Causes and Explanations Revisited

James D. Park
Computer Science Deptartment
University of California, Los Angeles, CA 90095
jd@cs.ucla.edu

Abstract

This paper reconsiders the notions of actual cause and explanation in functional causal models. We demonstrate that isomorphic causal models can generate intuitively different causal pronouncements. This occurs because psychological factors not represented in the model determine what criteria we use to intuitively determine causation. This partially explains the difficulty encountered in previous attempts to define actual cause. Freed from trying fit all examples to match intuition directly (which is not possible using only the information in causal models), we provide definitions for causation matching the different types of causal criteria we intuitively apply. This formulation avoids the difficulties associated with previous definitions, and allows a more refined discussion of what constitutes a cause in a given situation. The definitions of actual causality also allow for a formulation of explanation which can provide more information than previous versions.

1 Introduction

Identifying an actual cause in a specific situation is a normal part of every day human reasoning. From the specifics of the situation, we sift through the events, identifying those that actually caused the event in question. This is different than general claims about causal tendencies. In fact, the causal nature of events can actually run counter to the general causal tendency of the variables. For example, buying a lottery ticket has the general tendency to decrease wealth. However, in certain circumstances (the numbers on the ticket correspond to the winning numbers), it can cause a significant increase wealth. People are able to make these situation dependent pronouncements fairly easily.

Automating this reasoning process would be very useful. Automating causal determinations has a number of applications, including natural language processing, diagnosis, and planning. Unfortunately, formulating a successful notion of actual cause has proved elusive. Previous attempts were often hindered by trying to embed them in formalisms that are not appropriate for modeling causality (See [Pearl, 2000, Chapter 10] for a discussion of these). Recently, functional causal models [Pearl, 2000] have proven to be an effective formalism for modeling causal relationships. There have been previous attempts to use functional causal models as a basis for defining actual cause, but they have been shown to suffer from serious defects of their own [Hopkins & Pearl, 2003]. In this paper, we identify (at least partially) why previous methods have failed, and provide new definitions that avoid many of their problems. These definitions allow for a more refined determination of actual causation. Additionally, they provide a useful basis for producing explanations.

2 Review of causal models

In this section we review the basic definitions from [Pearl, 2000] that we will use. First we need some notation. Upper case letters (e.g. X, Y, Z) represent variables. Lower case letters (e.g. x, y, z) represent a particular value of the corresponding variable. Bold face upper case letters (e.g. $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$) represent sets of variables, and bold face lower case letters (e.g. $\mathbf{x}, \mathbf{y}, \mathbf{z}$) an instantiation of the corresponding variables.

A *causal model* is a triple $(\mathbf{U}, \mathbf{V}, \mathcal{F})$ where $\mathbf{U}$ is the set of background (or exogenous) variables (i.e. their values are determined outside the model), and $\mathbf{V}$ is the set of endogenous variables (i.e. those whose value is determined by model variables). $\mathcal{F} = \{f_X \mid X \in \mathbf{V}\}$ is a set of functions where f_X maps from $\mathbf{U} \cup (\mathbf{V} \setminus X)$ to X. Together, for any setting of the background variables, $\mathcal{F}$ determines the values of all variables in $\mathbf{V}$. The *parents* of a variable X are the variables that directly influence the value of X (i.e. the set of variables that is non-trivial in f_X). The causal diagram of a model is a directed graph with vertices $\mathbf{U} \cup \mathbf{V}$ where there is an edge from vertex X to vertex Y iff X is a parent of Y in the model. We will restrict our attention to models with acyclic causal diagrams. A submodel of causal model M under intervention $\mathbf{X} = \mathbf{x}$ is a causal model $M_{\mathbf{x}} = (\mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}})$ where $\mathcal{F}_{\mathbf{x}} = \{f_Y \mid Y \in \mathbf{V} \setminus \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$. This allows us to talk about what would have happened if $\mathbf{X}$ had been $\mathbf{x}$. We will write $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ to mean the value that $\mathbf{Y}$ takes on in the submodel $M_{\mathbf{x}}$ in context $\mathbf{U} = \mathbf{u}$. For a propositional formula φ, defined over primitives of the form $X = x$, $\varphi_{\mathbf{x}}(\mathbf{u})$ represents the value of the formula when the primitives are evaluated in model M_x in context $\mathbf{u}$.

