In *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, AAAI Press, Menlo Park, CA, pp. 83-89, 2003.

TECHNICAL REPORT
R-311
March 2003

Clarifying the Usage of Structural Models for Commonsense Causal Reasoning

Mark Hopkins and Judea Pearl
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
{mhopkins,judea}@cs.ucla.edu

Abstract

Recently, Halpern and Pearl proposed a definition of actual cause within the framework of structural models. In this paper, we explicate some of the assumptions underlying their definition, and re-evaluate the effectiveness of their account. We also briefly contemplate the suitability of structural models as a language for expressing subtle notions of commonsense causation.

Introduction

Providing an adequate definition for when one event causes another has been a troublesome issue in philosophy for centuries (Sosa & Tooley 1993). To partially illustrate the difficulties involved, consider the following example:

Example (Firing Squad) There are two riflemen (R_1 and R_2) in a firing squad. On their captain's order, they both shoot simultaneously and accurately. The prisoner dies.

From this story, we can ask causal queries such as: did R_1's shot cause the prisoner's death? We can also ask whether the captain's order caused the prisoner's death. For both of these queries, a satisfactory account of causation should answer "yes," to agree with our intuition. Most accounts attempt to capture the concept of causation by considering sufficiency of the candidate cause, necessity of the candidate cause, or some hybrid of the two. For instance, the captain's order is both necessary and sufficient for the prisoner's death (given that we assume the riflemen always obey the captain's order). Alternatively, R_1's shot is sufficient, but not necessary, for the prisoner's death to occur. Additionally, it is possible to derive candidates that are sufficient for the effect, but that we would not consider causes.

In a recent paper (Halpern & Pearl 2001), Halpern and Pearl propose a definition of cause (which they term *actual cause*) within the framework of structural causal models. Specifically, they express stories as a structural causal model (or more accurately, a causal world), and then provide a definition for when one event causes another, given this model of the story. Their definition is primarily necessity–based. The main idea is that a candidate C is an actual cause of an effect E when C and E have both occurred, and there exists some *counterfactual contingency* W under which E is *counterfactually dependent* on C. By this, we mean that *had* W *occurred*, C and E would still have occurred, but E would not have occurred were it not for C. For instance, in the above example, the prisoner's death is counterfactually dependent on R_1's shot, under the counterfactual contingency that R_2 did not shoot his rifle. Halpern and Pearl impose a few restrictions such that not every contingency W can be considered (we discuss this in greater detail in the next section).

In this paper, we make the following contributions:

1. We explicate some of assumptions underlying the usage of causal models for the commonsense causal reasoning addressed by Halpern and Pearl in (Halpern & Pearl 2001). Halpern and Pearl do not elaborate on how stories are mapped into appropriate causal models; rather, they simply use models that "seem right." Spelling out the formal implications of a given causal model is especially crucial in light of the fact that different (seemingly sensible) models of the same story can yield different answers for identical queries. This analysis also helps us to determine what problem Halpern and Pearl are actually addressing with their definition. Namely, what kind of information does their definition assume is encoded in the model? On what basis do their conclusions about causation rest?
2. We evaluate the counterfactual strategy that provides the foundation for Halpern and Pearl's definition. We provide evidence that this strategy (despite the attempts at restricting viable counterfactual contingencies) is far too permissive.
3. After highlighting some problematic aspects of Halpern and Pearl's account, we briefly address whether this is attributable to the framework in which it is based: the language of structural models. Essentially a propositional language, we consider whether the ontological commitment of this framework limits the ability to effectively capture notions that are essential to a valid account of causation.

Structural Models and a Definition of Cause

Halpern and Pearl define their notion of causation within the language of structural models. Essentially, structural models are a system of equations over a set of random variables. We

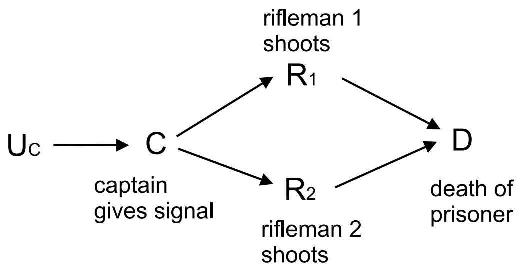

Figure 1: Causal model for the Firing Squad scenario. All variables are propositional. $C = U_C$; $R_1 = C$; $R_2 = C$; $D = R_1 \vee R_2$.

can divide the variables into two sets: endogenous (each of which have exactly one structural equation that determines their value) and exogenous (whose values are determined by factors outside the model, and thus have no corresponding equation).

