The Intransitivity of Causation Revealed in Equations and Graphs

Author(s): Christopher Hitchcock

Source: *The Journal of Philosophy*, Jun., 2001, Vol. 98, No. 6 (Jun., 2001), pp. 273-299

Published by: Journal of Philosophy, Inc.

Stable URL: https://www.jstor.org/stable/2678432

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/2678432?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

is collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Philosophy*

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

THE JOURNAL OF PHILOSOPHY

VOLUME XCVIII, NO. 6, JUNE 2001

THE INTRANSITIVITY OF CAUSATION REVEALED IN EQUATIONS AND GRAPHS*

We live in exciting times. By 'we' I mean philosophers studying the nature of causation. The past decade or so has witnessed a flurry of philosophical activity aimed at cracking this nut, and, surprisingly, real progress has been made. Two developments are especially worthy of note.[1] First, there has been a resurgence of interest in the counterfactual theory of causation, given its best-known formulation by David Lewis.[2] Second, there has been increasing philosophical interest in the techniques of causal modeling developed and employed within fields such as econometrics, epidemiology, and artificial intelligence.[3] These two developments have been largely independent and addressed to different sorts of problems. Work in the counterfactual tradition has been primarily concerned with issues involving "token" or "singular" causation, while work in the second tradition has tended to focus on issues concerning "type-level" or "general" causation.

* For discussion and comments, thanks go to Martin Barrett, Paul Bartha, Ellery Eells, Clark Glymour, Alan Hájek, Dan Hausman, Mark Kalderon, Henry Kyburg, Michael McDermott, Laurie Paul, Augustin Rayo, Jonathan Schaffer, Elliott Sober, Jim Woodward, Stephen Yablo, and especially Judea Pearl.

[1] I do not mean to disparage other important developments in the theory of causation, such as the development of a theory of causal processes in terms of conserved quantities—see Phil Dowe's *Physical Causation* (New York: Cambridge, 2000)—but I shall not talk about them here.

[2] "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556-67, reprinted in *Philosophical Papers, Volume II* (New York: Oxford, 1986), pp. 159-72. For examples of important recent work on counterfactual theories of causation, see the essays in this JOURNAL, XCVII, 4 (April 2000); and in John Collins, Ned Hall and L. A. Paul, eds. *Causation and Counterfactuals* (Cambridge: MIT, forthcoming).

[3] See, for example, the essays in Vaughan R. McKim and Stephen P. Turner, eds., *Causality in Crisis?* (Notre Dame: University Press, 1997).

0022-362X/01/9806/273–99 © 2001 The Journal of Philosophy, Inc.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

Fortunately, these two developments are just beginning to merge: the computer scientists Judea Pearl and Joseph Halpern[4] have developed theories of token causation in terms of structural equations of the sort used in causal modeling. The account I shall present here bears a very strong resemblance to the theories developed by Pearl and Halpern. I shall not present their account explicitly and contrast it with mine, but I urge philosophers to read this work on their own.

As a case study, I shall explore the problem of the transitivity of causation from within the structural equations framework. Causation is transitive if and only if, whenever *a* causes *b* and *b* causes *c*, then *a* causes *c* as well. Many philosophers, notably Lewis, have claimed that causation is transitive. Others[5] have raised powerful objections to the transitivity of causation. I side with the objectors. This is not to say that I shall offer a conclusive refutation of the transitivity thesis: the transitivity fetishist may well be able to preserve her cherished principle by tweaking my account in various ways. But I shall show that there is no *independent* motivation for accepting the transitivity thesis: all of the benefits of maintaining transitivity can be had without paying the costs.

By working within the structural equations framework, I shall not be abandoning the traditional counterfactual framework. Rather, I shall present structural equations as tools for representing patterns of counterfactual dependence. These tools allow us to make explicit the structural differences between those cases which appear to instantiate the transitivity of causation, and those which appear to be counterexamples. The structural equations framework draws attention to a certain kind of counterfactual whose importance to the analysis of causation has hitherto been ignored by philosophers. I shall use this type of counterfactual to define the notion of an *active route* between two events. If it is possible to give a reductive analysis of counterfactuals in purely *acausal* terms, as Lewis[6] believes, then I offer a reductive analysis of active routes. In those cases which appear to be counterexamples to the transitivity of causation, the putative cause and effect fail to be connected by an active route.

[4] Pearl, *Causality* (New York: Cambridge, 2000), chapter 10; Halpern and Pearl, "Causes and Explanations: A Structural-model Approach," Technical report R-266, Cognitive Systems Laboratory, University of California/Los Angeles, 2000.

[5] See especially Michael McDermott, "Redundant Causation," *British Journal for the Philosophy of Science,* XL (1995): 523-44.

[6] See, for example, "Counterfactual Dependence and Time's Arrow," *Noûs,* XIII (1997): 455-76, reprinted in *Philosophical Papers, Volume II,* pp. 32-52; and *Philosophical Papers, Volume II,* introduction.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

In focusing on the problem of transitivity, I shall have little or nothing to say about a variety of other problems in the theory of causation. I shall assume determinism, and hence leave unsolved important problems involving indeterministic causation. I shall discuss only briefly the problems of symmetric overdetermination, late preemption, and "trumping." My strategy will be to take Lewis's account of causation as a foil,[7] argue that my account does a better job on the issue of transitivity, and say just enough about these further issues to suggest that my account fares no worse in these arenas.

I. TRANSITIVITY: ANATOMY OF A PUZZLE

Our point of departure will be Lewis's original counterfactual theory of causation. Let *c* and *e* be distinct events that both occurred. Then *e counterfactually depends upon c* if and only if, if *c* had not occurred, *e* would not have occurred. Lewis takes counterfactual dependence to be sufficient but not necessary for causation: causation is defined as the *ancestral* of counterfactual dependence, rendering causation transitive by definition.

Lewis[8] offers a detailed account of what must be "held fixed" when evaluating the relevant counterfactuals. For our purposes, it is enough to point to two features that these counterfactuals must have. First, the counterfactuals must not *backtrack*. If *a* caused *c* (but not vice versa), then a counterfactual of the form 'If *c* had not occurred, then...' must hold *a* fixed. If not, the theory would incorrectly rule that *c* causes *a*. Second, the counterfactuals must "foretrack": if *c* causes *e*, we do not want to hold *e* fixed when evaluating the counterfactual 'If *c* had not occurred, then...'. If we do, the consequent of the conditional will obviously not be '*e* would not have occurred'.

Note that these are restrictions on what is to be *tacitly* held fixed when entertaining counterfactual antecedents; they are not restrictions on the antecedents that we are permitted to entertain. The necessity of foretracking from *c* to *e* does not prevent us from entertaining a counterfactual of the form 'If *c* had not occurred, but *e* had occurred anyway, then...'.[9] Let us call a counterfactual of this sort an *explicitly nonforetracking* or ENF counterfactual. In the sequel, I shall argue that ENF counterfactuals should play a central role in the analysis of causation.

[7] I do not deny that other counterfactual theories of causation would make worthy foils. Space limitations prohibit detailed comparison with all such rivals.

[8] "Counterfactual Dependence and Time's Arrow."

[9] Lewis himself makes this point in "Finkish Dispositions," *Philosophical Quarterly*, XLVII (1997): 143-58, here p. 150.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

Why does Lewis define causation as the *ancestral* of counterfactual dependence? There is certainly a strong pretheoretic intuition that causation is transitive, but this can be explained without building transitivity into the analysis. Let us call a case of causation *ordinary* if it has this structure: *e* depends counterfactually upon *d*, which in turn depends counterfactually upon *c*, and *e* also depends counterfactually upon *c*. *Most* cases of causation are ordinary, and in such cases we can explain why *c* counts as a cause of *e* just by identifying causation with counterfactual dependence.

