[This is the long version of a paper contributed to the symposium 'Causation and Bayesian Networks' at the PSA 2002. The paper submitted for publication in *Philosophy of Science* is a much truncated version of this paper]

CAUSAL MODELS, TOKEN-CAUSATION AND PROCESSES

Peter Menzies
Department of Philosophy
Macquarie University
Sydney, Australia
Email: peter.menzies@mq.edu.au

1. INTRODUCTION

It is regrettable that the structural equations and directed graph approach to the study of causation is not better known among philosophers. A partial explanation of this fact is that much of the discussion within the structural equations approach has been of type-causation, whereas philosophers have typically been pre-occupied with issues of token-causation.

However, philosophers now have no excuse for overlooking the important developments in this approach. For Judea Pearl in his landmark work (2000) has developed a theory that applies the structural equations approach to token-causation. Subsequently, in collaboration with David Halpern (2001), he has extended and simplified the theory. More recently, Christopher Hitchcock has written a series of intriguing papers exploring issues to do with token-causation within Pearl's structural equations framework (2001; forthcoming a, b). In this paper I want to continue the philosophical engagement with Pearl's work by examining whether his theory of token-causation can successfully handle a special class of counterexample, and, if not, how it might be modified to do so.

It strikes me that there are many opportunities for mutually beneficial collaboration between philosophers and the advocates of the structural equations approach in applying it to token-causation. Philosophers of causation, especially those working within the counterfactual tradition, have accumulated a stockpile of examples illustrating early and late pre-emption, symmetric overdetermination, trumping, preventive causation and failures of transitivity that might prove interesting test cases for a theory of token-causation. On the other hand, the structural equations framework offers a powerful new formal method that might solve problems that philosophers have found intractable with more orthodox philosophical methods.

2. PEARL'S THEORY OF TOKEN CAUSATION

Let us start by describing the basic features of Pearl's theory of token-causation. One thing many philosophers may find problematic about the theory is that it relativizes token-causal judgments to a causal model. Let us simply grant here that such relativization

makes sense so that we can see how much can be explained in terms of Pearl's theory. (For arguments in favour of such relativization see Menzies 2003a.)

A causal model is an ordered triple <*U*, *V*, *E*>, where *U* is a set of exogenous variables whose values are determined by factors outside the model; *V* is a set of endogenous variables whose values are determined by factors within the model; and *E* is a set of structural equations that express the value of each endogenous variable as a function of the values of the other variables in *U* and *V*.

It is best to illustrate the theory by way of an example. Let us consider an example illustrating what Lewis (1986) called late pre-emption (and subsequently renamed late cutting in Lewis 2000).

> *Example 1: Billy and Suzy*
> Billy and Suzy throw rocks at a bottle. Suzy's rock gets there first, shattering the bottle. Billy's throw arrives at the scene a split second later, encountering nothing but air and flying shards of glass where the bottle used to be. But Billy's throw, like Suzy's, was perfectly accurate so that his rock would have shattered the bottle if Suzy's had not.

A causal model represents this example in terms of a set of selected variables. In all the examples we shall consider the variables will be binary variables that take the values 1 or 0, representing the presence or absence of an event. To represent Example 1, we might choose the four variables *ST*, *BT*, *SH*, *BH,* and *BS*, having the following interpretations:

> *ST* = 1 if Suzy throws a rock, 0 if not.
> *BT* = 1 if Billy throws a rock, 0 if not.
> *SH* = 1 if Suzy's rock hits the intact bottle, 0 if not.
> *BH* = 1 if Billy's rock hits the intact bottle, 0 if not.
> *BS* = 1 if the bottle shatters, 0 if not.

In this example, the variables *ST* and *BT* are exogenous variables that have the following values:

> *ST* = 1
> *BT* = 1

As exogenous variables, their values are assumed to be given and out of the control of the modeler. The values of the endogenous variables—*SH*, *BH*, and *BS*—are determined by structural equations in the set *E* on the basis of the values of the exogenous variables and other endogenous variables. Each endogenous variable has its own structural equation, representing an independent causal mechanism by which its values are determined. A structural equation can be thought of as encoding a battery of non-backtracking counterfactuals. The convention is that the variables appearing on the right hand side of an equation figure in the antecedents of the corresponding counterfactuals, and those appearing on the left hand side figure in the consequents. Each equation asserts several counterfactuals, one for each assignment to the variables that makes the equations true.

The structure of causal mechanism in Example 1 can be captured using the following three structural equations:

$SH = ST$
$BH = BT \ \& \sim SH$
$BS = SH \text{ v } BH$

(Note that we are using familiar symbols from logic to represent mathematical functions on binary variables in the obvious way: $\sim X = 1 - X$; $X \text{ v } Y = \max\{X, Y\}$; $X \ \& \ Y = \min\{X, Y\}$.) The first of these equations encodes two counterfactuals, one for each possible value of *ST*. It asserts that if Suzy threw a rock, her rock hit the bottle; and if she didn't throw a rock, her rock didn't hit the bottle. The second equation encodes four counterfactuals, one for each possible combination of values for *BT* and ~*SH*. It asserts that if Billy threw a rock and Suzy's rock didn't hit the bottle, Billy's hit the bottle; but did not do so if one or other of these conditions was not met. The third equation encodes four counterfactuals, one for each possible combination of values for *SH* and *BH*. It asserts that one or other (or possibly both) of Suzy's rock or Billy's rock hit the bottle, the bottle shattered; but if neither rock hit the bottle, the bottle did not shatter.

The structural equations above can be represented using a directed graph with nodes corresponding to the variables *ST*, *BT*, *SH*, *BH*, and *BS*. An arrow is drawn from one variable *X* to another variable *Y* just in case *X* appears on the right side of an equation and *Y* on the left. In this case, *X* is said to be a parent of *Y*. Exogenous variables have no parents within a model, while endogenous variables do. The graph for Example 1 is depicted in figure 1.

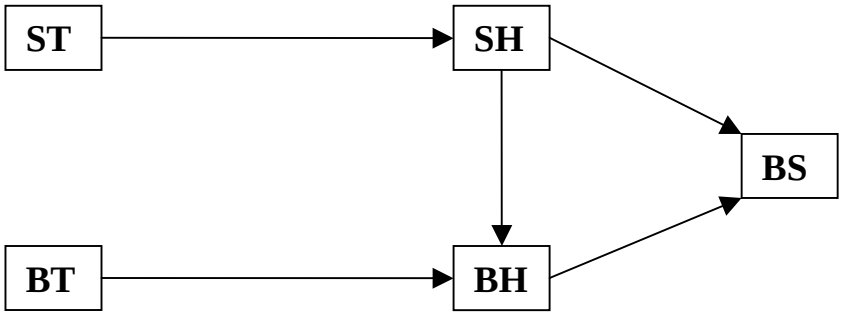

Figure 1

The arrows in this figure tell us that the bottle's shattering is a function of Suzy's rock hitting the bottle and Billy's rock hitting the bottle; that Billy's rock hitting the bottle is a function of Billy's throwing the rock and Suzy's rock hitting the bottle; and that Suzy's rock hitting the bottle is a function of her throwing the rock. (The existence of an arrow from one variable does not always signify a stimulatory connection. For example, the arrow directed from *SH* to *BH* is inhibitory.)

