L&PS – Logic & Philosophy of Science
Vol. VII, No. 1, 2009, pp. 115-127

Criteria for the characterization of token causation

Tyler J. VanderWeele
Departments of Epidemiology and Biostatistics, Harvard University
e-mail: tvanderw@hsph.harvard.edu

ABSTRACT. It is argued that the definition of an "actual cause" given by Halpern and Pearl (2005) is neither necessary nor sufficient for some event to constitute an actual cause. Only when Halpern and Pearl sufficiently elaborate, in a manner of their own choosing, a structural model does their definition in fact identify the actual cause. It is proposed that a characterization of actual or token causation might be considered adequate if one could articulate either (1) a definition for the actual cause which supplies a necessary condition for an actual cause and is such that an elaborated structural model can always be constructed which will only identify the event or events constituting the actual cause; or (2) a definition for the actual cause which supplies a sufficient condition for an actual cause and is such that an elaborated structural model can always be constructed which will identify the event or events constituting the actual cause.

KEYWORDS: actual causation, causal diagrams, structural models, token causation

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/~episteme – ISSN: 1826-1043

1. Introduction

The problem of providing a characterization of token causation or of an "actual cause" has long eluded philosophers (Hall and Paul, 2003; Collins et al., 2004). Although in every-day conversation we can often agree that one event was the cause of another, trying to articulate precisely the rules by which we come to this agreement proves difficult.

Halpern and Pearl (2005) recently proposed a definition for an "actual cause" which can be applied to causal diagrams which represent causal structural models as formulated by Pearl (1995, 2000) and Halpern (2000). They demonstrate that their definition, in conjunction with sufficiently elaborated structural models, handles well a number of examples which are often considered problematic in the philosophical literature, such as those described by Bennett (1993) and Hall (2004). Halpern and Pearl (2005) use "actual cause" to make clear that they are interested in statements of the form "X caused Y (in this particular case)" as opposed to statements of the form "X is a cause of Y (in general, but perhaps not in this particular case)." Statements of the former type are often referred to as instances of "token causation" and have been the subject of interest in many philosophical analyses as well as in legal decisions; statements of the latter type are often referred to as instances of "type causation" and often are of interest in the social and biomedical sciences.

In the examples that Halpern and Pearl consider, their definition can be used to correctly identify the event which intuition suggests is the actual cause. Often the crudest causal diagram or structural model to describe one of these examples is insufficient to identify correctly the actual cause of a particular event but an elaborated structural model represented by a causal diagram with additional nodes does allow the use of their definition to identify the actual cause. In this paper two examples are used to argue that the definition given by Halpern and Pearl is neither necessary nor sufficient for some event to constitute an actual cause. It is not the goal of this paper to show that every possible account of actual causation using structural equations will fail but merely that the account given by Halpern and Pearl is not adequate. It is furthermore proposed that a characterization of actual or token causation might be considered adequate if one could articulate either (1) a definition for the actual cause which supplies a necessary condition for an actual cause and is such that an elaborated structural model can always be constructed which will only identify the event

or events constituting the actual cause; or (2) a definition for the actual cause which supplies a sufficient condition for an actual cause and is such that an elaborated structural model can always be constructed which will identify the event or events constituting the actual cause.

We will begin with a brief description of the framework and the definition of actual cause proposed by Halpern and Pearl (2005). We will then consider two examples, one from Halpern and Pearl (2005) and one introduced here, to argue that the definition given by Halpern and Pearl is neither necessary nor sufficient for some event to constitute an actual cause. We then discuss Halpern and Pearl's position, implied by their definition, that whether an event is the actual cause of another event is relative to a particular causal model. Finally, we propose criteria for an adequate characterization of token causation. We do not give a definition that satisfies these criteria but merely indicate what criteria it seems a definition of "actual cause" ought to satisfy in order to be considered an adequate characterization of token causation.