Trouble arises in "extraordinary" cases. In many such cases, we judge that there is causation without counterfactual dependence. Here is a standard example.

> "Backup": an assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.

In this case, it seems that Trainee's shot caused the death of Victim, even though Victim's death does not counterfactually depend upon Trainee's shot. This is a case of *preemption*: by shooting at Victim, Trainee preempted a process that would itself have resulted in Victim's death. The standard solution is to invoke the transitivity of causation: Trainee's shot is a cause of Victim's death because there is a chain of counterfactual dependence running from the former to the latter. Consider a further event (call it '*b*') that was not mentioned explicitly in the foregoing description: the presence of a bullet en route from Trainee to Victim. Had Trainee not shot, *b* would not have occurred; and if *b* had not occurred, Victim would not have died. Note the importance of the "no backtracking" rule in the second counterfactual: if *b* had not occurred, Trainee would have shot anyway, so Supervisor would not have shot.

Other examples suggest, however, that causation is not transitive in general.

> "Boulder": a boulder is dislodged, and begins rolling ominously toward Hiker. Before it reaches him, Hiker sees the boulder and ducks. The boulder sails harmlessly over his head with nary a centimeter to spare. Hiker survives his ordeal.[10]

[10] This example is from an early draft of Hall's "Two Concepts of Causation," forthcoming in Collins et alia.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

> "Dog Bite": Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb. Shortly before noon, he is bitten by a dog on his right hand. Unable to use his right hand, he pushes the detonator with his left hand at noon. The bomb duly explodes.[11]

In each case, we have a chain of counterfactual dependence. Hiker would not have ducked if the boulder had not fallen; and given the trajectory of the boulder, he would not have survived if he had not ducked. The dog bite caused Terrorist to push the detonator with his left hand, and his pushing the detonator with his left hand caused the bomb to explode. But the fall of the boulder did not cause Hiker to survive and the dog bite did not cause the explosion—these are the verdicts of common sense.

We should be particularly troubled that we judge there to be a causal relationship in that case where the chain of counterfactual dependence is hardest to see. In "Backup," we must find an intermediate event that is not made salient in the presentation of the example, namely, *b*, and imagine it away while holding Trainee's shot fixed. We are to imagine the bullet vanishing into thin air or some such. This is no homey piece of counterfactual reasoning, but requires the efforts of a trained philosopher. In "Boulder" and "Dog Bite," by contrast, it is easy to see the needed intermediate events: Hiker's ducking and Terrorist's pushing the detonator with his left hand (respectively). These two events were specified as parameters of the examples. Moreover, the necessary counterfactual reasoning is not particularly straining: it is easy to imagine, for example, what would have happened if Hiker had not ducked, even holding fixed the boulder's fall. The defender of transitivity must maintain that we have inconsistently acute intuitive powers: we see causal relations that are underwritten by obscure chains of counterfactual dependence, and yet we are blind to causal relations that are underwritten by obvious chains of counterfactual dependence.

I shall argue that we can accept "Dog Bite" and "Boulder" as counterexamples and provide an alternative account of "Backup." Trainee's shot caused Victim's death, not because there is a chain of counterfactual dependence, but because there is an *active route* between Trainee's shot and Victim's death. The existence of this route is revealed by an ENF counterfactual: if Trainee had not shot, and Supervisor still did not shoot, then Victim would not have died. There are no comparable ENF counterfactuals in "Boulder" and "Dog Bite."

[11] From McDermott, "Redundant Causation."

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

II. FURTHER PROBLEMS

While my primary focus will be on the problem of transitivity, I shall briefly mention a few further problems with Lewis's theory along with his recent attempt[12] to solve some of them. The first sort of problem involves cases of *symmetric overdetermination*, in which two events have an equal claim to be causes of a third event, which does not depend counterfactually upon either of them. Lewis[13] claims that he has no clear intuitions about such cases, and thus dismisses them as being of no diagnostic value.

A second type of problem involves cases of preemption that have different structures from "Backup." Lewis[14] uses the term *early cutting* to describe this structure. In addition, he recognizes cases of *trumping*[15] and *late cutting*. Here is an example of the latter:

> Billy and Suzy both throw rocks at a bottle. Suzy's throw gets there first, shattering the bottle. Billy's throw arrives at the scene a split second later, encountering nothing but air where the bottle used to be.[16]

Suzy's throw is clearly a cause of the bottle's shattering, Billy's not. The shattering of the bottle does not counterfactually depend upon Suzy's throw, and there is no chain of counterfactual dependence from Suzy's throw to the bottle's shattering.[17]

Lewis responds to these problems by modifying his theory of causation. He first defines the notion of *influence*:

> Where c and e are distinct actual events,...c influences e if and only if there is a substantial range c_1, c_2,...of different not-too-distant alterations of c (including the actual alteration of c) and there is a range e_1, e_2,...of alterations of e, at least some of which differ, such that if c_1 had occurred, e_1 would have occurred, and if c_2 had occurred, e_2 would have occurred, and so on.[18]

(An alteration c' of c is a fine-grained event that is similar to c, but possibly different in matters of detail.) Suzy's throw influences the shattering of the bottle: had Suzy thrown slightly earlier, or aimed at a slightly different point, the bottle would have shattered slightly

[12] "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182-97.
[13] Both in "Causation" and in "Causation as Influence."
[14] "Causation as Influence."
[15] For a detailed discussion of trumping, see Schaffer, "Trumping Preemption," this JOURNAL, XCVII, 4 (April 2000): 165-81.
[16] This example appears in "Causation as Influence," but has been in the lore for some time.
[17] For a defense of the latter claim, see "Postscript E" to "Causation" in Lewis's *Philosophical Papers, Volume II*, pp. 193-213.
[18] "Causation as Influence," p. 190.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

earlier, or in a slightly different way. Lewis then defines causation as the ancestral of influence, citing the problem of early cutting preemption as his motivation for doing so.[19]

I conclude this section by pointing to an interesting feature of Lewis's new account: it is infected with context sensitivity. In order for c to influence e, there must be true counterfactuals involving a *substantial range* of *not-too-distant* alterations of c. It may be an objective matter whether a set of counterfactuals of the form 'If c_i had occurred, e_i would have occurred' are true, but there is a further question about whether we should describe this pattern by saying 'c causes e'. The answer to this question will depend, in part, upon which unactualized possibilities we consider "too distant" to take seriously. This point is underscored in Lewis's[20] brief discussion of *preemptive prevention*. The account developed below will agree with Lewis's new account on this point.

III. EQUATIONS AND GRAPHS FOR DUMMIES

The use of directed graphs to represent systems of causal relationships dates back at least to the work of Sewall Wright[21] in the early 1920s; the use of structural equations, pioneered by R. Frisch,[22] T. Haavelmo[23] and others in the 1930s and 1940s is not much younger. Pearl's *Causality* is representative of the current state of the art.[24]

A system of structural equations is a sequence of equations $\mathscr{E}$ relating the values of variables belonging to some set $\mathscr{V}$. An ordered pair $\langle \mathscr{V}, \mathscr{E} \rangle$ will be called a *causal model*. In the simplest case, a variable

[19] See especially "Causation as Influence," p. 194. Note that in Lewis's new account, transitivity would no longer be needed to handle "Backup": the precise time and manner of Victim's death does indeed depend upon the precise time and manner of Trainee's shot. Lewis maintains that transitivity will still be needed to handle fancier examples of early cutting. I shall continue to use "Backup" in unmodified form; this will not change the fundamental structure of our central problem.