As we have seen, the structural equations encode some counterfactuals directly. However, some counterfactuals that are not directly encoded can be derived from them. More

generally, to evaluate any counterfactual whose antecedent specifies the value of a variable, we replace the equation for the relevant variable with one that stipulates the new value of the variable. For example, to calculate what would have happened if Suzy's rock had not hit the bottle, we replace the structural equation for the endogenous variable *SH* with $SH = 0$, while keeping all the other structural equations unchanged. In effect, this creates a new set of structural equations in which *SH* is an exogenous variable. Graphically, the arrow directed into this variable is removed, while all other arrows remain intact. Instead of this variable having its value causally determined in the normal way, it is 'miraculously' set to its new hypothetical value. The result may be thought of as characterizing the closest possible worlds in which $SH = 0$ is true, or in other words, in which Suzy's rock didn't hit the bottle. With variable *SH* set at the value 0, we can compute that *BH* is equal to 1 and that *BS* is also equal to 1. In other words, in the closest possible worlds in which Suzy's rock didn't hit the bottle, the bottle shattered nonetheless (because Billy's rock would have hit the bottle and shattered it.)

The technique of replacing a structural equation with a hypothetical value set by 'a tiny miracle', as Lewis (1979) expresses it, enables us to capture the idea of counterfactual dependence.

> *Definition 1: Counterfactual Dependence*
> A variable Y counterfactually depends upon a variable X in a causal model M iff it is actually the case that $X = x$ and $Y = y$ and there exist $x' \neq x$ and $y' \neq y$ such that the result of replacing the equation for X with $X = x'$ yields $Y = y'$.

The basic idea of counterfactual theories of causation is to link token-causation with counterfactual dependence or, as in Lewis's (1973) original theory, chains of counterfactual dependence. However, examples of late pre-emption such as Example 1 show that the formulation of a counterfactual theory must proceed with care. For even though Suzy's throwing a rock caused the bottle to shatter, the bottle's shattering did not depend counterfactually on Suzy's throw (nor for that matter on Billy's throw). Hypothetically setting the value of *ST* at 0, while holding the value of *BT* fixed at its original value 1, does not yield a different value for *BS*.

Pearl's theory of token-causation can be called a counterfactual theory. In his (2000), he attempts to capture within the structural equations framework the notion of *quasi-dependence* that Lewis (1986) introduced as a tentative solution—though later discarded—to the difficulties that late pre-emption examples posed his original counterfactual theory. In such examples, Lewis recognized that there is no counterfactual dependence, nor chain of counterfactual dependence, between cause and effect. Nonetheless, he observed that in such cases there would be a counterfactual dependence between cause and effect but for the fact that their surroundings include the presence of a back-up cause. He also observed that, notwithstanding the lack of counterfactual dependence, there is a process connecting cause with effect that is intrinsically like the process that would connect them if their peculiar surroundings were different. In such cases where there is an intrinsic process of this kind but no actual counterfactual dependence, Lewis said there is a quasi-dependence. As a tentative solution to the problem of late pre-emption, Lewis proposed weakening his account of causation to allow

for chains of quasi-dependence, as well as counterfactual dependence, to ground causal relations.

In his (2000) Pearl observes that we might adopt as a test for the existence of a quasi-dependence between a putative cause and effect the following procedure: Look for an intrinsic process connecting putative cause with effect; suppress the influence of their nonintrinsic surroundings, and subject the cause to a counterfactual test to see whether changing the putative cause changes the effect. One way to suppress the influence of the nonintrinsic surroundings, he observes, is to 'freeze' these surroundings as they actually are and then subject the putative cause to the counterfactual test. So let us suppose that the nonintrinsic surroundings are represented by a set of variables $\{W_i\}$. Then Pearl supposes we can suppress the influence of these variables on the causal pathway between a putative cause $X = x$ and an effect $Y = y$ by hypothetically freezing them at their actual values. Then we can subject the event $X = x$ to a counterfactual test and check whether Y would change if X were not x. These ideas are captured by the following definitions, taken from Halpern and Pearl (2001), which improves the definitions of Pearl (2000).

> *Definition 2: A Process*
> A process between two variables X and Y in a model $<U, V, E>$ is an ordered sequence of variables $<X, Z_1,..., Z_n, Y>$ such that each variable in the sequence is in $U \cup V$ and is a parent of its successor in the sequence.

> *Definition 3: Active Causal Process*
> The process $<X, Z_1,..., Z_n, Y>$ is an active causal process relative to the model $<U, V, E>$ iff there exists a (possibly empty) set of variables $\{W_1,...W_m\}$ in $U \cup V \backslash \{X, Z_1,..., Z_n, Y\}$ with actual values $w_1,...,w_m$ such that Y depends counterfactually upon X within the new set of structural equations constructed from E as follows: for each W_i, replace the equation for W_i with a new equation that sets W_i equal to w_i.

These definitions are supposed to capture the notions mentioned in the informal motivating remarks above. For example, the values of the variables in the sequence $<X, Z_1,..., Z_n, Y>$ represent the intrinsic process connecting cause and effect. The set of variables $\{W_i\}$ represent the nonintrinsic surroundings of this process. The definition of an active process says that there is an active process between X and Y if there is a true counterfactual of the form: if the values of the nonintrinsic variables $\{W_i\}$ had been held fixed at their actual value but the value of X had been different, then the value of Y would have been different too. By hypothetically freezing the values of the variables in $\{W_i\}$, any influence these variables have on Y is eliminated. Consequently, the relevant counterfactual isolates the influence of X on Y along the process in question.

Pearl defines several causal notions in terms of these concepts. One important notion is that of actual causation.

> *Definition 4: Actual Causation*
> $X = x$ is an actual cause of $Y = y$ relative to the causal model $<U, V, E>$ iff there is an active causal process from X to Y.

In order to deal with examples of symmetric overdetermination Pearl defines a weaker notion of a *contributory cause*. His definition requires that there be an active causal process between *X* and *Y* under the weakened requirement that *Y* depends counterfactually on *X* under the hypothetical freezing of some possible, not necessarily actual, values of the nonintrinsic variables in the set $\{W_i\}$. As we shall not be considering any examples of symmetric overdetermination, we shall focus on the stronger definition of an actual cause.