2. The Definition of Actual Cause in Halpern and Pearl

The definition of an "actual cause" given by Halpern and Pearl (2005) makes use of structural models as proposed by Pearl (1995, 2000) and Halpern (2000) which are sometimes graphically represented by causal diagrams. Halpern and Pearl (2005) define a causal model as a set of exogenous variables $\mathscr{U}$ determined outside the model, a set of endogenous variables $\mathscr{V}$ determined inside the model, a set $\mathscr{R}$ which for each variable $Y \in \mathscr{U} \cup \mathscr{V}$ specifies a range of possible values for Y and a collection of functions $\mathscr{F}$ which for each $X \in \mathscr{V}$ contains a function F_X which maps a value in $\mathscr{U}$ and a value in $\mathscr{V} \backslash X$ to a value in X. If we let z be any value of the variables in $\mathscr{V} \backslash X$, then the function $F_X(u,z)$ can be thought of as the value of X that would be obtained if the variables in $\mathscr{U}$ took on the values u and the variables in $\mathscr{V} \backslash X$ were set to z. The causal relations amongst the endogenous variables in $\mathscr{V}$ are sometimes represented on a causal diagram in which there is an arrow from $X_1 \in \mathscr{V}$ to $X_2 \in \mathscr{V}$ if F_{X_2} depends on the value X_1 obtains.

For illustration we consider an example given in by Halpern and Pearl (2005, Example 4.2) and adapted from Hall (2004) which we will return to below. The example concerns two rocks thrown at a bottle, one thrown by Suzy and one

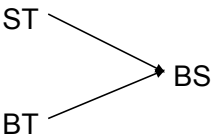

Figure 1: Example 4.2 from Halpern and Pearl (2005).

thrown by Billy; Suzy and Billy have perfect aim so that if either of them throw the rock the bottle will shatter. In the example of Halpern and Pearl, Suzy's rock hits the bottle first and is the actual cause of the bottle's shattering; however, if Suzy's rock had not been thrown, then Billy's rock would have hit the bottle and shattered it. Halpern and Pearl initially define three variables ST, BT and BS. The variable $ST = 1$ if Sally throws and 0 otherwise; the variable $BT = 1$ if Billy throws and 0 otherwise; the variable $BS = 1$ if the bottle shatters and 0 otherwise; all three variables are taken as endogenous. The variable $BS = 1$ if either $ST = 1$ or $BT = 1$ and 0 otherwise; the variable ST, however, does not depend on BT or BS and likewise the variable BT does not depend on ST or BS. The structural equations for the example are given by:

$$\begin{aligned} ST &= 1 \\ BT &= 1 \\ BS &= ST \vee BT. \end{aligned}$$

We thus could graphically represent the causal relations as given in Figure 1. Halpern and Pearl (2005) make the following proposal for a definition of an actual cause.

Definition 1 (Halpern and Pearl 2005) $X = x$ is the actual cause of some event ϕ if the following three conditions hold:

(AC1) Both $X = x$ and ϕ actually occur.

(AC2) There exists a partition (Z,W) of $\mathscr{V}$ with $X \subseteq Z$ and some setting (z',w') of (Z,W) such that if z^* denotes the actual value of Z then both

(a) changing (X,W) from its actual value (x,w) to (x',w') changes ϕ from true to false and

(b) (b) for all subsets W' of W and all subsets Z' of Z, setting W' to its value in w' and setting Z' to its value in z^* will leave ϕ true provided X is set to its actual value x.

(AC3) If X consists of multiple variables then no subset of X is such that conditions (AC1) and (AC2) can be satisfied for that subset.

If $X = x$ and ϕ satisfy conditions (AC1)–(AC3) then Halpern and Pearl (2005) say that $X = x$ is the actual cause of ϕ. Halpern and Pearl (2005) go on to discuss a possible refinement of the definition just given which restricts the endogenous variables to a set of "allowable settings." The definition just given above is then a special case of their refined definition. However, the comments that we make below apply also to their refined definition and for simplicity we will thus consider the definition of actual cause just given.