3 Previous work

Formulating the notion of cause has been a topic of philosophic debate for many years. Unfortunately, all of the philosophical works of which we are aware either suffer from serious problems, or are not precise enough to be falsifiable (See [Halpern & Pearl, 2001a] for a review of some of these problems).

There have been at least two previous attempts to formulate actual cause in functional causal models. The first was the causal beam approach [Pearl, 2000]. This notion of actual cause seemed to be too restrictive, classifying certain events as noncauses that intuitively clearly should have been causes (See for example the critique in [Halpern & Pearl, 2001a]). The other [Halpern & Pearl, 2001a] (which we will call ASC for Alternate situation Strong Cause) tried to remedy those problems, but introduced several others. [Hopkins & Pearl, 2003] provides a thorough analysis of its problems.

4 Why so hard?

The many formulations encountered, and the problems they exhibit begs the question: why is ascribing cause so difficult to formalize? Consider the following example from [Halpern & Pearl, 2001a]: Two arsonists light matches at opposite ends of the forrest. Both start fires, and the forrest burns down. Either arsonist alone was sufficient to burn down the forrest. Now, we ask whether arsonist 1 lighting his match caused the forrest to burn down. Clearly it did. The reason being that it was sufficient to burn it down, and participated in the actual causal process that brought it about.

Consider another example: A prisoner is scheduled for execution. The execution proceeds as follows: The executioner selects a gun from among several, which may or may not have a bullet in the chamber. The prisoner is given the option of death by lethal injection (which is always fatal), or face execution by gunshot. If the prisoner selects gunshot, the executioner fires the weapon. If the weapon he selected had a bullet, it kills the prisoner. If not, the execution fails and the prisoner goes free. In the situation we consider, the prisoner chooses execution by lethal injection. As it turns out, the selected gun was loaded so the prisoner would have died even if he had he chosen gunshot.

Now, we ask if the prisoner's choice of execution method caused his death. Obviously it did not. The reason being, that in this circumstance, regardless of his choice, he would still die. That is, his choice had no impact on the outcome.

Surprisingly, the obvious causal models in these two examples are in fact isomorphic as are the specific situations considered. The effect (forrest burns down $\leftrightarrow$ death) occurs in all but one possible instantiation of its direct causes (neither arsonist lights their match $\leftrightarrow$ execution by gunshot is selected and the gun is not loaded) and both sufficient conditions occurred in the situation under consideration. In other words, the information that rendered one a cause and the other a noncause is not contained in the causal model, or the specific situation. We achieved intuitively different results because we tested different notions of cause. In the forest fire example, we tested the ability to produce the fire, i.e. whether it was capable of bringing about the result. In the second example, we tested the consequentiality of the choice of execution methods, i.e. whether an alternate decision would produce a different result. The choice of what test to apply seems to be based on psychological factors outside of the causal model.

We believe that the difficulties associated with previous definitions are at least in part due to the fact that they failed to distinguish between different notions of cause.[1] In fact, we would argue that like many problems in philosophy, one reason cause eludes satisfactory definition is because the intuitive concept is fundamentally imprecise. What criteria to apply is often a subjective decision, witnessed by the ease with which one can often convince others that an alternate criteria is the appropriate one. For example, one may argue that it is inappropriate to say that arsonist 1 caused the fire since regardless of his action, the fire would still have occurred.