Applying these definitions to Example 1, we obtain the right results. Let us consider the model in which $U = \{ST, BT\}$, $V = \{SH, BH, BS\}$, and *E* is the set of five equations listed above. (For convenience I shall include the specification of the values of the exogenous variables in the set of structural equations.) Consider the process <*ST*, *SH*, *BS*>. *BT* and *BH* are not part of this process, so let them belong to the set $\{W_i\}$. Now construct a new set of structural equations, in which the variables *BT* and *BH* are frozen at their actual values. It is easy to see that *BS* counterfactually depends on *ST* in this new set of structural equations. But notice that this is not true for the process <*BT*, *BH*, *BS*> with *ST* and *SH* belonging to the set $\{W_i\}$. In the new set of equations with *ST* and *SH* frozen at their actual value 1, *BS* does not depend counterfactually on *BT*. The difference amounts to the contrast between the following counterfactuals: given that Billy threw and didn't hit the bottle, if Suzy had not thrown, the bottle would not have shattered; on the other hand, given that Suzy threw a rock and hit the bottle, if Billy had not thrown, the bottle would still have shattered.

3. SOME PROBLEM CASES

In this section we shall consider one class of counterexample to Pearl's theory that has been discussed by Halpern and Pearl (2001) and by Christopher Hitchcock (2001; forthcoming a, b). They argue that the theory can be defended against this kind of counterexample. I am much less sanguine about the prospects of the theory in its present form. In this section I shall explore a couple of especially recalcitrant counterexamples as a prelude to introducing an alternative way of elaborating the theory that is more successful in my view in dealing with these counterexamples.

The counterexamples we shall consider are among those usually cited as counterexamples to the transitivity of token-causation. Halpern and Pearl (2001) consider the following example.

> *Example 2: Rail Track Switching*
> 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 it 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 (via the left hand track or via the right hand track).

They invite us to consider what happens when we model this scenario using the following variables and structural equations:

$F = 1$ if you flip the switch, 0 if not.
$T = 1$ if the train goes down left hand track, 0 if it goes down right hand track.
$A = 1$ if the train arrives at point of reconvergence, 0 if not.

$F = 1$
$T = F$
$A = T \text{ v } \sim T$

The directed graph for this model is depicted in Figure 2.

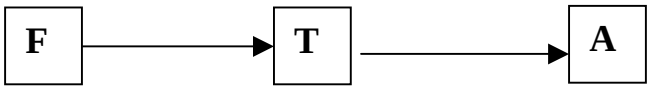

Figure 2

It is easy to see that flipping the switch ($F = 1$) does cause the train to go down the left hand track ($T = 1$), and the train's going down the left track causes it to arrive but flipping the switch does not cause it to arrive ($A = 1$), thanks to the fact that arrival does not counterfactually depend on flipping the switch: the train would arrive whether or not the switch was flipped.

However, Halpern and Pearl note that we can also model the scenario using two variables to represent the track and employing the following modified structural equations:

$LT = 1$ if train goes down left hand track, 0 if not.
$RT = 1$ if train goes down right hand track, 0 if not.

$F = 1$
$LT = F$
$RT = \sim F$
$A = LT \text{ v } RT$

The graph for this new model is depicted in Figure 3.

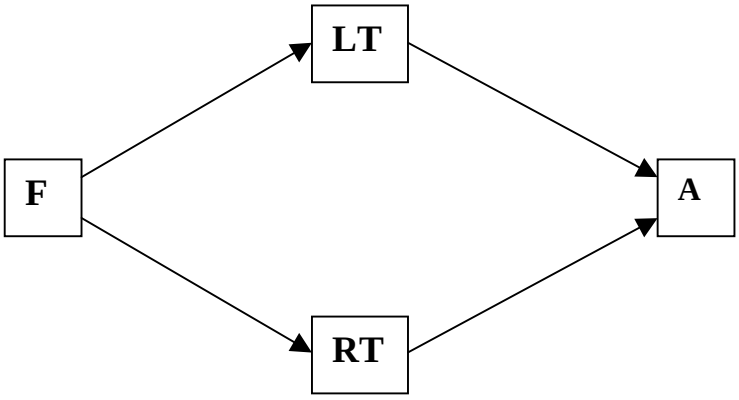

Figure 3

They observe that this change in the model results in a change of causal judgments. This model yields the conclusion that flipping the switch is a cause of train's arrival, since there is an active causal process from flipping the switch to the arrival: holding fixed the fact that the train did not go down the right hand track ($RT = 0$), the train's arrival counterfactually depends on flipping the switch.

Halpern and Pearl attempt to justify the fact that such a change in representation can turn a noncause into a cause. They argue that the change to a four variable from a three variable model represents a profound change in the story. The four variable model depicts the tracks as two independent mechanisms, thus allowing one track to be set to false without affecting the other. Specifically, this permits the mishap of flipping the switch while the left hand track is malfunctioning. Such abnormal eventualities are imaginable and expressible in the four variable model, but not the three variable model. The potential for such eventualities is precisely what renders flipping the switch a cause of the train's arrival in the four variable model.

There are, however, other counterexamples to Pearl's theory in which our causal judgments cannot be explained away in the same manner. Hitchcock has described one such counterexample (forthcoming a).

> *Example 3: Assassin and Bodyguard*
> An assassin slipped poison into the king's coffee. A bodyguard responded to the threat by pouring an antidote into the coffee. The antidote, by itself, is harmless. Nonetheless, the bodyguard would not have put the antidote into the coffee if the assassin had not poisoned it. The king drank the coffee and survived but he would have died if the poison had not been neutralized by the antidote.

Consider what happens when we model the scenario using the following variables and structural equations.

> $A = 1$ if the assassin pours poison into king's coffee, 0 if not.
> $G = 1$ if the bodyguard responds by pouring an antidote into the coffee, 0 if not.
> $S = 1$ if the king survives, 0 if not.
>
> $A = 1$
> $G = A$
> $S = \sim A \vee G$

The directed graph for this model is depicted in Figure 4.

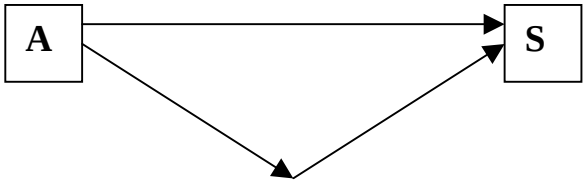

Figure 4

It is easy to check that Pearl's definitions support the commonsense judgments that the assassin's putting poison into the coffee caused the bodyguard to put in the antidote, which caused the king to survive, but the assassin's putting poison into the coffee did not cause the king to survive. For example, it can be seen that there is no active causal process from the assassin's pouring in the poison to the king's survival: given that the bodyguard poured in the antidote ($G = 1$), if the assassin had not poured the poison into the coffee, the king would still have survived.