[20] "Causation as Influence," p. 197. For more detailed discussion, see Collins, "Preemptive Prevention," this JOURNAL, XCVII, 4 (April 2000): 223-34.

[21] "Correlation and Causation," *Journal of Agricultural Research*, XX (1921): 557-85.

[22] "Statistical versus Theoretical Relations in Economic Macrodynamics," League of Nations Memorandum, 1938; reprinted as "Autonomy of Economic Relations" in D. Hendry and M. Morgan, eds., *The Foundations of Econometric Analysis* (New York: Cambridge, 1995), pp. 407-19.

[23] "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, XI (1943): 1-12, reprinted in Hendry and Morgan, pp. 454-63; "The Probability Approach in Econometrics," Supplement to *Econometrica*, XII, reprinted, in part, in Hendry and Morgan, pp. 440-53, 477-90.

[24] Philosophers will probably be most familiar with the work of Peter Spirtes, Clark Glymour, and Richard Scheines, especially their *Causation, Prediction, and Search* (New York: Springer, 1993). This work is very much in the same vein, although I shall not be using their probabilistic framework here.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

will have only two values, and will represent the occurrence or nonoccurrence of a certain event. For example, in representing the causal relations in "Backup," we shall use a variable T, which takes the value 1 when Trainee shoots, and takes the value 0 when he does not. As a general convention, variables will be represented by italicized capital letters. When a binary variable E is used to represent the occurrence or nonoccurrence of an event e, $E = 1$ will represent the occurrence of e, and $E = 0$ the nonoccurrence of e. But variables need not be binary. For example, the values of a variable could represent the mass of some object, or they could represent various alterations of a particular event. $\mathscr{V}$ contains both *exogenous* and *endogenous* variables.

Each equation in $\mathscr{E}$ expresses the value of one variable, appearing on the left-hand side of the equation. Each variable appears on the left-hand side of exactly one such equation. $\mathscr{E}$ is comprised of two subsets, $\mathscr{E}_e$ and $\mathscr{E}_n$. The former contains equations with exogenous variables on the left-hand side, while the latter contains equations with endogenous variables on the left hand side. Equations in $\mathscr{E}_e$ all take the simple form $X = x$: they simply state the actual value of the variable in question. Equations in $\mathscr{E}_n$ express the value of the endogenous variable as a function of the values of other variables in the set $\mathscr{V}$:

(*) $Z = f_Z(X, Y, \ldots, W)$

The syntax of such an equation is richer than that of ordinary mathematical equations. In particular, structural equations are not symmetric: (*) is not equivalent to $f_Z(X, Y, \ldots, W) = Z$. (In structural equations, side matters.) This is because structural equations encode counterfactuals. For example, (*) encodes a set of counterfactuals of the following form:

> If it were the case that $X = x$, $Y = y, \ldots, W = w$, then it would be the case that $Z = f_Z(x, y, \ldots, w)$.

These counterfactuals are to be understood along the lines discussed in section I above; in particular, they do not backtrack.

Equations in $\mathscr{E}_n$ must always be written in minimal form: if for all x, x', y, $z, \ldots, w$, $f_Z(x, y, \ldots, w) = f_Z(x', y, \ldots, w)$, then the value of Z does not depend upon the value of X at all, and the structural equation for Z must be rewritten $Z = f_Z(Y, \ldots, W)$. As written, then, the equation (*) says that the value of the variable Z depends counterfactually upon the values of the variables $X, Y, \ldots W$. By the same token, equations in $\mathscr{E}_n$ must always include as arguments any variables in $\mathscr{V}$ upon which Z counterfactually depends, given the values of the other variables. If, for some x, x', y, $z, \ldots, w$, $f_Z(x, y, \ldots, w) \neq f_Z(x', y, \ldots, w)$, then the value

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

of Z does depend upon the value of X, and $Z = f_Z(Y,..., W)$ is not in $\mathscr{E}_u$. The correct equation for Z can be arrived at by expressing the value of Z as a function of *all* other variables in $\mathscr{V}$, and then eliminating those variables whose values are redundant given every assignment of values to the other variables.

It will sometimes be helpful to use symbols familiar from sentential logic to represent relations between variables. The symbols $\neg$, $\vee$, and $\wedge$ will represent the following mathematical functions: $\neg X \equiv 1 - X$, $X \vee Y \equiv \max\{X, Y\}$, $X \wedge Y \equiv \min\{X, Y\}$. When the variables are binary, these functions behave much as the corresponding connectives do in sentential logic. For example, if $Z = X \vee Y$, then Z will take the value 1 if and only if either X or Y takes the value 1.

If the variable X figures as an argument on the right-hand side of the structural equation for the endogenous variable Z, then X is a *parent* of Z. Exogenous variables have no parents within a model, while endogenous variables do.

A system of structural equations can be given an elegant graphical representation. The variables in $\mathscr{V}$ form the nodes of a graph. The nodes are connected by directed edges or "arrows" according to the following rule: an arrow is drawn from X to Z if and only if X is a parent of Z. A *directed path* from variable X to variable Z is a sequence of arrows lined up "tip-to-tail" connecting X with Z. A variable is exogenous if there is no arrow directed into it.

If the structural equations can be ordered so that no variable appears on the left-hand side after having appeared on the right-hand side, then the system of equations is *acyclic*. Equivalently, a system of structural equations is acyclic if no directed path in the corresponding graph runs from a given node back into itself. Intuitively, an acyclic system of equations represents a causal structure in which there are no causal loops. All of the systems considered here will be acyclic. If a system of equations is acyclic, then the system of equations has a unique solution. That is, the values of the exogenous variables together with the other structural equations entail a unique value for every variable.

I shall illustrate structural equations and graphical representations using "Backup." The causal graph is depicted in figure 1. The variables are to be interpreted as follows. $T = 1$ corresponds to Trainee's shooting, $T = 0$ to his refraining; $S = 0$ or 1 depending upon whether Supervisor shoots; $B = 0$ or 1 depending upon the presence of a bullet in flight at some point along the line from Trainee to Victim; and $V = 0$ or 1 according to whether Victim dies. The set of structural equations is:

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

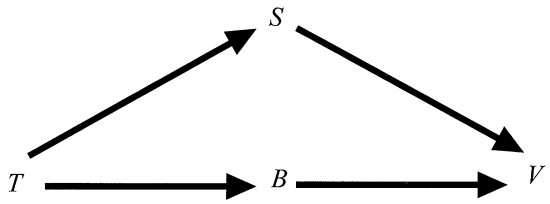

Figure 1

$\mathscr{SE}$: $T = 1$; $S = \neg T$; $B = T$; $V = B \vee S$

T is an exogenous variable. The equation, $V = B \vee S$ encodes the following counterfactuals: if either B or S were to take the value 1—if either *b* or Supervisor's shot were to occur—then Victim would die; if B and S were both to take the value 0, then Victim would not die. $\mathscr{SE}$ has the following unique solution:

$T = 1$; $S = 0$; $B = 1$; $V = 1$

That is: Trainee actually shot; Supervisor did not shoot; *b* occurred; and Victim died.