First we establish some preliminaries. We will generally use upper-case letters (e.g. X, Y) to represent random variables, and the lower-case correspondent (e.g. x, y) to represent a particular value of that variable. $Dom(X)$ will denote the domain of a random variable X. We will use bold-face upper-case letters to represent a set of random variables (e.g. $\mathbf{X}$,$\mathbf{Y}$). The lower-case correspondent (e.g. $\mathbf{x}$,$\mathbf{y}$) will represent a value assignment for the corresponding set.

Formally, a *structural causal model* (or *causal model*) is a triple $(\mathbf{U}, \mathbf{V}, \mathbf{F})$, in which $\mathbf{U}$ is a finite set of exogenous random variables, $\mathbf{V}$ is a finite set of endogenous random variables (disjoint from $\mathbf{U}$), and $\mathbf{F} = \{F_X | X \in \mathbf{V}\}$ where F_X is a function $Dom(\mathbf{R}) \to Dom(X)$ that assigns a value to X for each setting of the remaining variables in the model $\mathbf{R} = \mathbf{U} \cup \mathbf{V} \backslash \{X\}$. For each X, we can define $\mathbf{PA}_X$, the *parent set* of X, to be the set of variables in $\mathbf{R}$ that can affect the value of X (i.e. are non-trivial in F_X). We also assume that the domains of the random variables are finite.

Causal models can be depicted as a *causal diagram*, a directed graph whose nodes correspond to the variables in $\mathbf{U} \cup \mathbf{V}$ with an edge from Y to $X \in \mathbf{V}$ iff $Y \in \mathbf{PA}_X$.

Example In Figure 1, we see the firing squad scenario expressed as a causal model. Here, $\mathbf{U} = \{U_C\}$ and $\mathbf{V} = \{C, R_1, R_2, D\}$. All variables are propositional, with value 1 indicating a true proposition, and value 0 indicating that the proposition is false (this will be the convention for most causal models given as examples in this paper).

If we assume a particular value for the exogenous variables $\mathbf{U}$ (referred to as a *context*), then the resulting causal model is called a *causal world*. We generally assume that any particular value for $\mathbf{U}$ uniquely determines the values of the variables in $\mathbf{V}$. This always happens when the causal diagram is acyclic (such causal models are called *recursive*). Causal worlds are of interest since they represent a specific situation, while causal models represent a more general scenario. For instance, if we assume that $U_C = 1$ in our firing squad causal model, then the resulting causal world describes our story (given in the introduction). The more general model allows for the situation in which the captain does not signal.

To handle counterfactual queries, we define the concept of *submodels*. Given a causal model $M = (\mathbf{U}, \mathbf{V}, \mathbf{F})$, $\mathbf{X} \subseteq \mathbf{V}$, $\mathbf{x} \in Dom(\mathbf{X})$, the *submodel* of M under *intervention* $\mathbf{X} = \mathbf{x}$ is $M_{\mathbf{X}=\mathbf{x}} = (\mathbf{U}, \mathbf{V}, \mathbf{F}_{\mathbf{X}=\mathbf{x}})$, where $\mathbf{F}_{\mathbf{X}=\mathbf{x}} = \{F_Y | Y \in \mathbf{V} \backslash \mathbf{X}\} \cup \{\mathbf{X} = \mathbf{x}\}$. Intuitively, the submodel fixes the values of the variables in $\mathbf{X}$ at $\mathbf{x}$ (i.e, their values are no longer determined by their parents' values). Consequently, the values of the remaining variables represent what values they *would have had* if $\mathbf{X}$ had been $\mathbf{x}$ in the original model. $M_{\mathbf{X}=\mathbf{x}}$ and $\mathbf{F}_{\mathbf{X}=\mathbf{x}}$ are typically abbreviated $M_{\mathbf{x}}$ and $\mathbf{F}_{\mathbf{x}}$. The value of variable $Y \in \mathbf{V}$ in submodel $M_{\mathbf{x}}$ (under context $\mathbf{u}$) is represented as $Y_{M_{\mathbf{x}}}(\mathbf{u})$ (or simply $Y_{\mathbf{x}}(\mathbf{u})$).

Example (interventions) Consider the firing squad causal model under context $\mathbf{u} = \{U_C = 1\}$ and the question: would the prisoner be dead if we *make sure* that R_1 does not fire his gun? This corresponds to evaluating $D_{R_1=0}(\mathbf{u})$. In this case, the captain still signals, so rifleman 2 still shoots. Thus $D_{R_1=0}(\mathbf{u}) = 1$, and we conclude that the prisoner still dies in this counterfactual scenario.