While the choice of criteria for determining cause is imprecise and subject to factors not representable in a functional causal model, the same is not true once a criteria has been chosen. Each criteria is precise and completely decidable based on the causal model and the state of its variables. Formulating and illustrating appropriate definitions of these criteria is the focus of the next section.

5 A new proposal

In this section, we introduce basic definitions which we then compose to form criteria for evaluating causal relationships. As there are several ways in which an event may be considered a cause, we provide definitions for the different aspects. We believe that all causal tests are based on two basic ideas, which are related to , though distinct from, necessity and sufficiency. By combining these two criteria we can produce the different criteria used for making causal judgments.

First, we need some background and auxiliary definitions. Foremost among them is what we allow to be causes and effects.

Definition 1 *An instantiation* $\mathbf{x}$*, of variables* $\mathbf{X}$ *is an* event *in context* $\mathbf{u}$ *if* $\mathbf{X}(\mathbf{u}) = \mathbf{x}$*. A propositional formula* φ *defined over the endogenous variables of the model is an* effect *in context* $\mathbf{u}$ *if* $\varphi(\mathbf{u})$.

Any potential cause of an effect must be an event. Note that we force events and effects to be true in the situation being considered. This is merely a convenience meant to simplify subsequent definitions.

Since we are concerned with determining cause in a particular context, the causal diagram of a model is not very informative. A variable X that in general is necessary to determine the value of another variable Y (and hence has an edge from X to Y in the diagram), may become irrelevant in a particular context. That is a particular value of the background variables $\mathbf{u}$ may make the value of Y unresponsive to the value of X. This observation motivates the following definition:

Definition 2 *The* context diagram *of causal world* $M(\mathbf{u})$ *is the directed graph whose vertices are the endogenous variables of* M*, and includes each edge* (X, Y) *where the func-*

[1] [Hall, 2002] makes a similar argument, although the details differ from ours.

tion f_Y is nontrivial in X when the background variables are fixed to $\mathbf{u}$.

The context diagram simplifies reasoning about causal relationships in a specific context.

Another preliminary notion we will need is sufficiency.[2]

Definition 3 *An event* $\mathbf{x}$ *is* sufficient *to cause effect* φ *in context* $\mathbf{u}$ *if for all instantiations* $\mathbf{y}$ *of the nondescendants* $\mathbf{Y}$ *of* $\mathbf{X}$ *in the context diagram of* $M(\mathbf{u})$, $\varphi_{\mathbf{xy}}(\mathbf{u})$.

Note that it must hold only when the nondescendants are manipulated. Descendants are excluded from manipulation since we believe that the values they take on, including interplay between them, are an integral component of determining causation. Allowing them to be manipulated would upset and skew that process.

Sufficiency is not really strong enough for our purposes. For example, consider this example from [Hall, 2002]:

> Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw

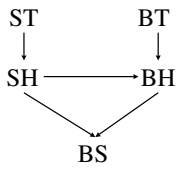

Figure 1: Rock throwing example

Figure 1 shows the causal diagram of a model of the story (the background variables, as well as the functional relationship between variables should be obvious in this story, as well as the others, so will not be explicated). BT (ST) represents Billy (Suzy) throwing BH (SH) represents Billy's (Suzy's) rock hitting the bottle, and BS stands for the bottle shattering. Billy throwing is sufficient to cause the bottle to shatter, but it doesn't actually participate in the causal process that made the bottle shatter. Suzy's action preempted Billy's. We expect not only sufficiency of the cause, but that in some sense, the event actually participates in the causal process. Defining this notion has been one of the primary difficulties in talking about actual cause. The following two definitions formalize these notions in a way that avoids the problems associated with previous approaches.