However, Hitchcock points out that when we enlarge the model by adding an extra variable Pearl's definitions deliver a different verdict. Let us consider what happens when we add the variable P into the model, where P refers to a time t shortly after the assassin puts the poison in the coffee but before the bodyguard has had time to react.

$P = 1$ if there is poison in the coffee at time t, 0 if not.

The structural equations now become:

$A = 1$
$P = A$
$G = A$
$S = \sim P \vee G$

The directed graph for this new model is depicted in Figure 5.

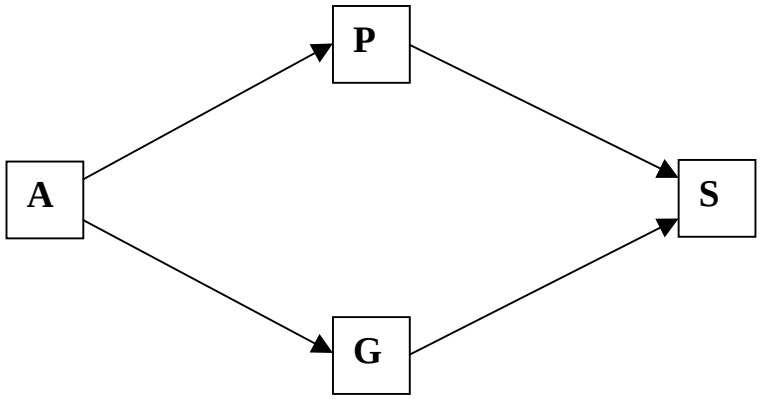

Figure 5

Hitchcock observes that Pearl's theory now yields the result that there is an active causal process from the assassin's pouring in the poison to the king's survival. It is the process <*A*, *G*, *S*>. We can see that this process is active because, holding fixed that fact that the poison is in the cup at time *t* (*P* = 1), the king's survival counterfactually depends on the assassin's pouring in the poison. Thus, given that there was poison in the cup, if the assassin had not poured poison into the cup (and so the bodyguard not responded), the king would not have survived.

Hitchcock (forthcoming a) tries to explain this anomalous result by saying that the three variable model of Example 3 is more natural than the four variable model. The omission of the interpolated variable *P* reflects a feature of the way we think about the example, he writes. We tend to overlook the variable because its inclusion introduces hypothetical possibilities that we consider to be too far-fetched to be relevant to the evaluation of the case. Once we introduce a variable *P*, in addition to *A*, we admit the possibility that these variables might take on values independently of one another. That is, we admit the possibility of entertaining counterfactual antecedents such as 'There is poison in the coffee at *t*, even though *A* did not put it there,' or 'There is no poison in the coffee at *t*, even though *A* poured poison into it'. The problem with these antecedents is that they describe possibilities that are too far-fetched or remote from actuality. On Hitchcock's interpretation, we judge that the assassin's putting the poison in the coffee is not a cause of the king's survival, because the counterfactual that would reveal the active causal process is deemed out of bounds.

These counterexamples to Pearl's theory have involved cases in which the causal verdicts of the theory change when the relevant model is enlarged by the addition of an extra variable. Consequently, it has been possible for the theory's defenders to explain away the anomalous verdicts on the basis of the plausibility or implausibility of adding the extra variable. However, we shall now see that there are counterexamples to Pearl's theory that do not require the interpolation of overlooked variables to make them into counterexamples to the theory. Below I describe two such examples.

> *Example 4: The Alarm System and Generators*
> The power for the alarm system in a building is supplied by a main electricity generator. It is important that the alarm system in the building is always functioning. So, if the main generator cuts out for whatever reason, a back-up generator is activated which supplies electricity to the alarm system. As it happens, the main generator functions normally, the back-up generator is not activated, and the alarm system stays on.

The following model is a natural one to use in reasoning about the causal structure of this example.

> *MG* = 1 if main generator is functioning, 0 if not.
> *BG* = 1 if back-up generator is activated, 0 if not.
> *A* = 1 if alarm system is on, 0 if not.
>
> *MG* = 1

$BG = \sim MG$
$A = MG \vee BG$

The directed graph for this model is depicted in Figure 6.

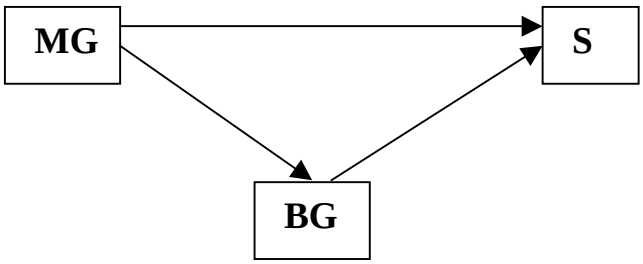

Figure 6

Without interpolating any further variables in this natural model, we can see that there is an active causal process from the functioning of the main generator to the alarm's being on. It is the process $\langle MG, A \rangle$. Given that the back-up generator was not activated, if the main generator hadn't been functioning, the alarm would not have been on. This verdict will strike many as incorrect. The main generator did not cause the alarm to be on because it would have stayed on even if the main generator had been malfunctioning.

Here is another example which presents a similar difficulty for Pearl's theory.

> *Example 5: The Deadly Antidote*
> An assassin puts poison in the king's coffee. The bodyguard responds by pouring an antidote in the king's coffee. If the bodyguard had not put the antidote in the coffee, the king would have died. On the other hand, the antidote is fatal when taken by itself and if the poison had not been poured in first, it would have killed the king. The poison and the antidote are both lethal when taken singly but neutralize each other when taken together. In fact, the king drinks the coffee and survives.

In the natural model for this scenario, the variables are A, G, and S, as described for Example 3. The graph is the same as that depicted in Figure 4. However, the difference in underlying causal mechanisms between Examples 3 and 5 is reflected in the different structural equations for this example:

$A = 1$
$G = A$
$S = (A \ \& \ G) \vee (\sim A \ \& \sim G)$

Testing for active causal processes, we see that the process from the assassin's pouring the poison in the coffee to the king's survival is active. Holding fixed the fact that the bodyguard poured the lethal antidote into the coffee, we note that the king would not have survived if the assassin had not put the poison in the coffee first. However, most people

judge that the assassin's action did not cause the king's survival, since the king would have survived even if the assassin had not performed this act.