Equipped with this background, we can now proceed to Halpern and Pearl's definition of actual cause:

Definition 1 *Let* $M = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ *be a causal model. Let* $\mathbf{X} \subseteq \mathbf{V}$, $\mathbf{Y} \subseteq \mathbf{V}$. $\mathbf{X} = \mathbf{x}$ *is an* actual cause *of* $\mathbf{Y} = \mathbf{y}$ *(denoted* $\mathbf{x} \propto \mathbf{y}$*) in a causal world* $(M, \mathbf{u})$ *if the following three conditions hold:*

(AC1) $\mathbf{X}(\mathbf{u}) = \mathbf{x}$ *and* $\mathbf{Y}(\mathbf{u}) = \mathbf{y}$.

(AC2) There exists $\mathbf{W} \subseteq \mathbf{V} \backslash \mathbf{X}$ *and values* $\mathbf{x}' \in Dom(\mathbf{X})$ *and* $\mathbf{w} \in Dom(\mathbf{W})$ *such that:*

(a) $\mathbf{Y}_{\mathbf{x}'\mathbf{w}}(\mathbf{u}) \neq \mathbf{y}$.

(b) $\mathbf{Y}_{\mathbf{x}\mathbf{w}}(\mathbf{u}) = \mathbf{y}$.

(c) $\mathbf{Y}_{\mathbf{x}\mathbf{w}\mathbf{z}}(\mathbf{u}) = \mathbf{y}$*, for all* $\mathbf{Z} \subseteq \mathbf{V} \backslash (\mathbf{X} \cup \mathbf{W})$ *such that* $\mathbf{z} = \mathbf{Z}(\mathbf{u})$.

(AC3) $\mathbf{X}$ *is minimal; no subset of* $\mathbf{X}$ *satisfies conditions AC1 and AC2.*

Intuitively, $\mathbf{x}$ is an actual cause of $\mathbf{y}$ if (AC1) $\mathbf{x}$ and $\mathbf{y}$ are the "actual values" of $\mathbf{X}$ and $\mathbf{Y}$ (i.e. the values of $\mathbf{X}$ and $\mathbf{Y}$ under no intervention), and (AC2) under some counterfactual contingency $\mathbf{w}$, the value of $\mathbf{Y}$ is dependent on $\mathbf{X}$, such that setting $\mathbf{X}$ to its actual value will ensure that $\mathbf{Y}$ maintains its "actual value," even if we force all other variables in the model back to their "actual values." (AC3) is a simple minimality condition.

Example In the firing squad example, we see that $R_1 = 1$ (the first rifleman's shot) is indeed an actual cause of $D = 1$ (death), since [AC1]$R_1(\mathbf{u}) = 1$, $D(\mathbf{u}) = 1$, [AC2(a)]$D_{R_1=0,R_2=0}(\mathbf{u}) = 0$, [AC2(b)]$D_{R_1=1,R_2=0}(\mathbf{u}) = 1$, and [AC2(c)]$D_{R_1=1,R_2=0,C=1}(\mathbf{u}) = 1$. Here, our $\mathbf{w}$ is $R_2 = 0$.

One useful theorem (proven by (Eiter & Lukasiewicz 2001) and (Hopkins 2002)) demonstrates that the minimality condition of the definition forces every actual cause to be an event over a single random variable (also called a *primitive event*).

Theorem 2 *Let $M = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ be a causal model. Let $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ and $\mathbf{x} \in Dom(\mathbf{X})$, $\mathbf{y} \in Dom(\mathbf{Y})$. If $\mathbf{x} \propto \mathbf{y}$ under $\mathbf{u}$, then $\mathbf{X}$ is a singleton.*

Explicating Assumptions

In the next two sections, we will attempt to answer the following question: what information does a causal world encode? Specifically, what information are we using when we decide that event A causes event B, given a causal world?

Halpern and Pearl (Halpern & Pearl 2001) give suggestive comments that help illuminate the path, but stop short of providing details:

> It may seem strange that we are trying to understand causality using causal models, which clearly already encode causal relationships. Our reasoning is not circular. Our aim is not to reduce causation to noncausal concepts, but to interpret questions about causes of specific events in fully specified scenarios in terms of generic causal knowledge such as what we obtain from the equations of physics.