Definition 4 *Let* $\mathbf{Z}$ *be the children of* $\mathbf{X}$ *in the context diagram of* $\mathbf{u}$, *and let* $\mathbf{Y}$ *be some non-empty subset of* $\mathbf{Z}$. *Let* $\mathbf{W}$ *be the intersection of the descendants of* $\mathbf{X}$ *(including* $\mathbf{X}$*) and the ancestors of* $\mathbf{Y}$ *(excluding* $\mathbf{Y}$*). Then event* $\mathbf{xy}$ *is a* child *of event* $\mathbf{x}$ *in context* $\mathbf{u}$ *if* $\mathbf{W} \subseteq \mathbf{X}$.

[2]Note that this definition differs from the definition for sufficiency presented in [Pearl, 2000]. That definition was used for a different purpose, and is not appropriate in this context. It is more closely related to, although distinct from the notion of sustenance [Pearl, 2000].

A child event can be thought of as a next step in an unfolding causal process. The criteria that $\mathbf{W} \subseteq \mathbf{X}$ is merely to enforce the intuition that before a variable takes on a value, its direct causes must be fixed. That is, a new value can not be added, if the relevant parents have not already been included.

Definition 5 *An event* $\mathbf{x}$ *is a* strongly sufficient cause *of effect* φ *in context* $\mathbf{u}$ *if either of the following conditions hold:*

- *For all instantiations* $\mathbf{z}$ *of* $\mathbf{Z} = \mathbf{V} \setminus \mathbf{X}$, $\varphi_{\mathbf{xz}}(\mathbf{u})$
- $\mathbf{x}$ *is a sufficient cause of* φ *and there exists some child event* $\mathbf{y}$ *of* $\mathbf{x}$ *which is a strongly sufficient cause of* φ *in context* $\mathbf{u}$

Effectively, what we require is that the event be sufficient for the effect, and that as the causal process progresses from the event, the sufficiency is maintained. This prevents events that are preempted from being strongly sufficient. For example, while {BT=T} is sufficient for {BS=T}, the only child event, {BT=T,BH=F} is not, so the event is not strongly sufficient. Suzy throwing, on the other hand, is strongly sufficient, witnessed by the sequence {ST=T}, {ST=T,SH=T}, {ST=T,SH=T,BH=F}, {ST=T,SH=T,BH=F,BS=T} .

Strong sufficiency is the first of the two primary notions used to evaluate causality in specific situations. In fact, strong sufficiency by itself corresponds to one of the primary notions of causation. Strong sufficiency corresponds to our intuitions about physical causality. The cause sets in motion a chain of events capable of bringing about the effect. Intuitively, we expect causation to be transitive. If $\mathbf{x}$ causes $\mathbf{y}$ and $\mathbf{y}$ causes $\mathbf{z}$, we expect $\mathbf{x}$ to cause $\mathbf{z}$.[3] It is not hard to see that strong sufficiency meets that expectation. In fact, strong sufficiency turns out to be the only transitive relation among our definitions of cause.

The other primary consideration when testing actual causality is the notion of consequentiality. In addition to participating in the causal process, we often require that an event be in some sense consequential in order to be considered a cause. That consideration motivates the following definition.

Definition 6 *Let* $\mathbf{Z}$ *be the nondescendants of* $\mathbf{X}$ *in the context diagram of* $\mathbf{u}$. *Then, an event* $\mathbf{x}$ *is* potentially consequential *with respect to effect* φ *in context* $\mathbf{u}$ *if for some instantiations* $\mathbf{x}'$ *of* $\mathbf{X}$, *and* $\mathbf{z}$ *of* $\mathbf{Z}$, $\varphi_{\mathbf{xz}}(\mathbf{u})$ *and* $\neg\varphi_{\mathbf{x}'\mathbf{z}}(\mathbf{u})$. *Event* $\mathbf{x}$ *is* actually consequential *with respect to effect* φ *if for some instantiation* $\mathbf{x}'$ *of* $\mathbf{X}$, $\neg\varphi_{\mathbf{x}'}(\mathbf{u})$.