Observe that once more this counterexample does not rely on the addition of any extra variable to the natural model suggested by the example. Hence, the defenses advanced by Halpern and Pearl, on the one hand, and by Hitchcock, on the other, are beside the point. In conclusion to this part of our argument, I suggest that we need to consider an alternative way of formulating Pearl's theory if it is to have some chance of successfully dealing with these examples.

4. DEFAULT WORLDS, CONDITIONALS, AND MODELS

By way of introducing an alternative formulation, let us consider more generally how counterfactuals are traditionally understood by philosophers and how they are understood in the structural equations framework. Traditionally, philosophers have developed semantics for counterfactuals in terms of similarity relations between possible worlds. One classic treatment is David Lewis's (1973) possible world semantics. (Pearl in his (2000, section 7.4) shows that the axioms of Lewis's theory follow from the axioms of his own structural semantics.) One central feature of Lewis's semantics is that it uses a system of nested spheres of possible worlds centered on the actual world. A sphere represents a set of possible worlds that are equally similar to the actual world, the smaller the sphere the more similar to the actual world are the possible worlds within it.

Built into this semantics is the Centering Principle to the effect that there is no world as similar to the actual world as the actual world itself. In terms of this system of spheres the truth condition for a counterfactual is stated as follows: $P \mathrel{\square\!\!\rightarrow} Q$ is true iff Q is true in every P-world in the smallest P-permitting sphere. It follows from the Centering Principle that $P \mathrel{\square\!\!\rightarrow} Q$ is true if P and Q are true.

In my paper (2003b) I propose a modified semantics for the kinds of conditionals that are relevant to token-causation. The approach differs from Lewis's in two ways. The first difference is that the similarity relation for the causally relevant conditionals is specified by reference to a contextually salient causal model. Such a causal model determines the relevant respects of similarity to be considered in evaluating a given conditional. A causal model within this semantics is understood in much the same way as it is understood in Pearl's framework. The second difference from Lewis's semantics is that system of spheres of possible worlds is centered, not on the actual world, but on a set of what I call default worlds. Adapted to the present framework, the default worlds generated by a causal model of an actual system are characterized as follows:

> *Definition 5: The Default Worlds Generated by a Causal Model*
> A causal model $<U, V, E>$ of an actual system generates a sphere of default worlds that consists of all and only worlds w such that:
> (i) w contains a counterpart system of the same kind whose exogenous variables in U are set at their default values;
> (ii) w evolves in accordance with the structural equations in E without any further intervention.

The intuitive idea is that the default worlds generated by the causal model exemplify a course of evolution that is normal, in a certain sense, for a system of the given kind. More particularly, they represent the way that system of the given kind would evolve from its default initial state without intervention or interference from outside the system. Of course, the crucial notion here is that of the default settings of the exogenous variables of a model. Roughly speaking, these represent the normal state of the system at the beginning of its evolution. It is difficult to specify this notion more precisely. For the way we select the default settings of the exogenous variables is affected by a range of considerations from the kind of system under investigation to the nature and the purpose of the investigation. Indeed, in many kinds of ordinary causal judgments the factors that influence our judgments about the appropriate default settings may include intersubjective background expectations and even normative judgments about what should happen in scenarios of the kind in question. I shall rely on examples to make clear how this notion is to be understood.

Let us consider how the notions of Definition 5 would apply to the scenario described in Example 1. What would a default world generated by the salient causal model for this scenario look like? As we have seen, the exogenous variables in the causal model are *ST* and *BT*. I suggest that it would be natural to set the default values of these variables at 0 to represent the state of affairs in which neither Suzy nor Billy throw a rock: in some sense, this represents the normal state of affairs in this scenario. A default world would evolve from this state of affairs in accordance with the structural equations in *E* so that the bottle does not shatter ($BS = 0$). This is also assuming that the world involves no intervention that 'freezes' any of the endogenous variables. So, in short, a default world would be one in which neither Suzy nor Billy throws rocks, no rock hits the bottle, and the bottle does not shatter.

The sphere of default worlds generated by a model is tied, in some sense, to the actual world. For worlds earn their membership in the sphere by virtue of their resemblance to the way the actual system under consideration would evolve in conformity with the structural equations. Nonetheless, it is important to note that the actual world need not itself belong to the sphere of default worlds. For these worlds represent how a normal system would evolve in conformity with the structural equations in the absence of outside intervention. In many cases, therefore, these worlds are ideal ones. The actual world, as we know, may be very far from ideal in that the actual system may not be normal and its course of evolution may be affected by external interferences. We see this illustrated by Example 1. The actual world is one in which both Suzy and Billy threw rocks and the bottle shattered, while the default worlds are ones in which neither Suzy nor Billy threw rocks and the bottle did not shatter. The sphere of default worlds within this framework includes the worlds that count as the closest worlds to the actual world. The fact that the actual world need not belong to this sphere means that Lewis's Centering Principle fails in this framework. This has some surprising implications for the logic of conditionals, including the failure of the Modus Ponens, but these must be explored elsewhere.

So far we have attended to the question of which worlds count as the default worlds generated by a causal model. But we need to provide truth-conditions for all causally

relevant conditionals, both those with factual and those with counterfactual antecedents; and so we need to specify which will be the closest-antecedent worlds for any conditional. In some cases, the antecedent of the conditional will overlap with the sphere of default worlds, and so the closest antecedent-worlds are simply specified as those antecedent-worlds that belong to this overlap. In other cases, however, the antecedent of the conditional will not overlap with the sphere of default worlds and the closest antecedent-worlds must be specified in some non-obvious way. In my paper (2003b) I explored one way in which a causal model might create an ordering of spheres of possible worlds. Here, however, I adopt an idea of Pearl's (2001, p.241) and propose the following method for ordering spheres of worlds in terms of a causal model.

> *Definition 6: Ordering of Spheres by a Causal Model*
> $\{S_0,..., S_n\}$ is a system of spheres ordered by the model $<U, V, E>$ iff S_0 is the sphere of default worlds generated by the model and S_i is a sphere of worlds such that a default world in S_0 is transformed into a world in S_i by a maximum number of i interventions in the structural equations of the model.

It is easy to see that this method of ordering the spheres of possible worlds ensures that they are centred on the sphere of default worlds and are nested within each other.

On the basis of this ordering of the spheres, the truth-conditions for causally relevant conditionals can be formulated in general terms as follows:

> *Definition 7: Truth-Conditions of Causally Relevant Conditionals*
> $P \boxdot\!\!\rightarrow_M Q$ is true in the actual world relative to a system of spheres ordered by a causal model M iff Q is true in all P-worlds in the smallest P-permitting sphere of the system.