3. The Implications of Two Examples

As noted above, in their example 4.2, Halpern and Pearl (2005) consider two rocks thrown at a bottle, one by Suzy and one by Billy; Suzy's rocks hits the bottle first and is the actual cause of the bottle's shattering; however, if Suzy's rock had not been thrown, then Billy's rock would have hit the bottle and shattered it. In the simplest causal diagram that represents these causal relationships as given in Figure 1, the definition of an actual cause given by Halpern and Pearl identifies the throwing of both rocks as actual causes of the bottle's shattering. If ϕ is taken as $BS = 1$ then for Sally's throwing $X = ST = 1$ if we let $W = BT$ and $(x', w') = (0, 0)$ we have (AC1) $ST = 1$ and $BS = 1$, (AC2a) $BS = 0$ if $ST = 0$ and $BT = 0$, (AC2b) $BS = 1$ regardless of whether $W = BT$ takes the value 0 or 1, provided $ST = 1$ and (AC3) holds trivially. Thus the definition of Halpern and Pearl implies that Sally's throwing is the actual cause of the bottle's breaking. However, applying the same reasoning to Billy's throwing $X = BT = 1$, if we let $W = ST$ and $(x', w') = (0, 0)$ we have (AC1) $BT = 1$ and $BS = 1$, (AC2a) $BS = 0$ if $BT = 0$ and $ST = 0$, (AC2b) $BS = 1$ regardless of whether $W = ST$ takes the value 0 or 1, provided $BT = 1$ and (AC3) holds trivially. Thus the definition of Halpern and Pearl applied to Figure 1 implies also that Billy's throwing is the actual cause of the bottle's breaking; this is contrary to intuition. To remedy this problem, Halpern and Pearl consider a more elaborate causal diagram which, when used in conjunction with their definition, correctly identifies Suzy's rock and not Billy's rock as the actual cause of the bottle's shattering. Consider the causal diagram given in Figure 2, for example, where $SH = 1$ if Sally's rock hits

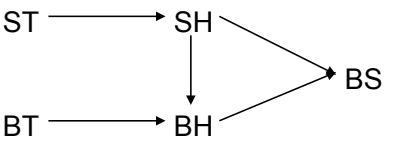

Figure 2: Elaborated causal diagram for Example 4.2 of Halpern and Pearl (2005).

the bottle and $SH = 0$ otherwise and $BH = 1$ if Billy's rock hits the bottle and $BH = 0$ otherwise.

The structural equations for the elaborated description of the causal relationships amongst the variables are given by

$$\begin{aligned} ST &= 1 \\ BT &= 1 \\ SH &= ST \\ BH &= BT \ \& \ (\text{not } SH) \\ BS &= SH \vee BH. \end{aligned}$$

Halpern and Pearl go on to show that if their definition is applied to Figure 2 then it correctly identifies Suzy's throwing the rock and not Billy's throwing the rock as the actual cause of the bottle's shattering. In any case, the simplest causal diagram for this example with two rocks demonstrates that the definition of Halpern and Pearl is not *sufficient* for a particular event to be an actual cause because, for the simplest causal diagram given in Figure 1, their definition implies that Billy's throwing the rock was an actual cause of the bottle's shattering. In defending their definition, Halpern and Pearl state (p. 845) that "the truth of every claim must be evaluated relative to a particular model of the world." We will return below to the question of whether it is reasonable to claim actual causation is relative in this way.

Note that there are certain cases in which two events in some way interact so that the effect comes about if and only if both events occur; in such cases, it is appropriate to speak of two actual causes and an adequate characterization of what constitutes an actual cause will be able to handle these cases with two or more actual causes. In the case of the bottle's shattering, however, we would

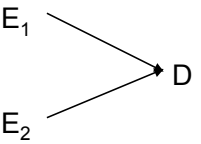

Figure 3: Example showing the definition of Halpern and Pearl (2005) is not a necessary condition for an event to be an actual cause.

ordinarily only say that Suzy's throwing her rock was the actual cause of the bottle's shattering.