Essentially, a causal world encodes information from two sources:

1. The choice of endogenous variables $\mathbf{V}$.
2. The set of structural equations over $\mathbf{V}$.

The latter item, the set of structural equations, gives us all counterfactual information regarding the variables of interest (i.e. the endogenous variables of the causal world). With it, we can answer any question of the form: "If $\mathbf{V}$ had been $\mathbf{v}$, what would the value of X have been?" Clearly, this counterfactual information is a cornerstone of Halpern and Pearl's definition. Furthermore, given a set of endogenous variables and a story, it is generally straightforward to formulate the appropriate structural equations.

Nevertheless, to answer questions of the form: "Did event $X = x$ cause event $Y = y$?", we need more information than simply a set of structural equations over X, Y, and an arbitrary set of other variables. To take an example, consider the following story, taken from (Halpern & Pearl 2000):

Example (Rock) Billy and Suzy both throw rocks at a bottle. Suzy's arm is better than Billy's, so her rock gets to the bottle first and shatters it. Billy's throw was perfectly accurate, so his rock would have shattered the bottle had Suzy's missed.

If we take the set of propositional random variables BT (Billy Throws), ST (Suzy Throws), and BS (Bottle Shattered), we can see that the structural equations over these three variables are equivalent to the structural equations over R_1, R_2, and D in the firing squad example. Nevertheless, in this instance we would like to conclude that $BT = 1$ is not a cause of $BS = 1$, whereas for the isomorphic query "Is $R_1 = 1$ a cause of $D = 1$?" in the firing squad example, we would like to conclude the opposite.

Thus we arrive at the second (and murkier) piece of information encoded by a causal world – the choice of endogenous variables.

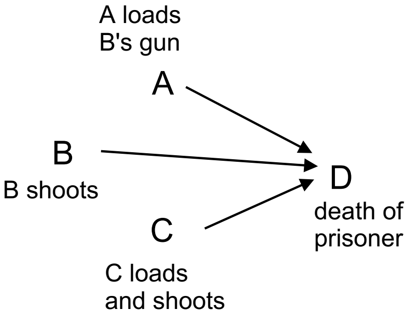

Figure 2: Causal diagram for the Loader scenario.

Variable Selection

Halpern and Pearl are well aware of the sensitivity of their definition to the choice of endogenous variables for the causal world. They are vague, however, when it comes to defining the semantics accorded to a particular choice of endogenous variables. In this section, we consider what their definition of cause assumes about local relationships between variables. Specifically, we examine a single variable in the causal world and its parents, and consider when a parent is considered to be a cause of its child, under their definition. Given local criteria, we can then establish guidelines about what constitutes an "incorrect" choice of endogenous variables.

Consider a variable Y, one of its parents X, and its structural equation $F_Y : Dom(\mathbf{PA}_Y) \rightarrow Dom(Y)$. Suppose that $X(\mathbf{u}) = x$ and $Y(\mathbf{u}) = y$. We want to extract conditions from F_Y that imply that x is a cause of y.

To do this, it is convenient to express the set of parent value assignments $\mathbf{p} \in Dom(\mathbf{PA}_Y)$ such that $F_Y(\mathbf{p}) = y$ as a (propositional) logical sentence (which we will denote $\Delta(Y = y)$) over literals of the form $Z = z$, for $Z \in \mathbf{PA}_Y$ and $z \in Dom(Z)$.

We illustrate this with a modified version of the firing squad example.

Example (Loader) For a firing squad consisting of shooters B and C, it is A's job to load B's gun. C loads and fires his own gun. On a given day, A loads B's gun. When the time comes, B and C shoot the prisoner.

Suppose we choose to model this as a causal world (whose causal diagram is pictured in Figure 2) over the following four propositional random variables: A (true iff A loads B's gun), B (true iff B shoots) C (true iff C loads and shoots), and D (true iff the prisoner dies).

In the model, $D(\mathbf{u}) = 1$, so consider the set of $\mathbf{p} \in Dom(\mathbf{PA}_D)$ such that $F_D(\mathbf{p}) = 1$. We can express this as the following logical sentence: $\Delta(D = 1) = (A = 1 \wedge B = 1 \wedge C = 1) \vee (A = 1 \wedge B = 1 \wedge C = 0) \vee (A = 0 \wedge B = 1 \wedge C = 1) \vee (A = 1 \wedge B = 0 \wedge C = 1) \vee (A = 0 \wedge B = 0 \wedge C = 1)$, where each conjunct is a full instantiation $\mathbf{p}$ of the parents of D such that $F_D(\mathbf{p}) = 1$.

More interesting is the prime implicant form of this sentence. Recall that an implicant of a sentence Δ is a term (conjunction of literals) that entails Δ. A prime implicant of Δ is an implicant of Δ that does not entail any other implicant of Δ (besides itself). The prime implicant form of a sentence is the disjunction of all of its prime implicants. Note that this form is unique.