On the basis of these truth-conditions, I suggest the following definition of conditional dependence.

> *Definition 8: Conditional Dependence Relative to a Model*
> Q conditionally depends on P relative to a causal model M of an actual scenario iff $P \boxdot\!\!\rightarrow_M Q$ and $\sim P \boxdot\!\!\rightarrow_M \sim Q$.

In my (2003b) paper I point out that this definition of conditional dependence covers two importantly different types of cases. It follows from the assumption of determinism that we are making that, for any given pair of conditionals of the form $P \boxdot\!\!\rightarrow_M Q$ and $\sim P \boxdot\!\!\rightarrow_M \sim Q$, one antecedent will overlap with the sphere of default worlds and the other antecedent will not. Let us call a conditional a 'default conditional' if its antecedent overlaps with the sphere of default worlds so that we do not have leave this sphere to find the closest antecedent-worlds. Let us call a conditional an 'intervention conditional' if its antecedent does not overlap with the sphere of default worlds so that we have to move to a sphere of worlds permitting interventions to find the closest antecedent-worlds. It follows from these definitions that one of the pair of $P \boxdot\!\!\rightarrow_M Q$ and $\sim P \boxdot\!\!\rightarrow_M \sim Q$ will be a default conditional and the other will be an intervention conditional. It is important to recognise that the distinction between these conditionals does not always line up with the distinction

between factual conditionals with true antecedents and counterfactual conditionals with false antecedents. Indeed, the two types of cases covered by the definition of conditional dependence correspond to two different possible ways in which the distinction could line up with the factual/counterfactual distinction between conditionals. Table 1 below represents the two possible types of conditional dependence.

Types of conditional dependence	*Conditional with true antecedent (factual)*	*Conditional with false antecedent (counterfactual)*
Type I:	Default conditional	Intervention conditional
Type II:	Intervention conditional	Default conditional

Table 1

There has been some discussion in the literature (see especially Hall (2003)) that there are in fact two concepts that influence our judgments about token-causation. One is a concept of causation as dependence and the other a concept of causation as production. In a discussion of this issue, Pearl (2000, p.316) makes a suggestion that in effect amounts to identifying causation as dependence with the first type of conditional dependence and causation as production with the second type. I doubt whether Pearl's interesting conjecture can account for all the different features that have become associated in the literature with these two conceptions. Nonetheless, the significant point for our discussion is that practically all counterfactual theories of causation have focused on the first type of conditional dependence. Even Pearl, who recognises the formal possibility of the second type, does not extensively investigate it on the grounds that the existence of a conditional dependence under the hypothesis that the exogenous variables have non-actual values could have no bearing on what happens in the actual world. We will need to address this important concern later. In the meantime, I simply observe that most philosophers do not even recognise the possibility of the second type of conditional dependence because they assume that the system of spheres of possible worlds must conform to the Centering Principle so that any conditional with true antecedent and consequent is trivially true. However, the present semantics for causally relevant conditionals rejects the Centering Principle, leaving open the possibility that a conditional may be false even when its antecedent and consequent are true. This in turn creates the opportunity for developing nontrivial truth-conditions for both factual and counterfactual conditionals, and so for distinguishing two ways in which a pair of conditionals of the form $P \boxdot\!\!\rightarrow_M Q$ and $\sim P \boxdot\!\!\rightarrow_M \sim Q$ might be true.

In the remainder of this paper we shall focus on the second type of conditional dependence, if only to counter the overwhelming preoccupation with the first type. This focus should not be construed as denial of the existence of the first type. The interactions between the two types of conditional dependence are subtle and complex, and a study of them will have to await another occasion. However, before continuing, let us pause to consider a passage from Hart and Honoré's classic discussion of the commonsense causal judgments, a discussion that emphasises the centrality of the second type of conditional dependence. They write:

> Human action in the simple cases, where we produce some desired effect by the manipulation of an object in our environment, is an interference in the natural course of events which *makes a difference* in the way these develop. In an almost literal sense, such an interference by human action is an intervention or intrusion of one kind of thing upon a distinct kind of thing. Common experience teaches us that, left to themselves, the things we manipulate, since they have a 'nature' or characteristic way of behaving, would persist in states or exhibit changes different from those we have learnt to bring about in them by our manipulation. The notion that a cause is essentially something which interferes with or intervenes in the course of events which would normally take place, is central to the commonsense concept of cause...Analogies with the interference by human beings with the natural course of events in part control, even in cases where there is literally no human intervention, what is identified as the cause of some occurrence; the cause, though not a literal intervention, is a difference to the normal course which accounts for the difference in the outcome. (Hart and Honoré 1985, p.29)

It is clear that the kind of case that Hart and Honoré are discussing is that in which a cause makes a difference to its effect because it is, or is analogous to, an intervention in a system which disturbs it from its normal course of evolution. This is precisely the kind of the connection between cause and effect that belongs to the second type of conditional dependence.

How is all of this relevant to modifying Pearl's framework so as to improve its fit with intuitive judgments? It is a simple matter to modify Pearl's framework so as to accommodate both types of conditional dependence described above. As we have seen above, the essential feature in specifying the default worlds is to fix the default setting of the exogenous variables and to plug these values into the structural equations without allowing for any further intervention. The default worlds will be simply the worlds which evolve from these initial settings in accordance with the structural equations without further intervention. Now I suggest that the obvious extension of this treatment is to relativize Pearl's definition of counterfactual dependence to a causal model which fixes the values of the exogenous variables at their default settings. Let us introduce the notion of a default causal model as follows:

> *Definition 9: A Causal Model with Default Settings of its Exogenous Variables (a Default Model for short)*
> A default causal model $<U, V, E>$ of an actual system is one in which U is a set of exogenous variables which are set at their default values, which may not coincide with their actual values, and V and E are the same as before.

The notion of counterfactual dependence is the same as before, except that it is now called conditional dependence to signify the nontrivial role of factual conditionals and it is now relativized to a special kind of causal model, a default model. If a conditional dependence holds between binary variables X and Y in a default causal model M in which $X = 1$ and $Y = 1$ represent the positive occurrence of events, then the definition implies that the following pair of conditionals holds:

$X = 1 \; \square\!\!\rightarrow_M \; Y = 1$ and $X = 0 \; \square\!\!\rightarrow_M \; Y = 0$

As we have seen above, the truth of these conditionals is compatible both with the possibility that the first is a default conditional and the second an intervention conditional and with the possibility the first is an intervention conditional and the second a default conditional.