We now consider a second example that demonstrates that the definition of Halpern and Pearl is also not a necessary condition for an event to be an actual cause. Suppose that a man is suffocating and if left untreated he will die at time t. A physician is present and has two possible treatments he can administer: injection 1 and injection 2. If injection 1 is given the man will be able to breathe; however, he is allergic to the chemical compounds constituting injection 1 and will thus die of heart failure, again at time t, resulting from an allergic reaction to injection 1. If injection 2 is given the man will once again be able to breathe but he is also allergic to the chemical compounds constituting injection 2 and thus if he is given injection 2 he will once again die of heart failure at time t. If he is given both injections 1 and 2, the two sets of chemicals interact and although the man is once more able to breathe, he will die of a failure of the nervous system at time t. The physician knows that the man is allergic to the chemical compounds constituting injection 2 but he does not know that the man is allergic to the chemical compounds constituting injection 1; the physician thus gives the man the first injection and the latter dies, at time t, of heart failure from injection 1. Quite clearly, the actual cause of this man's death is his receiving injection 1. A causal diagram could be constructed involving only the two injections, E_1 and E_2 respectively, and the outcome D with $D = 1$ indicating death at time t and $D = 0$ otherwise, with arrows from E_1 and E_2 pointing to D as in Figure 3.

The structural equations could be written as

$$\begin{aligned} E_1 &= 1 \\ E_2 &= 0 \\ D &= (E_1 \,\&\, E_2) \vee (E_1 \,\&\, \text{not } E_2) \vee (\text{not } E_1 \,\&\, E_2) \vee (\text{not } E_1 \,\&\, \text{not } E_2) = 1 \end{aligned}$$

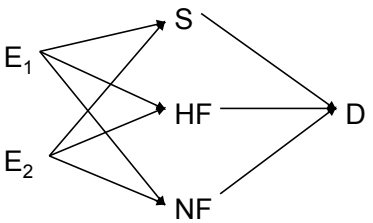

Figure 4: Elaborated causal diagram from Figure 3 showing the possible causes of death.

The definition of an actual cause given by Halpern and Pearl applied to Figure 3, will not identify either of the injections as the actual cause of death nor the combination of these injections nor the absence of them both. This is because the definition of Halpern and Pearl would require that under some set of values for the variables on the graph $D = 0$; however, in the scenario just given D will be 1 regardless of the values of E_1 and E_2. In this case, the definition of Halpern and Pearl delivers the peculiar conclusion that for this model there is no actual cause of the man's death.

It is possible to construct an elaborated causal diagram for this example such that the definition of Halpern and Pearl does correctly identify injection 1 as the actual cause of death. Consider for instance the causal diagram given in Figure 4 where S indicates suffocation, HF indicates heart failure and NF indicates failure of the nervous system.

The structural equations corresponding to the elaborated diagram can be written as

$$\begin{aligned} E_1 &= 1 \\ E_2 &= 0 \\ S &= \text{not } E_1 \ \& \ \text{not } E_2 \\ HF &= (E_1 \ \& \ \text{not } E_2) \vee (\text{not } E_1 \ \& \ E_2) \\ NF &= E_1 \ \& \ E_2 \\ D &= S \vee HF \vee NF. \end{aligned}$$

In actual fact, we have $E_1 = 1$, $E_2 = 0$, $S = NF = 0$ and $HF = D = 1$. Now let $X = E_1$ and $x = 1$, $Z = \{E_1, E_2, HF, D\}$, $W = \{S, NF\}$ and $(x', w') = (0,0,0)$ we then have (AC1) $E_1 = 1$ and $D = 1$ actually occur, (AC2a) $D = 0$ if E_1

is set to 0 and S and NF are both set to 0 with E_2 at its original value of 0, (AC2b) provided E_1 is set to 1, then $D = 1$ regardless of whether S and NF are set to 0 and regardless of whether E_2 or HF are set to their original values 0 and 1 respectively; also (AC3) holds trivially. Thus the definition of Halpern and Pearl applied to Figure 4 implies that injection 1 is the actual cause of the individual's death. However, the simplest causal diagram, given in Figure 3, and the corresponding structural causal model for this example, demonstrate that the conditions given in the definition of Halpern and Pearl are not *necessary* for an event to be an actual cause.

