but $\vec{X} = \vec{x}$ is not a cause of φ. Such spurious correlations are exculded by our suggested definition. (See [Pearl, 2000] for some examples showing that considering spurious correlations leads to bad outcomes.)

Again, (partial) explanations with higher explanatory power typically are more refined and, hence, less likely. than explanations with less explanatory power. There is no obvious way to resolve this tension. (See [Chajewska and Halpern, 1997] for more discussion of this issue.)

As this discussion suggests, our definition shares some features with that of Gärdenfors' 1988. Like him, we consider explanation relative to an agent's epistemic state. Gärdenfors also considers a "contracted" epistemic state characterized by the distribution $\Pr^-$. Intuitively, $\Pr^-$ describes the agent's beliefs before discovering φ. (More accurately, it describes an epistemic state as close as possible to $\Pr$ where the agent does not ascribe probability 1 to φ.) If the agent's current belief in φ came about as the result of an observation ψ, then

[1] Here and elsewhere, a formula such as $\vec{X} = \vec{x}$ is being identified with the set of contexts where the formula is true. Recall that, since all contexts in $\mathcal{K}$ are presumed to satisfy φ, there is no need to condition on φ; this probability is already updated with the truth of the explanandum φ. Finally, note that our usage of partial explanation is related to, but different from, that in [Chajewska and Halpern, 1997].

we can take Pr to be the result of conditioning $\Pr^-$ on ψ, as we have done above. However, Gärdenfors does necessarily assume such a connection between Pr and $\Pr^-$. In any case, for Gärdenfors, $\vec{X} = \vec{x}$ is an explanation of φ relative to Pr if (1) $\Pr(\varphi) = 1$, (2) $0 < \Pr(\vec{X} = \vec{x}) < 1$, and (3) $\Pr^-(\varphi|\vec{X} = \vec{x}) > \Pr^-(\varphi)$. (1) is the probabilistic analogue of EX1. Clearly, (2) is the probabilistic analogue of EX4. Finally, (3) says that learning the explanation increases the likelihood of φ. Gärdenfors focuses on the explanatory power of an explanation, but does not take into account its prior probability. As pointed out in [Chajewska and Halpern, 1997], Gärdenfors' definition suffers from another defect: Since there is no minimality requirement like EX3, if $\vec{X} = \vec{x}$ is an explanation of φ, so too is $\vec{X} = \vec{x} \wedge \vec{Y} = \vec{y}$.

In contrast to Gärdenfors' definition, the dominant approach to explanation in the AI literature, the *maximum a posteriori (MAP)* approach (see, for example, [Henrion and Druzdzel, 1990; Pearl, 1988; Shimony, 1991]), focuses on the probabability of the explanation, given the explanandum (i.e, $\Pr^-(\vec{X} = \vec{x}|\varphi) = \Pr(\vec{X} = \vec{x})$), but does not take explanatory power into account. The MAP approach is based on the intuition that the best explanation for an observation is the state of the world (in our setting, the context) that is most probable given the evidence. The most probable explanation for φ is then the context $\vec{u}^*$ such that $\Pr(\vec{u}^*) = \max_{\vec{u}} \Pr(\vec{u})$. Thus, an explanation is a (complete) context. This means that part of the explanation will include totally irrelevant facts (the agent sneezed). Moreover, it is quite sensitive to the description of the context (see [Chajewska and Halpern, 1997] for details) and does not directly take causality into account.

To some extent, these problems can be dealt with by limiting the set of candidate explanations to ancestors (of the explanandum) in the causal network; this also avoids many of the problems associated with non-causal approaches (although it requires there to be a causal network in the background). However, the MAP approach does not go far enough. One problem is that propositions with extremely high prior probabilities (e.g., that oxygen is present in the room) will also receive high posterior probabilities, regardless of how relevant they are to the events explained. To remedy this problem, more intricate combinations of the quantities $\Pr(\vec{X} = \vec{x})$, $\Pr^-(\varphi|\vec{X} = \vec{x})$, and $\Pr^-(\varphi)$ have been suggested to quantify the causal relevance of $\vec{X} = \vec{x}$ on φ but, as argued in [Pearl, 2000, p. 221], without taking causality into account, no such combination of parameters can work.