To continue our example, the prime implicant form of $\Delta(D = 1)$ is $(A = 1 \wedge B = 1) \vee (C = 1)$. Observe that $(A = 1 \wedge B = 1 \wedge C = 1)$ and $(A = 1 \wedge B = 1 \wedge C = 0)$, which are both implicants of $\Delta(D = 1)$, both entail implicant $(A = 1 \wedge B = 1)$, thus they do not appear in the prime implicant form.

The prime implicant form is interesting because it lends itself to a natural causal interpretation. Each prime implicant is a minimal set of value assignments to parents of FF that ensures that $FF = 1$. Specifically in this example, the prisoner will be guaranteed to die if A loads B's gun *and* B shoots, or alternatively, if C loads and shoots.

This logical form is reminiscent of the causal criterion laid out by John Mackie called the INUS condition (Mackie 1965). Consider an event C and an effect E. Rather than requiring that C is either necessary or sufficient (or both) to achieve E, Mackie instead requires that C is an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result. For instance, A loading B's gun is a necessary part of a sufficient condition to ensure the prisoner's death. In terms of the prime implicant logical form, sufficient conditions map to implicants. For instance, $A = 1 \wedge B = 1$ is a sufficient condition for $D = 1$. Furthermore, since $A = 1 \wedge B = 1$ is a *prime* implicate (hence no subset of its conjuncts is an implicate), we observe that both $A = 1$ and $B = 1$ are necessary parts of this sufficient condition. Hence any literal that appears in a prime implicate satisfies the INUS condition.

Interestingly, we can prove the following:

Theorem 3 *Consider a variable Y, one of its parents X, and its structural equation $F_Y : Dom(\mathbf{PA}_Y) \rightarrow Dom(Y)$. Suppose that $X(\mathbf{u}) = x$ and $Y(\mathbf{u}) = y$. Then if $X = x$ appears as an literal in any prime implicate of $\Delta(Y = y)$, then x causes y according to Halpern and Pearl's definition.*

Proof Clearly AC1 holds, since $X(\mathbf{u}) = x$ and $Y(\mathbf{u}) = y$. Furthermore AC3 holds trivially. It remains to show AC2 holds. We will prove an equivalent contrapositive. Specifically, we will suppose that AC2 does not hold, and prove that under this supposition, $X = x$ cannot appear as a literal in any prime implicate of $\Delta(Y = y)$.

Suppose, then, that AC2 does not hold. This implies that for the set of parents of Y not including X, $\mathbf{W} = \mathbf{PA}_Y \backslash \{X\}$, there is no instantiation $\mathbf{w} \in Dom(\mathbf{W})$ such that $Y_{x\mathbf{w}}(\mathbf{u}) = y$ and $Y_{x'\mathbf{w}}(\mathbf{u}) \neq y$ for $x' \neq x$. Hence for any implicate of $\Delta(Y = y)$ of the form $(X = x) \wedge C$, where C is a conjunction of literals, $(X = x') \wedge C$ is also an implicate, for every $x' \in Dom(X)$. Hence C is also an implicate, which means that any implicate containing the literal $X = x$ is not prime. ▮

The implications of the above theorem are surprising. It is encouraging that locally speaking, Halpern and Pearl's definition resembles the intuitively appealing criterion of Mackie. At the same time, the theorem exposes the over-permissiveness of Halpern and Pearl's definition. Observe that for x to cause y, $X = x$ need only appear in some prime implicate of $\Delta(Y = y)$. There is no requirement for $X = x$ to appear in a *satisfied* prime implicate. Consider the following alteration of the loader example for emphasis.

Example We have the same situation as in the Loader example above, except now B elects not to shoot. A still loads B's gun, C still loads and shoots, and the prisoner still dies.

This story can be modeled the same way as above (see Figure 2), except now $B(\mathbf{u}) = 0$. The prime implicate form of $D = 1$ is still $(A = 1 \wedge B = 1) \vee (C = 1)$. Notice that $A(\mathbf{u}) = 1$ and that $A = 1$ appears in $\Delta(D = 1)$, hence by Theorem 3, Halpern and Pearl's definition classifies $A = 1$ as a cause of $D = 1$. (alternatively, we can observe that the intervention $B = 1, C = 0$ satisfies AC2 of their definition).

Halpern and Pearl's definition classifies A loading B's gun as a cause of the prisoner's death because it is a necessary part of a sufficient condition to cause the prisoner's death, but it completely disregards the fact that this sufficient condition did not occur in the actual situation we are concerned with! This is what prompts the definition to draw such a counterintuitive conclusion. Given this observation, it is trivial to construct any number of situations for which Halpern and Pearl's definition returns an answer contrary to intuition.

