Causal Models, Token Causation, and Processes

Peter Menzies†

Judea Pearl (2000) has recently advanced a theory of token causation using his structural equations approach. This paper examines some counterexamples to Pearl's theory, and argues that the theory can be modified in a natural way to overcome them.

1. Introduction. It is a regrettable fact that the structural equations approach to the study of causation is not well 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 preoccupied 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 has developed a theory that applies the structural equations approach to token causation (2000). Subsequently, in collaboration with David Halpern, he has extended and simplified the theory (2001). 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, 2002a, 2002b). 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.

2. Pearl's Theory of Token Causation. Pearl's theory of token causation relativizes token-causal judgements to a causal model. A causal model is an ordered triple $\langle U, V, E\rangle$, 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;

†To contact the author, please write to: Department of Philosophy, Macquarie University, North Ryde NSW 2109, Australia; e-mail: peter.menzies@mq.edu.au.

Philosophy of Science, 71 (December 2004) pp. 820–832. 0031-8248/2004/7105-0014$10.00
Copyright 2004 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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 preemption.

> *Example 1: Billy and Suzy*. Billy and Suzy throw rocks at a bottle. Suzy's rock gets there first, shattering the bottle. Billy's rock 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 five 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. The structure of causal mechanisms in Example 1 can be captured using the following three structural equations:

$$SH = ST$$
$$BH = BT \,\&\, \sim SH$$
$$BS = SH \vee BH$$

A structural equation can be thought of as encoding a battery of nonbacktracking counterfactuals. The convention is that the variables appearing on the right-hand side of an equation figure in the antecedents

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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. For example, 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.

More generally, it is possible to derive a counterfactual from a set of structural equations even if the counterfactual is not directly encoded by them. 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. Instead of this variable having its value causally determined in the normal way, it is 'miraculously' set to its new hypothetical value. 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.

The technique of replacing a structural equation with a hypothetical value 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 $Y = y'$.

Pearl does not attempt to define token causation in terms of counterfactual dependence. For examples of late preemption such as Example 1 show that causation is not coextensive with counterfactual dependence. Suzy's throwing a rock caused the bottle to shatter, but 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*.

Rather than linking causation with counterfactual dependence directly, Pearl tries to capture within the structural equations framework the notion of *quasi dependence* that Lewis introduced as a tentative solution—though later discarded—to the difficulties that late preemption examples posed his original counterfactual theory (1986, 206). Pearl informally explains the notion of quasi dependence in terms of a test for token causation: Look for an intrinsic process connecting putative cause with effect, sup-

https://doi.org/10.1086/425057 Published online by Cambridge University Press

press the influence of their nonintrinsic surroundings, and subject the cause to a counterfactual test to see whether changing the putative cause changes the effect. More formally, he tries to capture the intuitive force of this test in the following series of definitions. (They are 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 $\langle U, V, E\rangle$ is an ordered sequence of variables $\langle X, Z_1, \ldots, Z_n, Y\rangle$ such that each variable in the sequence is in $U \cup V$ and appears on the right hand side of the structural equation for the variable that is its successor in the sequence.
>
> **Definition 3: An Active Causal Process**. The process $\langle X, Z_1, \ldots, Z_n, Y\rangle$ is an active causal process relative to the model $\langle U, V, E\rangle$ iff there exists a (possibly empty) set of variables $\{W_1, \ldots, W_m\}$ in $U \cup V \backslash \langle X, Z_1, \ldots, Z_n, Y\rangle$ with actual values $w_1, \ldots, w_m$ 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.
>
> **Definition 4: Actual Causation**. $X = x$ is an actual cause of $Y = y$ relative to the causal model $\langle U, V, E\rangle$ iff there is an active causal process from X to Y.

These definitions are intended to capture the notions mentioned in the informal test described above. For example, the values of the variables in the sequence $\langle X, Z_1, \ldots, Z_n, Y\rangle$ represent the intrinsic process connecting cause and effect. The set of variables $\{W_i\}$ represents 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 $\{W_i\}$ at their actual values, 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.