It has become common in the philosophical literature to represent causal relationships in terms of *neuron diagrams*. A circle is used to represent an event which may occur ("fire") or not. A shaded circle represents an event that did occur, while a hollow circle represents an event that did not occur. Neurons may be connected in one of two ways: by a stimulatory connection, represented by an arrow, or by an inhibitory connection, represented by an arrow with a backward head. A stimulatory connection between two events indicates that, if the first occurs, it will cause the second to occur; an inhibitory connection indicates that, if the first occurs, it will prevent the second from occurring (inhibitory connections override stimulatory ones). Causal graphs work similarly, but with several important differences. First, each vertex represents a *variable*. A variable may be binary, taking values of 1 or 0 according to whether some event occurs or not, but it may also be multivalued. Second, the causal graph does not itself specify the actual value of the variable; that information is contained in the accompanying set of structural equations. Third, an arrow from one variable to another says nothing about the *kind* of connection that exists between them—once one allows nonbinary variables, the possible connections cannot be classified neatly into

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

stimulatory and inhibitory.[25] I have argued elsewhere[26] that attempts to shoe-horn all causal relationships into these two types lead to a number of pseudo-problems in the theory of causation. Instead, the nature of the connection between variables is represented in the corresponding system of structural equations.

Each equation in $\mathscr{E}_{\mathscr{M}}$ encodes counterfactual information. Note, however, that the equations in $\mathscr{E}_{\mathscr{M}}$ do not directly represent *all* counterfactuals that are true of the system. Rather, $\mathscr{E}$ is a set of *fundamental* equations from which all other counterfactuals may be derived. In general, in order to evaluate the counterfactual 'If it were the case that $X = x$, $Y = y$,..., $W = w$, then...', we replace the equation for each of these variables with the identity stipulated; for example, we replace the equation for X with $X = x$. The equations for the other variables remain unchanged. In effect, this creates a new set of structural equations in which X, Y,..., W are exogenous variables. Graphically, the arrows directed into these variables are removed, while all other arrows remain intact. Instead of these variables having their values causally determined in the normal way, they are "miraculously" set to the new hypothetical values. The values of the remaining variables can then be computed. The result may be thought of as characterizing the "closest possible world(s)" where $X = x$, $Y = y$,..., $W = w$, are all true. A variable Z depends counterfactually upon a variable X in a system of structural equations if and only if in the actual solution, $X = x$, $Z = z$ and there exist $x' \neq x$ and $z' \neq z$ such that the result of replacing the equation for X with $X = x'$ yields $Z = z'$. This says that there is some possible value of X such that if X had taken that value, then the value of Z would have been different.

The standard treatment of "Backup" requires the counterfactual: 'If the bullet had not been in flight from Trainee to Victim, then Victim would not have died'. Modifying $\mathscr{SE}$ by setting $B = 0$, we get the following modified set of structural equations:

$$\mathscr{SE}'\text{: } T = 1;\ S = \neg T;\ B = 0;\ V = B \vee S$$

[25] Indeed, even with binary variables, this problem can arise as soon as you have two "neurons," each connected to a third. Suppose that the third neuron fires if exactly one of the others does; then the connections are neither purely stimulatory nor purely inhibitory.

[26] "A Generalized Probabilistic Theory of Causal Relevance," *Synthese*, XCVII (1993): 335-64; "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies*, LXXVIII (1995): 257-91; "Farewell to Binary Causation," *Canadian Journal of Philosophy*, XXVI (1996): 267-82; "The Role of Contrast in Causal and Explanatory Claims," *Synthese*, CVII (1996): 395-419.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

In the graphical representation of $\mathscr{SE}'$, the arrow from T to B is removed from figure 1. Solving, we have:

$T = 1$; $S = 0$; $B = 0$; $V = 0$

So if the bullet had not been in flight, Trainee still would have shot, Supervisor would not have fired, and Victim would not have died—just as required by the standard solution. This example illustrates how counterfactuals do not backtrack within the structural equations approach.

A system of structural equations is an elegant means for representing a whole family of counterfactuals of just the sort that Lewis's counterfactual theory of causation depends upon. The correctness of a set of structural equations, and of the corresponding graph, depends upon the truth of these counterfactuals. If, as Lewis believes, the truth values of counterfactuals supervene upon noncausal facts, then the correctness of a set of structural equations does as well.

IV. CAUSAL ROUTES

Consider the causal structure depicted in figure 2. What does the arrow from X to Z mean in such a diagram? This is a fundamental question, whose answer serves to highlight an important difference between the structural equations approach to causation, and the counterfactual approach more familiar to philosophers.

Note first the arrows from X to Y and from Y to Z. These indicate that the value of Z depends counterfactually upon the value of Y, which in turn depends upon the value of X. So far, so good. When these relationships hold, the traditional counterfactual approach to causation leads us to ask two further questions: Does the value of Z depend counterfactually upon the value of X? If not, is the value of X nonetheless a cause of the value of Z? But the arrow from X to Z in figure 2 does not correspond to an affirmative answer to either of these questions. Rather, the arrow from X to Z means that the structural equation for Z is of the form $Z = f_Z(X, Y)$, where X is an

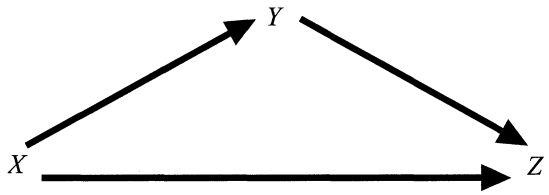

Figure 2

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

essential argument: there exist x, x', y, such that $f_Z(x, y) \neq f_Z(x', y)$. An arrow from X to Z thus means that the value of Z can depend counterfactually upon the value of X, *even holding fixed the value of* Y. The natural causal interpretation of this counterfactual is that the value of X can have an effect on the value of Z over and above the effect it has in virtue of causing the value of Y. There are two routes whereby X influences Z; one which runs through Y, and one direct route which bypasses Y. The overall effect of X on Z will depend upon both of these routes.

Pay attention to the counterfactual that reveals the direct route from X to Z in figure 2: it requires that we hypothetically change the value of X, while holding the value of Y fixed, and evaluate whether Z changes in response. This is an ENF counterfactual: it holds fixed Y, an effect of X, in supposing X to take on a different value; it does not allow the counterfactual to foretrack from X to Y.

"Traditional" counterfactual approaches to causation—those in the tradition of Lewis—do not employ ENF counterfactuals. They employ only counterfactuals with simple antecedents, antecedents that make stipulations about only one event. As a result, they are unable to detect the direct route from X to Z in figure 2. Indeed, despite the extensive use of neuron diagrams by philosophers working within the counterfactual tradition, neither I nor anyone I have polled has seen a diagram in which three neurons are configured in the triangular pattern shown in figure 2!

Whether a route is direct or not is relative to the variable set $\mathcal{X}$. It may be, for instance, that relative to the variable set $\mathcal{X}$, this route is direct; while in the richer variable set $\mathcal{X}'$, the route from X to Z that bypasses Y is mediated by some further variable W in $\mathcal{X}' \setminus \mathcal{X}$. If it is possible to interpolate a variable W along the direct route from X to Z, the traditional counterfactual approach *is* able to detect the existence of two routes from X to Z: one via Y and the other via W. This ability is attested to by the abundance of such "diamond-shaped" configurations in neuron diagrams. Since the traditional counterfactual approach is able to distinguish causal routes *only* by interpolating variables, it is hardly surprising that the standard solution to cases of preemption like "Backup" is to interpolate and find a chain of counterfactual dependence. But even when interpolation is possible and ENF counterfactuals are not needed to detect the *existence* of distinct causal routes, ENF counterfactuals provide crucial information about the *nature* of the distinct causal routes, and this is the key to solving our problem.

Returning to figure 2, note that the arrow from X to Z implies only that there is *some* value y of Y such that $f_Z(x, y) \neq f_Z(x', y)$—such that

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

Z depends counterfactually upon X while fixing Y at y. It may be, however, that given the *actual* value y of Y, Z does not depend upon X. In such a case, we shall say that the direct route from X to Z is *inactive*. Suppose, for example, that all three variables are binary and that $Z = X \wedge Y$. Then, if Y takes the value 1, the direct route from X to Z will be *active*: whether $Z = 0$ or 1 will depend upon whether $X = 0$ or 1 (while holding $Y = 1$ fixed). In such a case, the value of X does in fact play a role in determining the value of Z, over and above the role it plays by determining Y. On the other hand, if $Y = 0$, then the direct route from X to Z is inactive.