3.4 The General Definition

In general, an agent may be uncertain about the causal model, so an explanation will have to include information about it. (Gärdenfors 1988 and Hempel 1965 make similar observations). It is relatively straightforward to extend our definition of explanation to accommodate this. Now an epistemic state $\mathcal{K}$ consists not only of contexts, but of pairs $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$. Call such a pair a *situation*. Intuitively, now an explanation should consist of some causal information (such as "prayers do not cause fires") and the facts that are true. Thus, a *(general) explanation* has the form $(\psi, \vec{X} = \vec{x})$, where ψ is an arbitrary formula in our causal language and, as before, $\vec{X} = \vec{x}$ is a conjunction of primitive events. We think of the ψ component as consisting of some causal information (such as "prayers do not cause fires", which corresponds to a conjunction of statements of the form $F = i \Rightarrow [P \leftarrow x](F = i)$, where P is a random variable describing whether or not prayer takes place). The first component in a general explanation is viewed as restricting the set of causal models. To make this precise, given a causal model M, we say ψ is *valid in* M, and write $M \models \psi$, if $(M, \vec{u}) \models \psi$ for all contexts $\vec{u}$ consistent with M. With this background, it is easy to state the general definition.

Definition 3.7: $(\psi, \vec{X} = \vec{x})$ *is an explanation of* φ *relative to a set* $\mathcal{K}$ *of situations* if the following conditions hold:

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

EX2. For all $(M, \vec{u}) \in \mathcal{K}$ such that $(M, \vec{u}) \models \vec{X} = \vec{x}$ and $M \models \psi$, $\vec{X} = \vec{x}$ is a weak cause of φ in $(M, \vec{u})$.

EX3. $(\psi, \vec{X} = \vec{x})$ is minimal; there is no pair $(\psi', \vec{X}' = \vec{x}') \neq (\psi, \vec{X} = \vec{x})$ satisfying EX2 such that $\{M'' \in M(\mathcal{K}) : M'' \models \psi'\} \supseteq \{M'' \in M(\mathcal{K}) : M'' \models \psi\}$, where $\mathcal{M}(\mathcal{K}) = \{M : (M, \vec{u}) \in \mathcal{K} \text{ for some } \vec{u}\}$, $\vec{X}' \subseteq \vec{X}$, and $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$. Roughly speaking, this says that no subset of X provides a weak cause of φ in more contexts than those where ψ is valid.

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 $(M', \vec{u}') \in \mathcal{K}$. ∎

Note that, in EX2, we now restrict to situations $(M, \vec{u}) \in \mathcal{K}$ that satisfy both parts of the explanation $(\psi, \vec{X} = \vec{x})$, in that $M \models \psi$ and $(M, \vec{u}) \models \vec{X} = \vec{x}$. Furthermore, although both components of an explanation are formulas in our causal language, they play very different roles. The first component serves to restrict the set of causal models considered (to those with the appropriate structure); the second describes a cause of φ in the resulting set of situations.

Clearly Definition 3.3 is the special case of Definition 3.7 where there is no uncertainty about the causal structure (i.e., there is some M such that if $(M', \vec{u}) \in \mathcal{K}$, then $M = M'$). In this case, it is clear that we can take ψ in the explanation to be *true*.

Definition 3.7 can also be extended to deal naturally with statistical information of the kind considered by Gärdenfors and Hempel. Let a *probabilistic causal model* be a tuple $M_{Pr} = (\mathcal{S}, \mathcal{F}, \Pr)$, where $M = (\mathcal{S}, \mathcal{F})$ is a causal model and Pr is a probability measure on the contexts defined by signature $\mathcal{S}$ of M. Information like "with probability .9, $X = 3$" is a restriction on probabilistic models, and thus can be captured using a formula in an appropriate extension of our language that allows such probabilistic reasoning. With this extended language, the definition of explanation using probabilistic causal models remains unchanged.

As an orthogonal issue, there is also no difficulty considering a probability on the set $\mathcal{K}$ of situations and defining partial explanation just as before.