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 $\langle ST, SH, BS\rangle$. 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 coun-

https://doi.org/10.1086/425057 Published online by Cambridge University Press

terfactually depends on *ST* in this new set of structural equations. But notice that this is not true for the process $\langle BT, BH, BS\rangle$ 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 Christopher Hitchcock. He argues that the theory can be defended against this kind of counterexample (2001, 2002a, 2002b). I am much less sanguine about the prospects of the theory in its present form.

One of Hitchcock's examples (2002a) involves the following situation.

> *Example 2: 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 antidote into the coffee, 0 if not.
$S = 1$ if the king survives, 0 if not.

$$A = 1$$
$$G = A$$
$$S = {\sim}A \vee G$$

It is easy to check that Pearl's definitions support the commonsense judgements 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.

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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:

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

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 $\langle A, G, S\rangle$. We can see that this process is active because, holding fixed the 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 (2002a) tries to explain this anomalous result by saying that the three-variable model 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. We tend, he writes, 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. However, the following counterexample does not require the interpolation of any overlooked variable that introduces far-fetched possibilities.

> *Example 3: 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 2. However, the difference in underlying causal mechanisms between Examples 2 and 3 is reflected in the different structural equations for this example:

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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

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.

In conclusion to this part of the argument, I suggest 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, Counterfactual Dependence, and Intrinsic Processes. As a preliminary to introducing an alternative formulation, we need to reconsider some fundamental issues about the semantics of counterfactuals. A standard starting point for philosophical discussions of counterfactuals is David Lewis's possible world semantics (1973). Lewis's semantics 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: $P \mathbin{\square\!\!\rightarrow} Q$ is true iff Q is true in every P-world in the smallest P-permitting sphere.

In my paper (2003) I propose a modified semantics for the kinds of counterfactuals that are relevant to token causation. The semantics 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. The second difference from Lewis's semantics is that the 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 $\langle U, V, E \rangle$ of an actual system generates a sphere of default worlds that consists of all and only worlds w such that:

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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 a system of the given kind would evolve from its default initial state without 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 vast number of considerations ranging from the kind of system under investigation to the nature and the purpose of the investigation. I shall rely on examples to make clear how this notion is to be understood.

Consider, for example, 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

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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.

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 counterfactuals and to specify which will be the closest antecedent-worlds for all such counterfactuals. In some cases, the antecedent of the counterfactual 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 counterfactual will not overlap with the sphere of default worlds and the closest antecedent-worlds must be specified in some nonobvious way. Pearl makes a natural proposal for ordering spheres of worlds in terms of a causal model (2000, 241) and I shall adapt his proposal to the present framework as follows.

> **Definition 6: Ordering of Spheres by a Causal Model**. $\{S_0, \ldots, S_n\}$ is a system of spheres ordered by the model $\langle U, V, E\rangle$ 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 centered on the sphere of default worlds and are nested within each other.

How is all of this relevant to modifying Pearl's framework so as to improve its fit with intuitive judgements? Its relevance lies in the fact that the above changes in the semantics of counterfactuals suggest a way of redefining the central notion of counterfactual dependence that Pearl uses in his test for token causation. Clearly, it is a simple matter to modify Pearl's definition of counterfactual dependence by relativizing the truth conditions of counterfactuals to the sphere of worlds ordered by a causal model with its exogenous variables fixed at their default settings, and to redefine counterfactual dependence in terms of such counterfactuals.

> **Definition 7: Truth Conditions for Causally Relevant Counterfactuals**. $P \mathrel{\Box\!\!\rightarrow_M} Q$ is true in the actual world relative to the causal model M iff Q is true in all P-worlds in the smallest P-permitting sphere in

https://doi.org/10.1086/425057 Published online by Cambridge University Press

the system of spheres ordered by the model M with its exogenous variables fixed at their default settings.

Definition 8: Counterfactual Dependence Relative to a Model. Q counterfactually depends on P relative to a causal model M iff $P \mathbin{\Box\!\!\rightarrow_M} Q$ and $\sim P \mathbin{\Box\!\!\rightarrow_M} \sim Q$.