For the sake of concreteness, let us suppose that the pair of conditionals of this kind holds relative to a default causal model *M*, and that the pair are an intervention and a default conditional respectively. For example, let us reconsider Example 1 in which the default model fixes the values of the exogenous variables $ST = 0$ and $BT = 0$ and in which there is a conditional dependence between *ST* and *BS*:

$ST = 1 \; \square\!\!\rightarrow_M \; BS = 1$ and $ST = 0 \; \square\!\!\rightarrow_M \; BS = 0$

(Note that here the first conditional is a factual intervention conditional and the second a counterfactual default conditional.) Now let us return to Pearl's sceptical question about the relevance of this conditional dependence to the existence a causal relation in the actual situation between Suzy's throwing a rock and the bottle's shattering. How does the fact that this conditional dependence obtains in the hypothetical scenario in which *ST* and *BT* have the value 0 have any bearing on the actual scenario in which these variables actually have the value 1? In other words, why should the fact that the bottle's shattering would depend conditionally on Suzy's throwing a rock relative to the hypothetical situation in which neither Billy nor Suzy threw rocks have any relevance to the causal structure of the actual situation in which both actually threw rocks? The answer to this question is that when a conditional dependence of this type holds in such a hypothetical situation, it picks out an intrinsic process of a certain kind that is also present in the actual situation. For example, in relation to Example 1, consider what would be true of the closest worlds to the default worlds in which Suzy throws a rock. There will be an intrinsic process holding in this world that does not hold in any of the default worlds: the process consisting of Suzy's throwing a rock, the rock's hitting the bottle, and the bottle's shattering. While there is no conditional dependence in the actual world between Suzy's throwing a rock and the bottle shattering thanks to the presence of Billy's rock-throwing, nonetheless this same intrinsic process holds here too. So there is a feature of the actual world that grounds a causal judgment—the existence of an intrinsic process—but this process is identified as the relevant truth-maker by its occupying a certain functional role defined in terms of a conditional dependence.

We can specify the conditionally defined functional role that is occupied by the intrinsic process with some precision within the structural equations framework:

> *Definition 10: An Intrinsic Process Picked out by a Conditional Dependence (of Type II relative to a Default Model M)*
> Suppose that there is a conditional dependence of type II between *X* and *Y* relative to the default model *M*, where $X = x$ and $Y = y$ represent actually occurring events. Let $\langle X, Z_1, \ldots, Z_n, Y \rangle$ be a process connecting *X* to *Y*. The sequence of positive states $\langle X$

> $= x, Z_1 = z_1,..., Z_n = z_n, Y= y>$ is the intrinsic process picked out by this conditional dependence iff feeding X = x into the structural equations of the model yields the solutions $Z_1 = z_1,..., Z_n = z_n, Y= y$ while feeding in the default value assignment to X yields a different values for each of $Z_1, ..., Z_n, Y$.

Now we come to define the concept of causation:

> *Definition 11: (Productive) Cause*
> $X = x$ is a (productive) cause of the $Y = y$ relative to the default model M iff a conditional dependence (of Type II) holds between X and Y relative to M and an intrinsic process picked out by this conditional dependence links $X = x$ with $Y = y$ in the actual scenario.

I believe that this theory of causation is a more faithful rendering of Lewis's notion of quasi-dependence within Pearl's framework than is Pearl's own suggested rendering. Elsewhere (1996, 2003a) I have called this theory a functionalist theory of token-causation because it defines causal relations in terms of intrinsic processes that occupy certain conditionally defined functional roles.

In terms of this theory, we can explain why Suzy's rock-throwing, but not Billy's, is a productive cause of the bottle shattering. The natural default model of the scenario sets the default values of ST and BT at 0. In the default worlds generated by this model there is a conditional dependence of Type II between Suzy's throwing a rock and the bottle's shattering and furthermore the intrinsic process picked out by this counterfactual dependence—viz $<ST = 1, SH = 1, BS = 1>$ —holds in the actual situation. Contrast the situation with Billy's throwing a rock. Notice that in the default worlds generated by this model there is a conditional dependence of type II between Billy's throwing a rock and the bottle's shattering and furthermore this counterfactual dependence picks out an intrinsic process—viz, $<BT = 1, BH = 1, BS = 1>$. The crucial difference between Suzy's and Billy's actions, however, is that this second intrinsic process is not realised in the actual world.

5. THE PROBLEM CASES RECONSIDERED

Let us return to the examples which proved to be problematic for Pearl's theory and reconsider them in the light of the proposed theory of token-causation.

Let us start with Example 2: The Rail Track Switching Example. In this example our natural causal intuitions are that flipping the switch caused the train to go down the left track, which in turn caused the train to arrive at the point of reconvergence of the tracks, but the first event did not cause the last. As we have seen, Halpern and Pearl argue that these causal intuitions are supported by a three variable, but not by a four variable causal model of the scenario. Let us concentrate on their four variable model, where the divergence from commonsense intuition emerges sharply.

It is natural to interpret Example 2 in terms of a four variable default causal model that sets the value of the exogenous variable F at 0 so that the default worlds are ones in which

the switch is not flipped, the train goes down the right track, and arrives at the point of reconvergence. The proposed theory tells us that if we are considering this in terms of a type II conditional dependence, then we must consider whether the following pairs of conditionals hold corresponding to each of the three causal judgments in question:

$F = 1 \; \boxdot\!\!\rightarrow_M \; LT = 1$ and $F = 0 \; \boxdot\!\!\rightarrow_M \; LT = 0$
$LT = 1 \; \boxdot\!\!\rightarrow_M \; A = 1$ and $LT = 0 \; \boxdot\!\!\rightarrow_M \; A = 0$
$F = 1 \; \boxdot\!\!\rightarrow_M \; A = 1$ and $F = 0 \; \boxdot\!\!\rightarrow_M \; A = 0$

It is easy to check that the first two pairs of conditionals hold and that these conditional dependences (trivially) pick out intrinsic processes that hold in the actual situation, so vindicating the causal judgments that flipping the switch caused the train to go down the left track and that the train's going down the left track caused its arrival. However, the fact that the third pair fails to hold means there is no conditional dependence between flipping the switch and the train's arrival, so that there is no causal relation between them. This is in accord with the intuitive causal judgment and so marks a difference between the present theory and Pearl's theory. It is also easy to check that this causal judgment is also delivered by the three variable version of the default model as well. Therefore, it does not matter to the present theory whether a three or four variable model is used, the causal verdict delivered is the same in both cases.