Let us formulate this distinction between active and inactive routes more precisely, at the same time generalizing to include cases in which there are more than two routes between two variables, and where the route being evaluated is not direct. Let $\mathscr{E}$ be a system of structural equations on the variables in set $\mathscr{V}$. A *route* between two variables X and Z in $\mathscr{V}$ is an ordered sequence of variables $\langle X, Y_1, \ldots, Y_n, Z\rangle$ such that each variable in the sequence is in $\mathscr{V}$, and is a parent of its successor in the sequence. Graphically, a route between X and Z is a directed path from X to Z. A variable Y, distinct from both X and Z, is *intermediate between X and Z* if and only if it belongs to some route between X and Z. Then:

> "Act": The route $\langle X, Y_1, \ldots, Y_n, Z\rangle$ is *active* in the causal model $\langle \mathscr{V}, \mathscr{E}\rangle$ if and only if Z depends counterfactually upon X within the new system of equations $\mathscr{E}'$ constructed from $\mathscr{E}$ as follows: for all $Y \in \mathscr{V}$, if Y is intermediate between X and Z, but does not belong to the route $\langle X, Y_1, \ldots, Y_n, Z\rangle$, then replace the equation for Y with a new equation that sets Y equal to its actual value in $\mathscr{E}$. (If there are no intermediate variables that do not belong to this route, then $\mathscr{E}'$ is just $\mathscr{E}$.)

The activity of a route is entailed by the truth of a certain kind of ENF counterfactual. That is, the route is active if there is a true counterfactual of the form: if the value of X had been x', and the value of variables that lie along *other* routes from X to Z were held fixed, then the value of Z would have been different. By holding fixed intermediates along other routes, any influence of X on Z along those other routes is eliminated—hypothetical changes in X are not allowed to foretrack along those routes. The relevant counterfactual thus isolates the influence of X on Z along the route in question.

In analogy with Lewis's definition of influence, we could emend "Act" to require that there be "substantial" counterfactual dependence of Z on X in the new system of equations. This would allow us to deny that a route is active, for example, if X is multivalued and very few values of X lead to a different value of Z, or if Z is multivalued, and

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

changes in the value of X lead only to minimal changes in the value of Z. Such an emendation would sacrifice precision, but better fit the coarse grain of our ordinary causal judgments. I shall leave this choice to the reader; the emendation will not be necessary for any of our central examples.

My central proposal is:

> Let c and e be distinct occurrent events, and let X and Z be variables such that the values of X and Z represent alterations of c and e respectively. Then c is a cause of e if and only if there is an active causal route from X to Z in an appropriate causal model $\langle \mathscr{F}, \mathscr{E} \rangle$.[27]

What makes a causal model $\langle \mathscr{F}, \mathscr{E} \rangle$ appropriate? There are at least three requirements. The first two are objective: the equations in $\mathscr{E}$ must entail no false counterfactuals, and they must not represent counterfactual dependence relations between events that are not distinct. The third component is pragmatic: $\mathscr{F}$ should not contain variables whose values correspond to possibilities that we consider to be too remote. I shall discuss these issues at greater length in sections VII and VIII below. For now, note that these restrictions are present in Lewis's recent account and hence are not peculiar to the present account.

In sections V, VII, and VIII below, I shall return to our three central examples. I shall show that in "Backup," there is a causally active route from Trainee's shot to Victim's death. This gives us the result that Trainee's shot causes Victim's death without the need to invoke transitivity. This opens to door to the possibility of accepting "Boulder" and "Dog Bite" as counterexamples to the transitivity of causation, which I shall do. I shall show that there is no active causal route from the dog bite to the explosion, or from the boulder's fall to Hiker's survival.

V. "BACKUP" REVISITED

Let us represent the causal structure of "Backup," this time omitting the esoteric variable *B*. I claim as a virtue for my account that consideration of this variable is unnecessary (albeit harmless). The causal graph is depicted in figure 3, and the set of structural equations is:

[27] In "A Tale of Two Effects" (forthcoming), I take a slightly different line. Sometimes, our judgments of causation attach to individual active routes; at other times, we focus on the net effect of one variable on another along all routes taken together (and indeed there may even be cases where we attend to the effect along some intermediate number of routes). Nonetheless, it is the notion of dependence along a route captured by "Act" that seems to be at work in our judgments of token causation in the cases at issue here.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

$\mathscr{BH}$: $T = 1$; $S = \neg T$; $V = T \vee S$

The interpretation of the variables is the same as that given in section III above. Figure 3 clearly shows two distinct routes from T to V, and it follows from the equations $\mathscr{BH}$ that these two routes "cancel": V does not depend counterfactually upon T.

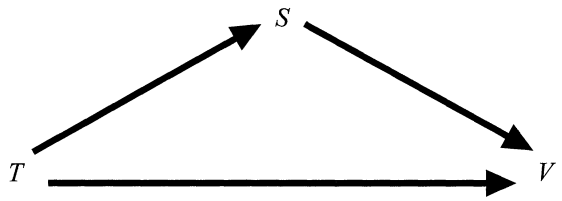

Figure 3

In order to see that the route $\langle T, V \rangle$ is active, we must show that V depends counterfactually upon T while holding the value of S fixed. In fact, $T = 1$, $S = 0$, and $V = 1$. So we must show that if T had been 0, and S had also been 0, then V would have been 0. We must solve for the new structural equations:

$\mathscr{BH}'$: $T = 0$; $S = 0$; $V = T \vee S$

It is easy to see that within this new system of equations, V must equal 0.

We have just demonstrated the truth of the following ENF counterfactual: if Trainee had not shot, and Supervisor had not shot either, then Victim would not have died. The truth of this counterfactual is straightforward, and could have been evaluated intuitively without the use of formal apparatus. It is in virtue of the truth of this counterfactual that Trainee's shot is a cause of Victim's death. No wonder, then, that we have such an easy time judging that Trainee's shot is a cause of Victim's death; this treatment of "Backup" is much simpler than that required by the standard solution. And for those[28] who worry about such other-worldly possibilities, the present account (unlike Lewis's) would work equally well if the assassins' guns did not

[28] For example, McDermott ("Redundant Causation") and Schaffer ("Trumping Preemption") discuss cases involving magic spells that act at a spatiotemporal distance.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

kill by firing bullets, but rather by some sort of unmediated action at a distance.

Note that in order to know that S must be held fixed, we do *not* need to know that Trainee's shot *caused* Supervisor not to shoot. We only need to know that the pattern of counterfactual dependence is as described by $\mathscr{BU}$. Thus the need to hold fixed an event that is intermediate between Trainee's shot and Victim's death does not undermine the reduction of causation to patterns of counterfactual dependence.

VI. FURTHER PROBLEMS REVISITED

Before examining the two counterexamples to transitivity, I wish to return briefly to the problems discussed in section II. It should be clear that Lewis's solution to the problems of late cutting and trumping preemption are readily adapted to the present framework. We could let the variable S take on different values representing different alterations of Suzy's throw; likewise, we could let values of the variable B represent differences in the time and manner of the bottle's shattering. Then the value of B will depend counterfactually upon the value of S, and the route from S to B will be active.

If the values of two variables X and Y symmetrically overdetermine the value of Z (for example, if all the variables are binary and $Z = X \vee Y$) then "Act" does not indicate an active route from either X or Y to Z. To the extent that this is a shortcoming of "Act," it is a shortcoming of Lewis's theory as well. Thus, these three further problems do not accord Lewis's theory any advantage over that presented here.