By way of illustrating these definitions, consider again Example 1. In the natural model of this example, the default settings of the exogenous variables ST and BT are 0, and this model with these settings generates an ordering of worlds relative to which the following counterfactual dependences hold:

$$ST = 1 \mathbin{\Box\!\!\rightarrow_M} BS = 1 \text{ and } ST = 0 \mathbin{\Box\!\!\rightarrow_M} BS = 0;$$
$$BT = 1 \mathbin{\Box\!\!\rightarrow_M} BS = 1 \text{ and } BT = 0 \mathbin{\Box\!\!\rightarrow_M} BS = 0.$$

Nonetheless, it is reasonable to ask, as Pearl does, about the relevance of these counterfactual dependences to the existence of causal relations in the actual situation (2000, 316). For instance, why does the counterfactual dependence between Suzy's rock throwing and the bottle's shattering that obtains in the hypothetical scenario in which ST and BT have the value 0 have any bearing on causal relations in the actual scenario in which these variables actually have the value 1? I suggest that the answer to this question is that when a counterfactual 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 counterfactual 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. There is a feature of the actual world that grounds a causal judgement—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 counterfactual dependence.

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

Definition 9: An Intrinsic Process Picked Out by a Counterfactual Dependence. Suppose that there is a counterfactual dependence between X and Y relative to the model M, where $X = x$ and $Y = y$

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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

At last we can define the concept of causation as follows.

Definition 10: Causation. $X = x$ is a cause of $Y = y$ relative to the model M iff a counterfactual dependence holds between X and Y relative to M and an intrinsic process picked out by this counterfactual 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 I have called this theory a functionalist theory of token causation because it defines causal relations in terms of intrinsic processes that occupy certain counterfactually defined functional roles (1996).

In terms of this theory, we can explain why Suzy's rock throwing, but not Billy's, is a cause of the bottle shattering. The natural 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 counterfactual dependence between Suzy's throwing a rock and the bottle's shattering, and furthermore the intrinsic process picked out by this counterfactual dependence—viz., $\langle ST = 1, SH = 1, BS = 1\rangle$—holds in the actual situation. Contrast the situation with Billy's throwing a rock. In the default worlds generated by this model there is also a counterfactual dependence between Billy's throwing a rock and the bottle's shattering and furthermore this counterfactual dependence picks out an intrinsic process—viz., $\langle BT = 1, BH = 1, BS = 1\rangle$. The crucial difference between Suzy's and Billy's actions, however, is that this second intrinsic process is not realized 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 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

https://doi.org/10.1086/425057 Published online by Cambridge University Press

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, there is no poison in the cup at time t, the bodyguard does not put the harmless antidote into the coffee, and the king survives.

The intuitive causal judgements 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. The proposed theory tells us that we must consider whether the following pairs of counterfactuals hold corresponding to each of the three causal judgements in question:

$$A = 1 \,\Box\!\!\rightarrow_M G = 1 \text{ and } A = 0 \,\Box\!\!\rightarrow_M G = 0$$
$$G = 1 \,\Box\!\!\rightarrow_M S = 1 \text{ and } G = 0 \,\Box\!\!\rightarrow_M S = 0$$
$$A = 1 \,\Box\!\!\rightarrow_M S = 1 \text{ and } A = 0 \,\Box\!\!\rightarrow_M S = 0$$

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

Finally, let us turn to consider Example 3: 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 judgements about this example are the same as those in Example 2. 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 an antidote into the coffee, and the king survives, we can see that the counterfactual dependences corresponding to these causal judgements hold or fail to hold appropriately. Whether the antidote is lethal or harmless does not seem to affect our

https://doi.org/10.1086/425057 Published online by Cambridge University Press

causal judgements 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

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

Hitchcock, Christopher (2001), "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy* 98: 273–314.

——— (2002a), "Token Causation and the Principle of Sufficient Reason", unpublished manuscript.

——— (2002b), "Active Routes and Token Causation", unpublished manuscript.

Lewis, David (1973), *Counterfactuals*. Oxford: Basil Blackwell.

——— (1986), *Philosophical Papers*, vol. 2. Oxford: Oxford University Press.

Menzies, Peter (1996), "Probabilistic Causation and the Pre-emption Problem", *Mind* 105: 85–117.

——— (2003), "Difference-Making in Context", in John Collins, Ned Hall, and Laurie Paul (eds.), *Counterfactuals and Causation*. Cambridge, MA: MIT Press.

Pearl, Judea (2000), *Causality*. Cambridge: Cambridge University Press.

https://doi.org/10.1086/425057 Published online by Cambridge University Press