Let us now turn to consider Example 3: The Assassin and the Bodyguard. Hitchcock argues that in this case Pearl's theory delivers different results depending on whether a three or four variable model is employed. The three variable model yields the correct result that the assassin's putting the poison in the king's coffee was not a cause of the king's survival, whereas the four variable model yields the opposite result. Hitchcock tries to explain away this anomalous result by explaining why the four variable model is not an appropriate one to use in this case. Let us, therefore, concentrate on the four variable model as it is the more contentious one. Let us also assume that it is appropriate to interpret the example with a four variable model that sets the default value of the exogenous variable A at 0. In the default worlds generated by this model, the assassin does not put poison in the coffee, the bodyguard does not put the harmless antidote into the coffee, and the king survives.

The intuitive causal judgments in this example are that the assassin's action caused the bodyguard's action, which caused the king to survive, but the assassin's action did not cause the king's survival. Now the proposed theory tells us that if we are considering this in terms of a type II conditional dependence, we must consider whether the following pairs of conditionals hold corresponding to each of the three causal judgments in question:

$A = 1 \; \boxdot\!\!\rightarrow_M \; G = 1$ and $A = 0 \; \boxdot\!\!\rightarrow_M \; G = 0$
$G = 1 \; \boxdot\!\!\rightarrow_M \; S = 1$ and $G = 0 \; \boxdot\!\!\rightarrow_M \; S = 0$
$A = 1 \; \boxdot\!\!\rightarrow_M \; S = 1$ and $A = 0 \; \boxdot\!\!\rightarrow_M \; S = 0$

It is easy to check that when the variable A is given the default setting of 0, then the first two pairs of counterfactual dependences hold and they trivially pick out intrinsic processes that hold in the actual situation, so vindicating the judgments that the assassin's

action caused the bodyguard's action and that the bodyguard's action caused the king's survival. But the failure of the last conditional dependence shows that the assassin's action did not cause the king's survival. The interpolated variable *P* does not figure in these counterfactuals, but the way the structural equations determine its values does not materially affect the truth-values of these counterfactuals. In any case, it can be shown that applying the present theory to the three variable model does not change any the causal verdicts of the theory. This suggests that the introduction of this fourth variable is immaterial to our causal judgments, whatever implications it may have for Pearl's theory.

Let us turn to consider Example 4: the Alarm and the Generators. As we saw, Pearl's theory implies that functioning of the main generator is the cause of the alarm's being on. I assume, somewhat controversially I grant, that this is contrary to commonsense judgment. Nonetheless, let us see how this example is interpreted in terms of the present theory. The natural setting of the default value for the exogenous variable *MG* is 1 so that the default worlds are ones which make it true that the main generator is functioning, the backup generator is not activated and the alarm system is on. Now we need to test the claim that the functioning of the main generator was or was not a cause of the alarm's being on. When the default conditions are set up in this way, the relevant conditional dependence to be evaluated is of type I, so this example is different from others considered above:

$MG = 1 \; \square\!\!\rightarrow_{M} A = 1$ and $MG = 0 \; \square\!\!\rightarrow_{M} A = 0$

It is apparent that this conditional dependence fails, as the alarm would be on whether or not the main generator was functioning. Accordingly, the functioning of the main generator did not cause the alarm to be on.

I must admit that this verdict is slightly puzzling. One might be inclined to say that there must be some cause of the alarm's being on and the only possible answer is that the cause is the functioning of the main generator that supplies the electricity to the alarm system. I agree that this view has some plausibility. But it is a virtue of the present kind of framework that it can supply an explanation of its plausibility. So far we have supposed that the appropriate model is one which includes a variable indication of the activation or inactivation of the back-up generator. However, since this generator is rarely activated and certainly not on the occasion in question, it is appealing to abstract away from the presence of this generator to focus on the system that includes just the circuit from the main generator to the alarm system. This is clearly a system that is capable of functioning independently of the back-up generator. An appropriate model for this system might contain the following variables

MG = 1 if the main generator is functioning, 0 if not.
E = 1 if electricity is present in the circuit from main generator to alarm, 0 if not.
A = 1 if the alarm is on, 0 if not.

The appropriate structural equations, with the default setting of the exogenous variable, are:

$MG = 1$
$E = MG$
$A = E$

It is easy to show that in this model there is a conditional dependence of Type I between the functioning of the main generator and the alarm's being on and that this conditional dependence picks out an appropriate intrinsic process that holds in the actual situation.

Finally, let us consider Example 5: The Deadly Antidote. This example appears to be a straightforward counterexample to Pearl's theory. For in this example we judge that the assassin's pouring the poison into the king's coffee did not cause the king to survive, but Pearl's theory has the opposite implication. Remember that in this example the antidote and the poison are both lethal when taken by themselves but neutralize each other when taken in combination. The intuitive causal judgments about this example are the same as those in Example 3. If we take the default worlds again to be those worlds in which the assassin does not put poison in the coffee, the bodyguard does not put antidote into the coffee, and the king survives, we can see that the conditional dependences corresponding to these causal judgments hold or fail to hold appropriately. Whether the antidote is lethal or harmless does not seem to affect our causal judgments about the particular occasion, though it affects the implications of Pearl's theory.

On the basis of these explanations, there are grounds for optimism that the present theory is on the right track, though this can only be confirmed by more extensive investigations. In the meantime, I simply claim that it presents a much more plausible way of elaborating Pearl's own framework in application to the case of token-causation.

REFERENCES

Hall, N. 2003. "Two Conceptions of Causation" in J. Collins, N. Hall. L. Paul (eds.), *Counterfactuals and Causation*. Cambridge, Mass.: MIT Press.

Halpern, J. and Pearl, J. 2001. *Causes and Explanations: A Structural Model Approach—Part I: Causes*, Technical Report R-266, Cognitive Systems Laboratory, University of California, Los Angeles.

Hart, H. and Honoré, A. 1985. *Causation in the Law*: 2nd edition. Oxford: Clarendon Press.

Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy*, 98, 273-314.

Hitchcock, C. Forthcoming a. "Token Causation and the Principle of Sufficient Reason".

Hitchcock, C. Forthcoming b. "Active Routes and Token Causation".

Lewis, D. 1973. *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, D. 1979. "Counterfactual Dependence and Time's Arrow", *Nous*, 13, pp.455-76. Reprinted in Lewis 1986.

Lewis, D. 1986. *Philosophical Papers: Volume II*. Oxford: Oxford University Press.

Menzies, P. 1996. "Probabilistic Causation and the Pre-emption Problem", *Mind*, 105, 85-117.

Menzies, P. 2003a. "The Causal Efficacy of Mental States", in S. Walter (ed.) *Current Issues in Mental Causation*, Mentis.

Menzies, P. 2003b. "Difference-Making in Context", in J. Collins, N. Hall. L. Paul (eds.), *Counterfactuals and Causation*. Cambridge, Mass.: MIT Press.

Pearl, J. 2000. *Causality*. Cambridge: Cambridge University Press.