In fact, locally speaking, Halpern and Pearl's definition is even more permissive than Theorem 3 suggests. The following counterexample demonstrates that the converse of Theorem 3 does not hold.

Example Suppose we have a causal world with three random variables A, B, C such that $Dom(C) = \{0, 1\}$, $Dom(A) = Dom(B) = \{0, 1, 2\}$. Define the structural equations such that $B = A + 1(mod3)$ and such that $C = 1$ iff $(A = 0 \wedge B = 0) \vee B = 2$ (note that this is $\Delta(C = 1)$). In the actual world, let $A = 1$ (hence $B = 2$ and $C = 1$). According to Halpern and Pearl's definition, $A = 1$ is a cause of $C = 1$ (letting $\mathbf{W} = \emptyset$). Observe that $A = 1$ does not appear as a conjunct of any prime implicate of $C = 1$.

Counterexamples can also be constructed for the case where all variables are restricted to be propositional.

These observations shed light on what is considered to be an appropriate choice of endogenous variables, under Halpern and Pearl's definition. Namely, the variables must be chosen in such a way that the structural equations have a *strong causal semantics*. By this, we mean that every prime implicate for event $X = x$ must *unconditionally* be a cause of $X = x$. Observe that every term of a satisfied prime implicate of $\Delta(X = x)$ will be considered to be a cause of $X = x$, regardless of the values of the other variables in the model. Often, structural equations do not possess these strong causal semantics.

As one example, we can revisit the Rock story. Here we observe that although $\Delta(BS = 1)$ is $(ST = 1) \vee (BT = 1)$, it is not the case that $BT = 1$ is unconditionally a cause of $BS = 1$, since $BT = 1$ is not a cause of $BS = 1$ in the event that $ST = 1$.

An alternative selection of variables for the Rock story adds two additional variables to our previous choice: SH

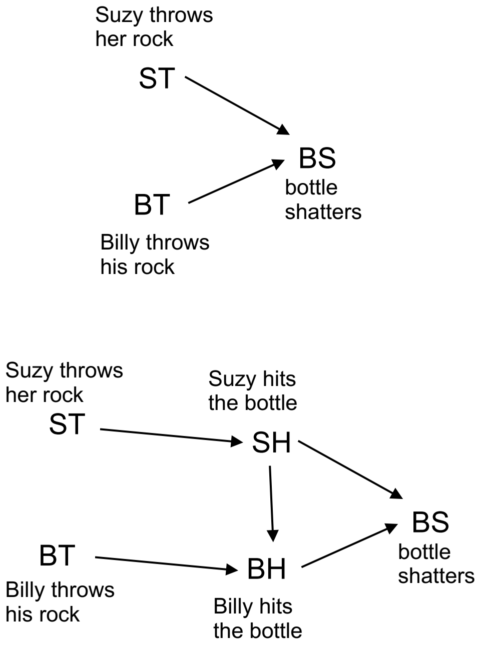

Figure 3: Causal diagrams for two formulations of the Rock scenario.

(Suzy Hits), which is true if and only if Suzy's rock hits the bottle, and BH (Billy Hits), which is analogous. The causal diagram for this alternate variable choice is depicted in Figure 3. In their paper, Halpern and Pearl consider both of these models, and reject the former because it does not carry all the information from the story. Notably, the latter is the only one which strong causal semantics. Specifically, notice that if either Billy's rock hits the bottle or if Suzy's rock hits the bottle, then both are causes of the shattering, no matter what the values of the other variables in the model happen to be.

From this perspective, Halpern and Pearl's definition can be viewed as a means of extracting more complex causal relationships from simple causal relationships. The implicit assumption is that every local interaction is causal (in the strong sense expressed above), and given this, the problem is to extract causal relationships between events that are not directly linked.

The Counterfactual Strategy

We now turn our attention to the counterfactual strategy used by Halpern and Pearl's definition and evaluate its validity.

A fuzzy way to define causality (Yablo 2000) between two events is to say: event C causes event E iff for some *appropriate* G, E is counterfactually dependent on C when we hold G fixed. Here, G is any imaginable statement about the world. For instance, in the Rock story, if we hold fixed that Billy does not throw his rock, then the bottle being shattered is counterfactually dependent on Suzy throwing her rock.