We note that the critique offered here differs from that offered by Menzies (2004). Menzies considers an earlier proposal of Halpern and Pearl (2001) and points out that their definition does not handle well cases in which token causation is not transitive. Menzies gives an example in which an assassin's poisoning the king's coffee causes a guard to put in an antidote into the coffee; the antidote without the poison would also be lethal but the antidote with the poison neutralizes the poison; the guard's putting the antidote into the coffee thus causes the king's survival. We would not ordinarily say that the assassin's poisoning the coffee caused the king's survival but Menzies shows that the definition given in Halpern and Pearl (2001) would identify the assassin's poisoning the coffee as the actual cause of the king's survival. Halpern and Pearl, in their (2005) paper, give some attention to such examples using their refined definition of actual cause restricting the endogenous variables to a set of "allowable settings" but they do not explicitly discuss Menzies' example.

4. The Claim of the Relativity of Actual Causation

The two examples in the preceding section indicate that the definition of an "actual cause" given by Halpern and Pearl is neither necessary nor sufficient for an event to constitute an actual cause. Pearl and Halpern seem to attempt to get around this issue by proposing that whether an event is an actual cause of another event is relative to the structural model under consideration. They are unconcerned by the fact that their definition, in the structural model represented in Figure 1, identifies Billy's stone as an actual cause of the bottle's shattering because their position is that (p. 845), "according to our definition, the truth of every claim must be evaluated relative to a particular model of the world; that is

our definition allows us to claim only that C causes E in a (particular context in a) particular structural model." They further note that, "It is possible to construct two closely related structural models such that C causes E in one and C does not cause E in the other." They consider this "a feature of our model, not a bug."

Their position is problematic for a number of reasons. First, it seems as though the definition of Halpern and Pearl is missing an important feature of our discourse concerning actual causation. When we are presented with insufficient information to determine what the actual cause is, our judgement is usually one of agnosticism. For example, if, along the lines of Figure 1, we were only told that the bottle shatters if and only if either Sally or Billy throw a rock and in fact they both threw a rock, we would ordinarily conclude that we didn't have enough information to determine whether Sally's throwing or Billy's throwing was the actual cause of the bottle's shattering. We would respond agnostically; we would not generally conclude that the throwing of both were actual causes and then revise our assessment once further information was available. It would seem that an adequate definition of an actual cause would allow for such agnosticism and would in fact deliver a judgement of agnosticism precisely when we ordinarily do so in every-day reasoning. As we have seen above, Halpern and Pearl's definition does not provide this judgement of agnosticism, at least not in the case considered.

Second, it is not at all clear that the implications of the relativity of Halpern and Pearl's definition of actual causation cohere with the sorts of conclusions we ordinarily come to concerning the use of expressions like "C caused E". In every-day discussion, we generally achieve near universal agreement concerning which event or events constitute an actual cause; certainly, simple examples such as those presented above would not by themselves lead us to think that the notion of "actual cause" is relative in the way that Halpern and Pearl claim. It might be possible come up with examples in which intuitions about which event is the actual cause are not clear so that it is not possible to achieve consensus i.e. examples in which, if looked at one way, some event seems as though it were the actual cause whereas if looked at another way, a different event seems like the actual cause rather than the first. However, the examples considered above are not of this variety; in these cases, the intuition is clear as to what the actual cause in fact is. It seems that a definition should not have relativistic implications for examples in which the language and reasoning we in fact use about actual causation is clear.

Finally, as a related point, it seems that if our ordinarily use of the expressions like "actual cause" and "*C* caused *E*" are not obviously relative, then a case would need to be made that they are in fact so. Again it might be possible to construct examples in which the relativity of the use of the expressions "actual cause" and "*C* caused *E*" is made clear; however, this has not been accomplished in their paper. Halpern and Pearl seem to take the relativity of actual causation as their premise rather than their conclusion.