It is worth briefly noting that the structural equations approach suggests novel solutions to the problems of late cutting preemption and symmetric overdetermination. Detailed accounts are provided by Halpern and Pearl.[29] Their formulations are slightly different from mine, but their proposals are readily adapted to the account sketched above. Consider first the case of late cutting preemption. Even without considering fine differences in the way Suzy throws her rock, or in the way the bottle shatters, it can be shown that there is an active route from Suzy's throw to the bottle's shattering. This is revealed by the following ENF counterfactual: *given* that Billy's rock did not hit the bottle, if Suzy had not thrown, the bottle would have remained intact throughout the incident.

The problem of symmetric overdetermination can be handled by weakening "Act." For reasons of space, I will provide only a compact

[29] In Pearl, *Causality*; and Halpern and Pearl, "Causes and Explanations."

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

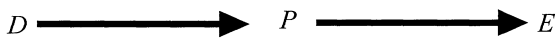

Figure 4

presentation of the technical details. Let $\langle X, Y_1,..., Y_n, Z\rangle$ be a route between the variables X and Z. Suppose that the actual values of the variables on the route (other than X) are: $Y_1 = y_1,..., Y_n = y_n$, $Z = z$. Let $\{W_1,..., W_m\}$ be a set of variables in $\mathscr{V}$ that does not belong to the route $\langle X, Y_1,..., Y_n, Z\rangle$; these variables need not be intermediate between X and Z. Let $w_1,..., w_m$ be possible values of the variables $W_1,..., W_m$, respectively. The values $w_1,..., w_m$ lie in the *redundancy range* of the variables $W_1,..., W_m$ for the route in question if the following counterfactual is true: if it were the case that $W_1 = w_1,..., W_m = w_m$, then it would be the case that $Y_1 = y_1,..., Y_n = y_n$, $Z = z$. In other words, we can set the values of the variables W_i to w_i without disturbing the variables (other than possibly X) on the route $\langle X, Y_1,..., Y_n, Z\rangle$. It should be obvious that the actual values of $W_1,..., W_m$ will lie in the redundancy range, but there may be nonactual values of these variables that do as well. Indeed, cases of symmetric overdetermination arise precisely because nonactual values of these variables lie in the redundancy range. We may now provide our generalization of "Act":

> *WA*: The route $\langle X, Y_1,..., Y_n, Z\rangle$ is *weakly active* relative to $\langle \mathscr{V}, \mathscr{E}\rangle$ if and only if there exists a set (possibly empty) of variables $\{W_1,..., W_m\}$ in $\mathscr{V} \setminus \{X, Y_1,..., Y_n, Z\}$, and values $w_1,..., w_m$ that lie within the redundancy range of these variables for this route, such that Z depends counterfactually upon X within the new system of equations $\mathscr{E}'$ constructed from $\mathscr{E}$ as follows: for each W_i, replace the equation for W_i with a new equation that sets W_i equal to w_i.[30]

"Act" characterizes a special case where each W_i lies on a route from X to Z and is set to its actual value. The reader may verify that *WA* is met in cases of symmetric overdetermination. I fully grant that *WA* is less intuitive and less well-motivated than "Act." Since none of our central examples involves symmetric overdetermination, I shall continue to use "Act" in what follows.[31]

VII. "DOG BITE" REVISITED

"Dog Bite" is represented graphically in figure 4. The interpretation of the variables is as follows: $E = 0$ or 1 depending on whether the explosion occurs; $P = 0$, 1, or 2, depending on whether Terrorist does

[30] Note the similarity to the definition of *actual cause* in Pearl and Halpern.
[31] The reader may verify that *WA* is not met in either "Dog Bite" or "Boulder."

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

not push the detonator at noon, pushes it with his right hand, or pushes it with his left hand (respectively); and $D = 0$ or 1 depending whether the dog bites his right hand or not, shortly before noon. Note that the variable P is not binary: it does not merely represent whether he pushes the detonator with his left hand or not, but represents whether he pushes the detonator at all, and if so with which hand. The structural equations for this system are:

$\mathscr{DB}$: $D = 1$; $P = 1 + D$; $E = P \wedge 1$

In words: the dog does bite Terrorist's right hand shortly before noon; Terrorist will push the detonator with his right hand at noon unless the dog bites his right hand, in which case he will push the detonator with his left hand; and the bomb will explode just in case he pushes the detonator with his right or left hand at noon. In fact, the values of these variables are $D = 1$, $P = 2$, $E = 1$. Suppose the dog had not bitten his hand. Then we set $D = 0$, yielding the values $P = 1$ and $E = 1$. That is, if the dog had not bitten his right hand, Terrorist would have pushed the detonator with his right hand at noon and the bomb would still have exploded. This system of equations yields the intuitively correct result that the explosion does not counterfactually depend upon the dog bite.

The dog bite is not a cause of the explosion: there is only one route from D to E, namely, $\langle D, P, E \rangle$ and that route is not causally active. In disanalogy with "Backup," the failure of E to counterfactually depend upon D is not due to cancellation along different routes, but rather to what we might call a *failure of composition*. While the function f_P from the variable D to P is nontrivial, and the function f_E from P to E is nontrivial, the composite function $f_E \circ f_P$ from D to E is trivial.

L. A. Paul[32] argues that "Dog Bite" (or rather a case structurally identical to it) does not constitute a violation of the transitivity of causation. She maintains that the effect of the dog bite is different from the cause of the explosion; one is a push qua thing done with left hand, the other a push qua push of detonator button. I take this to be a proposal about how to use causal language to describe the pattern of dependence captured by $\mathscr{DB}$. Perhaps it is possible to regiment our causal language in such a way that failures of composition need not be described by triples of the form: 'a causes b', 'b causes c', and 'a does not cause c'. But since the transitivity of causation is not a principle of which we have independent need (section v), and since there are counterexamples to the transitivity of

[32] "Aspect Causation," this JOURNAL, XCVII, 4 (April 2000): 235-56.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

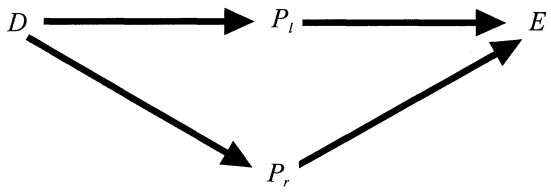

Figure 5

causation which do not rest on failures of composition (section VIII), the desire to preserve the transitivity of causation provides a poor motive for so regimenting our language.

We have shown that there is no active path from D to E in the model $\langle\{D, P, E\}, \mathscr{DB}\rangle$. Perhaps there is another choice of $\langle \mathscr{V}, \mathscr{E} \rangle$ that is equally appropriate? Suppose, for example, that we represent the situation using the structure depicted in figure 5. Here $P_l = 0$ or 1 depending upon whether Terrorist pushes the button with his left hand at noon, and $P_r = 0$ or 1 depending upon whether Terrorist pushes the button with his right hand at noon. The corresponding set of structural equations would be:

$\mathscr{DB}'$: $D = 1$; $P_l = D$; $P_r = \neg D$; $E = P_l \vee P_r$

This structure gives us two routes from D to E. Holding P_r fixed at its actual value of 0, E *does* counterfactually depend upon D. This implies that there is an active route, $\langle D, P_l, E\rangle$, from D to E.