Like strong sufficiency, actual consequentiality alone is a common criteria for causation. For example, in the execution example from Section 4 the choice of lethal injection was not considered a cause of his death precisely because it was not consequential. If instead the gun had not been loaded, which would make the decision consequential, we would consider it a cause. Note that a consequential event necessarily participates in the actual causal process. That is, for every actually consequential event, there is a minimal strongly sufficient event that contains some member of the consequential event.

Potential consequentiality on the other hand is not of itself enough for an event to be considered a cause. However,

[3]In fact, this intuition is so strong that some theories of causation start with the transitivity of cause as a starting point [Lewis, 1986].

Test Strong Sufficiency	Test Consequentiality		
	Actual	Potential	No
Yes	Strong	Sustaining	Strongly Sufficient
No	Consequential	–	–

Table 1: Different types of actual cause can be tested by combining tests for strong sufficiency and consequentiality.

combining it with strong sufficiency produces a very common causal criteria.

Definition 7 *An event* $\mathbf{x}$ sustains *effect* φ *in context* $\mathbf{u}$ *if* $\mathbf{x}$ *is strongly sufficient and potentially consequential for* φ.

The notion of sustenance defined here is a modified version of the one that appeared in [Pearl, 2000]. It differs in two ways. First, it requires strong sufficiency as opposed to merely sufficiency. Second, it fixes the variables $\mathbf{W}$, which $\mathbf{x}$ must be sufficient relative to, while the previous definition has $\mathbf{W}$ as one of the parameters.

Sustenance is a very commonly applied criteria for causation. Most of the examples in [Halpern & Pearl, 2001a] which were deemed causes correspond to our notion of sustaining cause. In the forrest fire example in Section 4, arsonist 1 was considered a cause of the fire because it sustained the forrest fire.

An even more restrictive causal criteria is also sometimes useful.

Definition 8 *An event* $\mathbf{x}$ *is a* strong cause *of* φ *in context* $\mathbf{u}$ *if* $\mathbf{x}$ *is a strongly sufficient and actually consequential cause of* φ.

A strong cause can be thought of as a complete explication of why the effect occurred.

Often we talk about an event helping to cause an event even though it is neither consequential nor sufficient. It is instead part of some cause. That consideration motivates the following definition.

Definition 9 *An event* $\mathbf{xy}$ *(*$\mathbf{y}$ *may be empty) is a* contributory cause *of effect* φ *in context* $\mathbf{u}$ *if* $\mathbf{x}$ *is potentially consequential, and for some event* $\mathbf{z}$, $\mathbf{xz}$ *is a minimal sustaining cause of* φ.

Contributory causation is the closest in spirit to the ASC definition. If we insist on minimality, it reduces to singletons just as that definition does. Additionally, in all but one example included in [Halpern & Pearl, 2001a], contributory causation matches the pronouncements of the ASC definition. We consider that example among others in the next section.

6 Examples

The single example presented in [Halpern & Pearl, 2001a] where the ASC definition differs from our definition of contributory causation comes from the following story (along with some commentary), quoted from [Hall, 2002]:

> You are standing at a switch in the railroad tracks. Here comes the train: If you flip the switch, you'll send the train down the left-hand track; if you leave where it is, the train will follow the right-hand track. Either way, the train will arrive at the same point, since the tracks reconverge up ahead. Your action is not among the causes of this arrival; it merely helps to determine how the arrival is brought about (namely, via the left-hand track, or via the right-hand track).

Two models which produced different causal pronouncements were given, demonstrating the impact that model choice has on the determination of causation. The model that classified it as a cause is the one we are interested in. Figure 2 shows the model. F represents the state of the switch, LT (RT) whether the train goes down the left (right) track, and A represents whether the train arrives. The justification given was

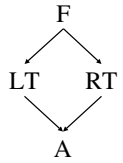

Figure 2: A possible model for the train story

that under normal circumstances, we may not call it a cause, but if something abnormal were to happen (for example, one of the tracks malfunctions, or is destroyed) then the action would be a cause. Our new definitions say that flipping the switch is a strongly sufficient cause (In the words of Hall it "helps to determine how the arrival is brought about"), but it is not even potentially consequential, so is not a contributory cause. This seems a more satisfactory result.