In the end it is not wholly clear what their definition contributes to reasoning. The approach Halpern and Pearl seem to take in making use of their definition is the following: an event may be considered an actual cause if one can construct a structural model such that their definition identifies the actual cause which agrees with intuition. However, as a characterization of what is meant by "actual cause" it is not clear what their approach contributes beyond relying entirely on intuition from the start.

5. Criteria for the Characterization of Token Causation

An advance beyond Halpern and Pearl would perhaps be made if it were possible to develop a definition of an actual cause that satisfied one of two sets of criteria. On the one hand, it might be possible to devise a definition for an actual cause which, (1a) when applied to any structural model corresponding to the description of the causal relationships, however crude, will always identify some set of events amongst which is included the event or events constituting the actual cause (i.e. for which the definition is a necessary condition for the actual cause) and for which (1b) an elaborated structural model can always be constructed which will identify only the event or events constituting the actual cause. On the other hand, it might be possible to devise a definition for the actual cause which, (2a) when applied to any structural model corresponding to the description of the causal relationships, however crude, would be satisfied only for events which were actual causes (i.e. for which the definition is a sufficient condition for the actual cause) and for which (2b) an elaborated structural model can always be constructed which will in fact identify the event or events constituting the actual cause. The example given above concerning the man suffocating demonstrates that the definition of Halpern and Pearl does not satisfy the first of these two sets of criteria. The example concerning the two rocks and

the bottle's shattering demonstrates the definition of Halpern and Pearl does not satisfy the second of these two sets of criteria.

Definitions which satisfied either the first or the second set of criteria would have the potential to give non-relative answers in accordance with intuition in cases in which such intuition was clear. However, whether they in fact did so would of course have to be examined once a definition were proposed. A definition satisfying the second set of criteria above would also have the potential to deliver judgements of agnosticism, discussed above, in that in simple models in which there is insufficient information to determine the actual cause, it might then not identify any event as the actual cause; however, since (2a) requires the definition is simply a sufficient condition, the conclusion would not then be that there is no actual cause but simply that nothing can be said in this case without further information (e.g. without an elaborated structural model). If definitions satisfying the first or the second set of criteria were proposed, it would remain to be determined whether examples could be constructed in which the event or events constituting the actual cause were relative, both in intuitive reasoning and as assessed by the definition.

A definition which satisfied one of the two sets of the criteria above would constitute a considerable advance in the problem of characterizing what constitutes token causation or an "actual cause." If two definitions could be provided, each satisfying one of the two sets of criteria given above, the concept of token causation could perhaps then be considered adequately characterized. Whether or not this is possible remains to be seen.

REFERENCES

BENNETT, J. (1993). "Event causation: the counterfactual analysis". In E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, pp. 217–33.

COLLINS, J., HALL, N. AND PAUL, L. A. (2004). "Counterfactual and causation: history, problems and prospects". In J. Collins, N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, Cambridge, MA: MIT Press, pp. 1–58.

HALL, N. (2004). "Two concepts of causation". In J. Collins, N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, Cambridge, MA: MIT Press, pp. 225–76.

HALL, N. AND PAUL, L. A. (2003). "Causation and preemption". In P. Clark and K. Hawley (eds.), *Philosophy of Science Today*, Oxford, Oxford University Press, pp. 100–29.

HALPERN, J. Y. (2000). "Axiomatizing causal reasoning". *Journal of Artificial Intelligence* 12, pp. 317–37.

HALPERN, J. Y. AND PEARL, J. (2001). "Causes and explanations: a structural model approach. Part I: Causes". Technical Report R-266, Cognitive Science Laboratory, Los Angeles: University of California.

HALPERN, J. Y. AND PEARL, J. (2005). "Causes and explanations: a structural-model approach. Part I: Causes". *British Journal of the Philosophy of Science* 56, pp. 843–87.

MENZIES, P. (2004). "Causal models, token causation, and processes". *Philosophy of Science* 71 pp. 820–832.

PEARL, J. (1995). "Causal diagrams for empirical research". *Biometrika* 82 pp. 669–688.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, Cambridge University Press.