As an example of an inappropriate G, consider the fact that the occurrence of a full moon is counterfactually dependent on whether you brushed your teeth this morning if we hold it fixed that *a full moon occurs only if you brushed your teeth this morning*. To consider a less trivial example, in the Suzy-Billy story, the bottle being shattered is counterfactually dependent on Billy throwing his rock, given that we hold fixed that Suzy does not throw her rock (still we should not conclude that Billy's throw causes the bottle to shatter).

Thus the key element of any counterfactual strategy is how it identifies which G are appropriate to hold fixed. Intuitively, we would like to screen out the other causes of E, such that the only causal mechanism responsible for E is C. Unfortunately, issues such as preemption make it extremely difficult to systematically define which choices of G are appropriate.

In Halpern and Pearl's definition, they essentially allow G to be anything that can be expressed as a conjunction of primitive events involving any endogenous variable which is neither a cause nor an effect variable. Naturally, this definition is too permissive (basically it allows for any imaginable G, given a suitable choice of endogenous variables), thus they make an effort (through AC2(c)) to restrict the permissiveness of the definition. Unfortunately, this restriction has two defects. Firstly, it is non-intuitive. Secondly, it is not restrictive enough.

This permissiveness was pointed out by Halpern and Pearl using the following example.

Example (Loanshark) Larry the Loanshark contemplates lurking outside (LL) of Fred's workplace to cut off his finger(LC), as a warning to him to repay his loan quickly. Something comes up, however, so he does not do so ($LL = 0$ and $LC = 0$). That same day, Fred has his finger severed ($FS = 1$) by a machine at the factory. He is rushed to the hospital, where the finger is reattached, so if Larry had shown up, he would have missed Fred. At day's end, Fred's finger is functional ($FF = 1$), which would not have been true had Larry shown up and Fred not had his accident.

In this case, Halpern and Pearl's definition unintuitively classifies Fred's accident as a cause of his finger being functional at day's end. To remedy this problem, they propose a scheme wherein "fanciful contigencies" are excluded from consideration. Thus, given that the prior odds of Larry showing up are slim, they conclude that Fred's accident is not a cause of his finger's functionality. Nevertheless, this is a rather unsatisfactory remedy to the problem. Consider what happens if the story is amended such that Larry fully intends to show up at the factory, but is improbably struck by lightning such that he doesn't arrive. Hence the prior probability of $LL = 1$ is high, and yet we still intuitively would like to conclude that Fred's accident did not cause his finger's functionality. In fact, we would *only* want to conclude that Fred's accident was a cause of his finger being functional at day's end in the event that Larry shows up *in actuality*.

In fact, the problem illustrated by this example is simply a representative of any number of situations where it is possible to choose an inappropriate G to keep fixed.

Example (Bomb) Billy puts a bomb under Suzy's chair. Later, Suzy notices the bomb and flees the room. Still later, Suzy has a prearranged medical checkup and is pronounced healthy (Yablo 2000).

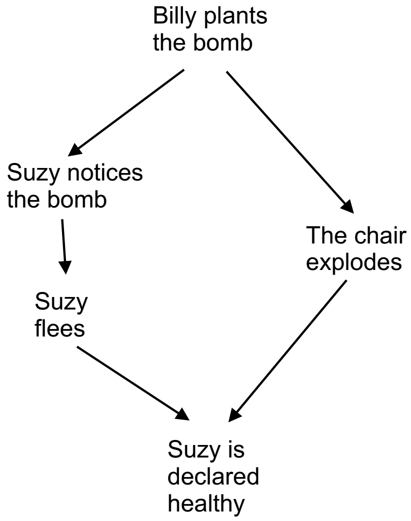

Figure 4: Causal diagram for the Bomb scenario.

Given the causal world (with strong causal semantics) whose causal diagram is shown in Figure 4, Halpern and Pearl's definition classifies Billy putting the bomb under Suzy's chair as a cause of Suzy being pronounced healthy. Clearly this is an unintuitive result, however if we hold fixed the fact that Suzy's chair explodes, then Suzy being pronounced healthy is counterfactually dependent on Billy planting the bomb (otherwise, she would not have any warning that the chair would explode and would not flee).

The moral of the story is that a definition of cause based exclusively on counterfactual contingencies must be considerably less permissive than Halpern and Pearl's definition. Whether it is feasible to propose such a definition at all in the structural model framework (without being overly restrictive) is considered in the next section.

Ontological Concerns

The question remains: although Halpern and Pearl's definition is problematic, is it feasible to propose a satisfactory definition of cause within the structural model framework? Consider the following example from Jonathan Schaffer (Yablo 2000) which parallels the Rock scenario.