Unlike the ASC definition, these definitions do not always reduce to singleton causes even if minimality is enforced. Additionally, the different criteria produce causes that correspond to different ways that people think about cause.

For example, consider a situation where 5 people vote regarding a ballot measure. The votes are tallied, and the measure passes if 3 or more vote for it. The associated diagram appears in Figure 3. Suppose that 4 vote for the measure, and 1 votes against. According to the ASC definition, each individual who votes for the measure is an actual cause of the measure passing. In the causal beam approach, nothing qualifies as a cause of the measure passing. In this approach, it depends what type of cause we are considering.

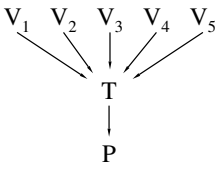

Figure 3: A voting example

Three votes for the measure constitute a strong cause of the measure passing, while two do not. This corresponds to our notion that a majority of the votes is required to cause it to pass.

Two votes for the measure constitute a consequential cause. Changing those votes to vote against the measure would make the measure fail. This corresponds to the way we often talk about causes in elections. Often election outcomes are attributed to the voting behavior of some minority of the electorate (in U.S. politics these include "Reagan Democrats", "soccer moms", "angry white men" , "African Americans", etc). While they are far from sufficient to win an election, we say that they caused the election result because if they had voted significantly differently the outcome of the election would have been different. On the other hand, most people understand that in the vast majority of elections their single vote did not cause the outcome. This is reflected in our definitions by the fact that none of the criteria yield a single voter as a cause in this situation. A single vote for the measure is a contributory cause however, as one would expect.

[Halpern & Pearl, 2001a] provides a number of examples that have proved problematic for other methods. As most of the examples correspond either to issues of sustenance or consequentiality[4] (and with the exception of the train example considered above, all of the positive examples are contributory causes, and the negative examples are not), we will not review them all here. We would like to revisit one example that was considered troublesome. The first version of the story is as follows: A machine severed Fred's finger while he was at work. He was rushed to the hospital and it was reattached. A month later his finger worked just as before. We would of course not consider getting his finger cut off a cause of it being fully function later (because it is not consequential). Now, suppose we add another component to the story: It is possible that Larry the loanshark may be waiting outside for Fred, so that he can cut off his finger to make an example of him for not paying his loan promptly. If Larry cuts off the finger, he will destroy it so that it can not be reattached. If Fred severs his finger at the factory, he will be rushed to the hospital, so Larry will not have an opportunity to do it himself. The causal diagram appears in Figure 4. Now, consider the

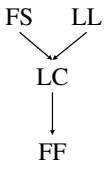

Figure 4: Finger severing example. FS represents severing his finger. LL represents whether Larry is lurking. LC represents whether Larry cuts off his finger and FF represents whether his finger is fully functional

situation in which Larry turns out not to be waiting outside, and Fred does sever his finger. Was Fred cutting his finger off a cause of his finger being fully functional the next month? According to the ASC definition it is. This was considered disturbing because adding a provision for a contingency that did not occur changed the the event from noncause to cause. In light of our new definitions, we do not see that particular consideration a problem. According to our definition, it again depends on what you are looking for. It is sustaining, but it is not consequential. Typically, when we are testing if an event caused something to remain unchanged, we are looking for a consequential cause. That is clearly the case in this example. However, it is still reasonable to say that cutting off his finger at the factory was a sustaining cause, because that event makes certain the effect, regardless of what Larry does or does not do.