What makes this an inappropriate model for "Dog Bite"?[33] One minimum criterion of adequacy for a model is that it entail only true counterfactuals. As it turns out, the very counterfactual that reveals the active route $\langle D, P_l, E\rangle$ in this model is false, or at best indeterminate. That ENF counterfactual is: given that Terrorist did not push the detonator button with his right hand, if the dog had not bitten Terrorist's right hand, then (he would not have pushed the button with his left hand either, and) the bomb would not have exploded.[34] Given that he did not push the button with his right hand, why should the dog bite make any difference to whether he pushes it with the left? He wanted the bomb to explode, would he not push the button with his left hand regardless of whether the dog bit his right? At any

[33] The following discussion has benefited especially from comments by Hausman and Sober.

[34] Note that $\mathscr{DB}$ does not entail that this counterfactual is true or that it is false. Substituting $P \neq 1$ does not yield a determinate solution for either P or E.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

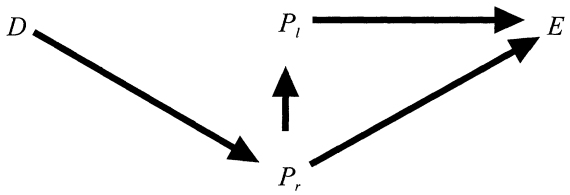

Figure 6

rate, it seems wrong to say that he would definitely *not* have pushed the button with his left hand.

The problem with figure 5 and $\mathscr{DB}'$ is that they represent the dog bite as having two distinct effects: it prevented Terrorist from pushing the button with his right hand, and caused him to push it with his left. These effects are not genuinely distinct: the dog bite caused Terrorist to push the button with his left hand only insofar as it prevented him from pushing it with his right. This intuitive distinction is made precise using ENF counterfactuals: Terrorist's pushing the button with his left hand ceases to depend counterfactually upon the dog bite when we specify whether or not he pushed the button with his right hand.

We might alternately represent "Dog Bite" by $\mathscr{DB}''$ and figure 6:

$\mathscr{DB}''$: $D = 1$; $P_r = \neg D$; $P_l = \neg P_r$; $E = P_l \vee P_r$

This captures the idea that the dog bite causes Terrorist to push the button with his left hand *by* preventing him from pushing with his right.

Figure 6 and $\mathscr{DB}''$ do not entail the troublesome ENF counterfactuals discussed above, but they are inappropriate for a different reason. It was specified in the example that Terrorist had to press the detonator *at noon* in order for the bomb to explode. Accordingly, the variables P_l and P_r must represent whether or not Terrorist presses the button with the relevant hand *at noon*, or else the equation for E will be false. It is thus inappropriate to think of these variables as standing in a *causal* relationship. Terrorist pushed the button at noon with his left hand rather than his right, but his failure to push the button with his right hand did not *cause* him to push the button with his left hand. The two events (or rather, the one omission and the one event) fail to be distinct in the way required for the one to be a cause of the other. Invoking the terminology of Lewis's new theory, pushing with the left hand and pushing with the right are alterations

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

of the same event. We capture this idea by representing the two as distinct values *of a single variable,* as we did in figure 4 and $\mathscr{DB}$.

Figure 6 and $\mathscr{DB}''$ would be an appropriate representation for a slightly different case. Suppose that there is a window of time during which Terrorist can push the button. Suppose, moreover, that his initial instinct is to push the button with his right hand, and only after he finds himself unable to do so will he push the button with his left. In this variant of the story, Terrorist's failure to push the button with his right hand does indeed cause him to push the button with his left hand a moment later. In this version of the example, we still have the intuition that the dog bite did not cause the explosion, and I owe an account of this. It turns out that this structure (or rather a crucial part of it) is isomorphic to that of "Boulder," which I discuss in the next section.

Consider one final alternative representation. Suppose we replace D with a three-valued variable: the dog might do nothing, bite Terrorist's right hand, or maul Terrorist severely. If the dog mauls Terrorist, he will not push the button (being totally incapacitated) and the bomb will not explode. Now the route $\langle D, P, E\rangle$ in figure 4 will be active: whether the bomb explodes will depend upon what the dog does. But this is as it should be: if we are willing to take seriously the possibility that the dog might have mauled Terrorist severely, we might well judge that the dog's (merely) biting his right hand (rather than mauling him) is a cause of the explosion.[35] Assume, however, that this is not a possibility that we wish to take seriously (did *you* consider this possibility before I mentioned it?): then this new model is inappropriate for pragmatic reasons. This mirrors Lewis's claim that in order for the event c to influence e, the alterations of c that yield different alterations of e must be "not-too-distant."

Of course, I cannot rule out every rival model of "Dog Bite." I maintain, however, that of those which spring to mind, only the model captured by figure 4 and $\mathscr{DB}$ is appropriate. In this model, there is no active route from the dog bite to the explosion. This captures our intuitive judgment that the dog bite is not a cause of the explosion. In the course of discussing this example, we have illustrated three different ways in which a model might fail to be appropriate: it might license counterfactuals that are not true; it might posit causal connections between events that are not distinct; or it

[35] This is the sort of case where the contrastive approach that I have developed elsewhere (see the references in footnote 26 above) is helpful.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

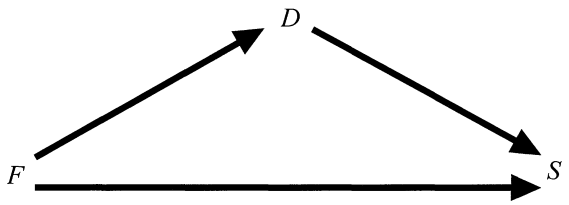

Figure 7

might encompass possibilities that are too distant from actuality to take seriously.

VIII. "BOULDER" REVISITED

Let us set up a model for "Boulder" (figure 7). The variable $F = 0$ or 1 depending upon whether the boulder falls; $D = 0$ or 1 depending upon whether Hiker ducks; and $S = 0$ or 1 depending upon whether Hiker survives. The structural equations are:

$$\mathscr{BC}\text{: } F = 1;\ D = F;\ S = \neg F \vee D$$

(To see the isomorphism with $\mathscr{DB}''$, identify P_r with $\neg F$, P_l with D, and E with S.)

In this model, there are no active routes from F to S. It is relatively easy to see that the direct route $\langle F, S\rangle$ is inactive. Holding D fixed at its actual value of 1, there is no counterfactual dependence of S on F: holding fixed that Hiker ducked, he would have survived if the boulder had not fallen. The more important issue, however, is whether the indirect route $\langle F, D, S\rangle$ is active. It is along this route that we have a chain of counterfactual dependence, and hence causation on Lewis's account. There are no intermediates between F and S that do not lie along this route, and hence nothing to hold fixed in evaluating whether this route is active. Since S does not depend on F, the route is inactive. Intuitively, the falling boulder does not save Hiker's life because without it, Hiker's life would not have been endangered in the first place. This is just what is indicated by the absence of an active route from F to S: there is no scenario in which the boulder does not fall and Hiker does not survive.

Note that while the causal diagram for "Boulder" is isomorphic to that in "Backup" (figure 3), the structural equations are not isomorphic. Thus "Backup," "Dog Bite," and "Boulder" all exhibit different causal structures: counterfactual dependence fails for different rea-

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

sons in each case. By contrast, Lewis[36] claims that all of the (putative) counterexamples to transitivity have a common structure. This is not surprising: it is only by considering ENF counterfactuals that the difference in structure is revealed. If we consider only counterfactuals with single antecedents, as Lewis does, then all three cases have the same structure: there is a chain of counterfactual dependence, but no dependence of the last event on the first.