Example (Magic) Imagine that it is a law of magic that the first spell cast on a given day matches the enchantment that midnight. Suppose that at noon Merlin casts a spell (the first that day) to turn the prince into a frog, that at 6pm Morgana casts a spell (the only other that day) to turn the prince into a frog, and that at midnight the prince becomes a frog.

Intuitively, Merlin's spell is a cause of the prince's transformation and Morgana's is not. In this case, although there is preemption, there are no intermediating events that we can really play with and model. Spells work directly, and without Merlin's spell, the prince's transformation would have occurred at precisely the same time and in the same manner. Hence it is far from clear how we could model this story appropriately with a structural model. One concise way to express this story uses first-order constructs. For example, we could neatly express the rule that a spell works iff *there does not exist* a previous spell cast that day.

Perhaps then, we should be looking beyond the ontological commitment of structural models (which essentially are built on propositional logic) to richer languages in which subtle points of causality can be more easily expressed. This is not to say that such a definition would be impossible within the structural model framework, but the framework does seem to overly limit our ability to do so.

Besides adding first-order constructs, we could also benefit from adding other features to the structural model framework, described briefly here (for a more detailed discussion, see (Hopkins & Pearl 2002)):

Temporal constructs: Time plays a critical role in our perceptions of causality, and yet it has no explicit representation in structural models. Time can be modeled within this framework in the similar fashion as dynamic Bayesian networks, yet this can often lead to counterintuitive conclusions. It is important to distinguish when an event causes another event, as opposed to an event hastening another event (as in a strong wind that causes Suzy's rock to hit the bottle earlier than anticipated, but does not cause the bottle to shatter). It is difficult to phrase this fine distinction within the structural model framework. (Halpern & Pearl 2000)

Distinction between condition and transition: Consider the difference between an enduring condition (e.g., the man is dead) versus a transitional event (e.g., the man dies). While we may consider a heart attack the cause of a man dying, we may be reluctant to consider the same heart attack as the cause of the man *being dead* in the year 3000. Such distinctions can be modeled by adding specialized classes of random variables to the structural model framework, but such classes are not part of the framework as it stands.

Distinction between presence and absence of an event: We often apply stronger criteria when deciding whether the absence of an event (e.g. a bystander's inaction) is a cause, as opposed to the presence of an event (e.g. a rifleman's shot). The underlying issue here is a matter of production – the latter plays an active role in bringing about an effect (e.g. a victim's death), while the former does not. In the structural model framework (where all events are value assignments to random variables), such distinctions are lost. We can regain such distinctions by adding *distinguished values* to the structural model framework (e.g., giving the value assignment 0 a special semantics).

Conclusions and Outlook

In this paper, we have attempted to explicate some of the latent assumptions made in (Halpern & Pearl 2001) and to evaluate their account of causality on two bases:

1. The effectiveness of their strategy of counterfactual dependence modulo a set of facts which are kept fixed.
2. The suitability of the structural model framework to capture the subtleties involved in commonsense causation.

In the process, we have highlighted fundamental stumbling blocks for their definition. One of the key results of this paper relates their definition to the prime implicate form of the structural equations, which lays bare some of the definition's problematic aspects. Furthermore, we have provided further evidence that the strategy of counterfactual dependence employed is much too permissive.

The most promising direction for future research seems to be finding ways to embed a definition of actual causation in a richer, more expressive language. Ideally, these definitions would benefit from the positive features of Halpern and Pearl's account.

References

Eiter, T., and Lukasiewicz, T. 2001. Complexity results for structure-based causality. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, 35–40. San Francisco, CA: Morgan Kaufmann.

Halpern, J., and Pearl, J. 2000. Causes and explanations: A structural-model approach. Technical Report R–266, UCLA Cognitive Systems Laboratory.

Halpern, J., and Pearl, J. 2001. Causes and explanations: A structural-model approach – part i: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 194–202. San Francisco, CA: Morgan Kaufmann.

Hopkins, M., and Pearl, J. 2002. Causality and counterfactuals in the situation calculus. Technical Report R–301, UCLA Cognitive Systems Laboratory.

Hopkins, M. 2002. A proof of the conjunctive cause conjecture in 'causes and explanations: A structural-model approach'. Technical Report R–306, UCLA Cognitive Systems Laboratory.

Mackie, J. L. 1965. Causes and conditions. *American Philosophical Quarterly* 2(4):245–255, 261–264.

Sosa, E., and Tooley, M. 1993. *Causation*. Oxford University Press.

Yablo, S. 2000. Advertisement for a sketch of an outline of a proto-theory of causation. *http://www.mit.edu/ yablo/advert.html*.