Our definitions also handle correctly the examples in [Hopkins & Pearl, 2003] avoiding the problems of ASC, but because of space limitations we will not review them all here. We will consider a single example of the problem to demonstrate how these definitions avoid it. Consider the following story: Before an execution, the captain loads one of the guns of the two members of his firing squad, putting a blank in the other gun. Both gunmen then fire simultaneously and the prisoner dies. In the current situation, the captain put the bullet in the gun of gunman 1, and both men shoot. The model consists of the loader, and the two gunmen as direct causes of the the variable representing whether the prisoner is alive. Now, we ask if gunman 2 shooting caused the prisoner's death. According to the ASC definition, it did. Gunman 2 did not cause the prisoner's death according to any of our definitions of cause. It is clearly not consequential, nor sufficient, so it can not be a cause. It is not a contributory cause either because it is not a member of any minimal sufficient cause. According to these definitions, gunman 1 shooting is a consequential cause, gunman 1 shooting and the captain loading his gun is a strong cause, and gunman 1 shooting and the captain loading his gun are both contributory causes.

7 Explanations

If we have a probability distribution quantifying our uncertainty about what context will actually occur, we can use that distribution, combined with our model to produce explanations for an observed effect.

Causation forms an integral part of evaluating explanations (see [Chajewska & Halpern, 1997; Pearl, 2000; Halpern & Pearl, 2001b] for critiques of methods that ignore causal considerations when producing causes). In spirit, we agree with the intuitive notion of explanation put forth in [Halpern & Pearl, 2001b]. That is, an explanation is fundamentally just a hypothesized event which if true would be a possible cause of the observed phenomenon. Our new view of actual causation, however, makes the particulars somewhat different. Our definitions of actual cause form the basis for evaluating possible explanations. We allow any instantiation of endogenous variables to be an explanation and compare them based on how well they explain the effect. According to our view, there are different types of explanation, just as there are different types of causes. Which type of explanation to choose depends on its intended purpose. This is in contrast to the approach using ASC as the basis for causal explanation. We believe that this approach provides more flexibility and is more general than

[4] In fact, most of the difficult examples play off of the conflicting interpretation of cause between sustenance and consequentiality. That is, the event in question is typically either consequential or sustaining, but not both.

Explanation Variable	Posterior =0	Posterior =1	Explanatory Power =0	Explanatory Power =1
X_1	.52	.48	.35	.65
X_2	.67	.33	.40	.60
X_3	.74	.26	.42	.58

Table 2: Posteriors and explanatory power for possible explanations in the circuit example according to ASC definition.

ASC based explanations.

By way of example, consider the following situation: We have a circuit consisting of a single 3 input parity gate, with inputs X_1, X_2 and X_3, and output Y (i.e. $Y = 1$ when an odd number of the inputs are 1). Each input has an associated background variable, independent of the others, where $X_j = U_j$ and $\Pr(U_j = 1) = 1/(j+2)$. So, typically $Y = 0$ and $X_j = 0$. Now, suppose that we observe that $Y = 1$, and we want an explanation in terms of the X variables. The first question is what constitutes an explanation (or even a partial explanation in the ASC definition). According to the ASC based definition, the only possible explanations (or partial explanations) are singletons.[5] Now, how can we compare between them? The ASC approach provides three measures of the quality of an instantiation $\mathbf{x}$ as an explanation of effect φ in light of evidence $\mathbf{e}$ (which must entail φ). The first, named goodness is $\Pr(\mathbf{x}$ causes $\varphi|\mathbf{xe})$.[6] The second, the posterior probability is $\Pr(\mathbf{x}|\mathbf{e})$. The third, explanatory power is $\Pr(\mathbf{x}$ causes $\varphi|\mathbf{x})$. The goodness of all explanations in this example are 1. Also, the explanatory power, $\Pr(\mathbf{x}$ causes $y|\mathbf{x}) = \Pr(\mathbf{x}y|\mathbf{x})$ since every instantiation is a weak cause according to the ASC definition. Table 2 shows the explanatory power and posterior probability of the different possible explanations. It is not clear how to choose between them in this case. One thing to notice about this example is that every variable is consequential in every context, so notions of consequential causation (or by extension, contributory causation) become less interesting. We would argue that in this case, what we are really interested in is sustenance. Our definitions allow that to be considered easily. For a particular instantiation we can calculate the probabilities $\Pr(\mathbf{x}$ sustains $\varphi|\mathbf{e})$, $\Pr(\mathbf{x}$ strongly causes $\varphi|\mathbf{e})$, $\Pr(\mathbf{x}$ is a consequential cause of $\varphi|\mathbf{e})$ from the model and the probability distribution over the background variables. For a particular causal criteria $c(\mathbf{x}, \varphi, \mathbf{u})$, the probability that $\mathbf{x}$ is a cause of φ given $\mathbf{e}$ using criteria c is just