There is a sense in which the inactivity of the route $\langle F, D, S \rangle$ rests on a loophole in the definition of an active route. The guiding idea behind that definition was to isolate the influence of one variable on another along a given route by factoring out any counteracting influences along other routes. This was accomplished by holding fixed the values of intermediate variables along those other routes. This technique will not work, however, when one of the other routes in question is direct: a direct route has no intermediate variables, so its role cannot be factored out. I maintain that this is no weakness in the definition; rather, it is the nature of direct routes that they are too intimately bound up with other routes to be separated from them. Any effect the falling boulder may have made on Hiker's survival by causing him to duck cannot be conceptually severed from the risk that the falling boulder posed to him in the first place. Thus "Boulder" is analogous to "Backup" in that S fails to depend counterfactually upon F because of cancellation along two different routes; but "Boulder" is unlike "Backup" in that an active route cannot be isolated.

Is it possible to interpolate a variable along the route $\langle F, S \rangle$? If so, then it may be possible to hold *that* variable fixed, and thus isolate the route $\langle F, D, S \rangle$ and show it to be active. This can be done, but it turns out to be much more difficult than it might first appear. The trick is to find a variable that does not also lie along the route $\langle F, D, S \rangle$. The possibility of doing this rests upon the contingent fact that it takes Hiker a finite amount of time to react to the boulder and get into a safe position. There will be a point on the boulder's trajectory—let us say one meter from Hiker's head—such that by the time the boulder reaches that point, it is too late for Hiker to duck if he has not done so already. Let $B = 1$ represent the boulder's presence at this point, and $B = 0$ its absence. Let us add the clarifying remark that $D = 0$ or 1 according to whether Hiker ducks *at the appropriate time*—sufficiently early to avoid being hit by the boulder. The value of D tells us nothing about whether Hiker attempts (too late) to duck in response

[36] "Causation as Influence," pp. 194-95.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

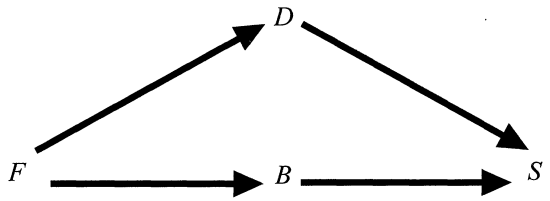

Figure 8

to the boulder when it is only one meter from his head. Note how we have carefully chosen B so that neither B nor D affect each other. Here is the new set of structural equations:

$$\mathscr{BC}'\text{: } F = 1;\ D = F;\ B = F;\ S = \neg B \vee D$$

The corresponding causal graph is shown in figure 8. In fact, B took the value 1. Holding this fixed, we must determine whether Hiker's survival depends counterfactually upon the boulder's fall. Setting $F = 0$ and $B = 1$, we get $D = 0$, and thus $S = 0$. Relative to the system of structural equations $\mathscr{BC}'$, the route $\langle F, D, S\rangle$ is indeed active.

Must we then conclude that the boulder's fall did save Hiker's life after all? Perhaps this would not be so bad. On my account, unlike Lewis's, it is at least easy to explain why our pretheoretic intuition yields the "wrong" answer in this case. The interpolated variable B is not easy to find, and the ENF counterfactual that reveals the active causal route from F to S is not at all intuitive. In words, the relevant piece of counterfactual reasoning would go as follows: suppose that the boulder had been present at a point one meter from Hiker's head and flying toward him, and suppose moreover that it had never fallen in the first place. Since it never fell, Hiker would not have seen it coming and would not have ducked; since it would have been there, one meter from his exposed head, it would have hit him and he would not have survived. This counterfactual reasoning is correct, but bizarre. If the boulder never fell, how did it get to be there, one meter from Hiker's head? We are to imagine, presumably, that the boulder was mysteriously and instantaneously transported to a position immediately in front of Hiker's head. This is the sort of counterfactual reasoning that only trained philosophers engage in; unaided intuition is not to be faulted for failing to "see" the relevant ENF counterfactual.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

I believe, however, that it is possible to give a less apologetic response.[37] Figure 8 and $\mathscr{BC}'$, while more complete than figure 7 and $\mathscr{BC}$, do not constitute an appropriate representation of "Boulder." As we have noted, our causal judgments depend, in part, upon which unactualized possibilities we are willing to take seriously, and which we consider too remote. The variables we choose to include in a causal model should reflect these concerns. We included the variables *F* and *D* in our original model of "Boulder." This choice reflects our willingness to take seriously the possibility that the boulder does not fall, and the possibility that Hiker does not duck. Moreover, it reflects our willingness to take seriously the possibility Hiker does not duck *even though the boulder falls*. That is, it reflects our willingness to view Hiker's life as being at risk. Why are we willing take this possibility seriously, even though it is stipulated that Hiker is determined to duck if the boulder falls? Plausibly, it is because we recognize the nomic connection between the boulder's fall and Hiker's duck to be highly contingent upon unspecified details of the case. Hiker might not have ducked if he had been looking in the wrong direction, if his reactions were slowed by tired muscles, if he were less prone to reacting coolly in times of crisis, and so on. It is not necessary, however, that we have any one of these explanations firmly in mind in order to take seriously the possibility that Hiker does not duck when the boulder falls.

When we exclude the variable *B* from our model, it is not because we are unwilling to take seriously the possibility that the boulder was not present at that point (one meter from Hiker's head). We take that possibility seriously when we entertain the possibility that the boulder does not fall in the first place. Rather, we are not willing to take seriously the possibility that the boulder (or *a* boulder of similar size and shape) comes to be in that position *even though the boulder does not fall in the first place*. This possibility is just too far-fetched. (Did *you* consider this possibility before I mentioned it?) Perhaps one could tell a story that would lead us to take this possibility seriously—perhaps Hiker has inadvertently walked in front of a boulder launcher that is carefully camouflaged against the hillside.[38] But in just such a case, we should take the original causal claim seriously: by causing Hiker to duck in plenty of time, the fall of the boulder down the hillside does indeed save Hiker's life.

[37] The following paragraphs owe a great deal to discussion with Pearl.

[38] To get the causal structure to come out right, more details would be needed. In particular, it would have to be the case that on this occasion, the launching of a boulder is triggered by the original boulder falling down the hillside.

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms

IX. CONCLUSION

If one subscribes to the thesis that causation is transitive, there are benefits: one can explain how there can be causation without counterfactual dependence in cases of preemption such as "Backup." But there are costs as well: the defender of transitivity is committed to unintuitive causal claims in cases like "Dog Bite" and "Boulder." The benefits are not worth the costs: they can be had cost-free. The use of ENF counterfactuals allows us to accommodate causation without counterfactual dependence in cases of preemption, without committing us to the counterintuitive consequences of transitivity. The use of ENF counterfactuals is independently motivated by the techniques employed in causal modeling.

In addition to reproducing our intuitions in standard test cases, the approach recommended here has a number of further advantages. It reproduces our intuitive causal judgments in a very natural way: the counterfactuals appealed to are psychologically natural counterfactuals rather than philosophically technical ones. Moreover, it reproduces our causal judgments without introducing events or variables beyond those explicitly presented in the various scenarios (such as intermediate stages in the trajectory of the bullet in "Backup" or the boulder in "Boulder"). In short, it reproduces our intuitive judgments using only those resources that are available to unaided intuition. The approach recommended here allows for an elegant representation of patterns of counterfactual dependence. We can explicitly represent the structural differences between superficially similar cases. These representations also allow us to make explicit some of our underlying assumptions about a particular case (such as that certain possibilities are too far-fetched to be taken seriously).

Perhaps there is still room for the defender of transitivity to maneuver. But when there is an alternative account with such striking advantages, why bother?

CHRISTOPHER HITCHCOCK

California Institute of Technology

This content downloaded from
192.58.125.1 on Sun, 08 Jan 2023 23:57:48 UTC
All use subject to https://about.jstor.org/terms