Example 3.8: Using this general definition of explanation, consider Scriven's 1959 famous paresis example. Paresis develops only in patients who have been syphilitic for a long time, but only a small number of patients who are syphilitic in fact develop paresis. Furthermore, according to Scriven, no other factor is known to be relevant in the development of paresis.[2] This description is captured by a simple causal model M_P. There are two endogenous variables, S (for syphilis) and P (for paresis), and two exogenous variables, U_1, the background factors that determine S, and U_2, which intuitively represents "disposition to paresis", i.e., the factors that determine, in conjunction with syphilis, whether or not paresis actually develops. An agent who knows this causal model and that a patient has paresis does not need an explanation of why: the agent knows without being told that the patient must have syphilis and that $U_2 = 1$. On the other hand, for an agent who does not know the causal model (i.e., considers a number of causal models of paresis possible), $(\{M_P\}, S = 1)$ is an explanation of paresis. ∎

4 Discussion

We have given a formal definition of explanation in terms of causality. As we mentioned earlier, there are not too many formal definitions of explanation in terms of causality in the literature. One of the few exceptions is Lewis 1986a, who defends the thesis that "to explain an event is to provide some information about its causal history". While this view is compatible with our definition, there is no formal definition given to allow for a careful comparison between the approaches. In any case, if were to define causal history in terms of Lewis's 1986b definition of causality, we would inherit all the problems of that definition. As we said earlier, our definition avoids these problems.

So what are the problems with our definition? For one thing, it inherits whatever problems our definition of causality has. As observed in [Halpern and Pearl, 2001], our definition at times declares certain events to be causes (and hence candidate explanations) that, intuitively, should not be causes because they should fail AC2(a). The only reason that they do not fail AC2(a) is because of extremely unlikely structural contingencies. To some extent, we can avoid this problem by simply ignoring structural contingencies that are extremely unlikely (this is essentially the solution suggested in [Halpern and Pearl, 2001] in the context of causality). Of course, we can do this in the context of explanation too. Another possibility is to take the probability of the structural contingency into account more directly when computing the probability of the explanation. We are currently exploring this option.

We have mentioned the other significant problem of the definition already: dealing with disjunctive explanations. Disjunctions cause problems in the definition of causality, which is why we do not deal with them in the context of explanation. As we pointed out earlier, it may be possible to modify the definition of causality so as to be able to deal with conjunctions without changing the structure of our definition of explanation. We are currently exploring this.

[2] Apparently there are now other known factors, but this does not change the import of the example.

Finally, our definition gives no tools for dealing with the inherent tension between explanatory power, goodness of partial beliefs, and the probability of the explanation. Clearly this is an area that requires further work.

Acknowledgments:

Thanks to Riccardo Pucella and Vicky Weissman for useful comments.

References

[Chajewska and Halpern, 1997] U. Chajewska and J. Y. Halpern. Defining explanation in probabilistic systems. In *Proc. UAI '97*, pages 62–71, 1997.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1–2):9–43, 1997.

[Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.

[Halpern, 2000] J. Y. Halpern. Axiomatizing causal reasoning. *Journal of A.I. Research*, pages 317–337, 2000.

[Halpern and Pearl, 2001] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach– Part I: Causes. Available at http://www.cs.cornell.edu/home/halpern, 2001.

[Hempel, 1965] C. G. Hempel. *Aspects of Scientific Explanation*. Free Press, 1965.

[Henrion and Druzdzel, 1990] M. Henrion and M. J. Druzdzel. Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In *Uncertainty in Artificial Intelligence 6*, Elsevier Science, pages 17–32, 1990.

[Lewis, 1986a] D. Lewis. Causal explanation. In *Philosophical Papers*, volume II, pages 214–240. Oxford University Press, 1986.

[Lewis, 1986b] D. Lewis. Causation. In *Philosophical Papers*, volume II, pages 159–213. Oxford University Press, 1986. The original version of this paper, without numerous postscripts, appeared in the *Journal of Philosophy* **70**, 1973, pp. 113–126.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasining, and Inference*. Cambridge University Press, 2000.

[Salmon, 1989] W. C. Salmon. *Four Decades of Scientific Explanation*. University of Minnesota Press, 1989.

[Scriven, 1959] M. J. Scriven. Explanation and prediction in evolutionary theory. *Science*, 130:477–482, 1959.

[Shimony, 1991] S. E. Shimony. Explanation, irrelevance and statistical independence. In *Proc. AAAI '91*, pages 482–487, 1991.

[Woodward, 2001] J. Woodward. Explanation. In *The Blackwell Guide to the Philosophy of Science*. Basil Blackwell, 2001. To appear.