$$\Pr(\mathbf{x} \text{ "c caused" } \varphi|\mathbf{e}) = \frac{\sum_{\mathbf{u}\in\{\mathbf{u}|c(\mathbf{x},\varphi,\mathbf{u})\wedge\mathbf{E}(\mathbf{u})=\mathbf{e}\}} \Pr(\mathbf{u})}{\sum_{\mathbf{u}\in\{\mathbf{u}|\mathbf{E}(\mathbf{e})=\mathbf{e}\}} \Pr(\mathbf{u})}.$$

The probability of causation can then be used as the measure of explanation quality. In our circuit example, for example, $\Pr(X_1 = 1, X_2 = 0, X_3 = 0$ sustains $Y = 1|Y = 1) = .444$, which is the maximal probability for sustenance. This explanation (which is not allowed in the ASC version) seems more satisfying than any of the single variable explanations.

This is not to say that the measures used in [Halpern & Pearl, 2001b] are not useful. In fact they can be very useful. We simply note that our definitions allow a further analysis that is useful in other situations. They prove especially useful when we are looking for a sufficient or sustaining explanations. Because our definitions can handle different notions of cause, the explanations they produce can be more informative than the explanations formed when using the methods in [Halpern & Pearl, 2001b], where they are restricted to one weak form of cause.

8 Conclusion

We demonstrated that the intuitive choice of causal criteria is subjective, and is influenced psychological factors not representable in a causal model. The criteria, once chosen however can be evaluated from the information contained in the causal model. We presented new criteria covering the different notions of causation. Fundamental among these is the definition of strong sufficiency, which provides a simple and intuitively satisfying definition for actually participating in the active causal process that generated the effect. These definitions avoid the problems of previous attempts, and allow for a more refined discussion of actual causality. Additionally, they provide a foundation for evaluating explanations, which proves to be more satisfactory than previous attempts to use functional causal models for explanation.

References

[Chajewska & Halpern, 1997] Chajewska, U., and Halpern, J. 1997. Defining explanation in probabilistic systems. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, 62–71.

[Hall, 2002] Hall, N. 2002. Two concepts of causation. In *Casuation and Counterfactuals*. MIT Press.

[Halpern & Pearl, 2001a] Halpern, J., and Pearl, J. 2001a. Causes and explanations: A structural-model approach–part I: Causes. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, 194–202.

[Halpern & Pearl, 2001b] Halpern, J., and Pearl, J. 2001b. Causes and explanations: A structural-model approach–part II: Explanations. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 27–34.

[Hopkins & Pearl, 2003] Hopkins, M., and Pearl, J. 2003. Clarifying the usage of structural models for comonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

[Lewis, 1986] Lewis, D. 1986. Causation. In *Philosophical Papers*, volume II, 159–213. Oxford University Press.

[Pearl, 2000] Pearl, J. 2000. *Causality*. Cambridge University Press.

[5] Note that this is not always the case. This example was specifically crafted to exhibit such behavior.

[6] In the ASC definition of explanation, it "causes" would mean is a weak cause. We write it in a general form, because these measures are applicable to the new method as well