CHAPTER TEN

The Actual Cause

And now remains
That we find out the cause of this effect,
Or rather say, the cause of this defect,
For this effect defective comes by cause.
Shakespeare (Hamlet II.ii. 100–4)

Preface

This chapter offers a formal explication of the notion of "actual cause," an event recognized as responsible for the production of a given outcome in a specific scenario, as in: "Socrates drinking hemlock was the actual cause of Socrates death." Human intuition is extremely keen in detecting and ascertaining this type of causation and hence is considered the key to constructing explanations (Section 7.2.3) and the ultimate criterion (known as "cause in fact") for determining legal responsibility.

Yet despite its ubiquity in natural thoughts, actual causation is not an easy concept to formulate. A typical example (introduced by Wright 1988) considers two fires advancing toward a house. If fire *A* burned the house before fire *B*, we (and many juries nationwide) would surely consider fire *A* "the actual cause" of the damage, though either fire alone is sufficient (and neither one was necessary) for burning the house. Clearly, actual causation requires information beyond that of necessity and sufficiency; the actual process mediating between the cause and the effect must enter into consideration. But what precisely is a "process" in the language of structural models? What aspects of causal processes define actual causation? How do we piece together evidence about the uncertain aspects of a scenario and so compute probabilities of actual causation?

In this chapter we propose a plausible account of actual causation that can be formulated in structural model semantics. The account is based on the notion of *sustenance,* to be defined in Section 10.2, which combines aspects of necessity and sufficiency to measure the capacity of the cause to maintain the effect despite certain *structural* changes in the model. We show by examples how this account avoids problems associated with the counterfactual dependence account of Lewis (1986) and how it can be used both in generating explanations of specific scenarios and in computing the probabilities that such explanations are in fact correct.

10.1 INTRODUCTION: THE INSUFFICIENCY OF NECESSARY CAUSATION

10.1.1 Singular Causes Revisited

Statements of the type "a car accident was the cause of Joe's death," made relative to a specific scenario, are classified as "singular," "single-event," or "token-level" causal

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

statements. Statements of the type "car accidents cause deaths," when made relative to a type of events or a class of individuals, are classified as "generic" or "type-level" causal claims (see Section 7.5.4). We will call the cause in a single-event statement an *actual cause* and the one in a type-level statement a *general cause*.

The relationship between type and token causal claims has been controversial in the philosophical literature (Woodward 1990; Hitchcock 1995), and priority questions such as "which comes first?" and "can one level be reduced to the other?" (Cartwright 1989; Eells 1991; Hausman 1998) have diverted attention from the more fundamental question: "What tangible claims do type and token statements make about our world, and how is causal knowledge organized so as to substantiate such claims?" The debate has led to theories that view type and token claims as two distinct species of causal relations (as in Good 1961, 1962), each requiring its own philosophical account (see, e.g., Sober 1985; Eells 1991, chap. 6) – "not an altogether happy predicament" (Hitchcock 1997). In contrast, the structural account treats type and token claims as instances of the same species, differing only in the details of the scenario-specific information that is brought to bear on the question. As such, the structural account offers a formal basis for studying the anatomy of the two levels of claims, what information is needed to support each level, and why philosophers have found their relationships so hard to disentangle.

The basic building blocks of the structural account are the functions $\{f_i\}$, which represent lawlike mechanisms and supply information for both type-level and token-level claims. These functions are type-level in the sense of representing generic, counterfactual relationships among variables that are applicable to every hypothetical scenario, not just ones that were realized. At the same time, any specific instantiation of those relationships represents a token-level claim. The ingredients that distinguish one scenario from another are represented in the background variables U. When all such factors are known, $U = u$, we have a "world" on our hands (Definition 7.1.8) – an ideal, full description of a specific scenario in which all relevant details are spelled out and nothing is left to chance or guessing. Causal claims made at the world level would be extreme cases of token causal claims. In general, however, we do not possess the detailed knowledge necessary for specifying a single world $U = u$, and we use a probability $P(u)$ to summarize our ignorance of those details. This takes us to the level of probabilistic causal models $\langle M, P(u)\rangle$ (Definition 7.1.6). Causal claims made on the basis of such models, with no reference to the actual scenario, would be classified as type-level claims. Causal effects assertions, such as $P(Y_x = y) = p$, are examples of such claims, for they express the general tendency of x to bring about y, as judged over all potential scenarios.[1] In most cases, however, we possess partial information about the scenario at hand – for example, that Joe died, that he was in a car accident, and perhaps that he drove a sports car and suffered a head injury. The totality of such episode-specific information is called "evidence" (e) and can be used to update $P(u)$ into $P(u \mid e)$. Causal claims derived from the model $\langle M, P(u \mid e)\rangle$ represent token claims of varying shades, depending on the specificity of e.

[1] Occasionally, causal effect assertions can even be made on the basis of an incomplete probabilistic model, where only $G(M)$ and $P(v)$ are given – this is the issue of identification (Chapter 3). But no token-level statement can be made on such basis alone without some knowledge of $\{f_i\}$ or $P(u)$ (assuming, of course, that x and y are known to have occurred).

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

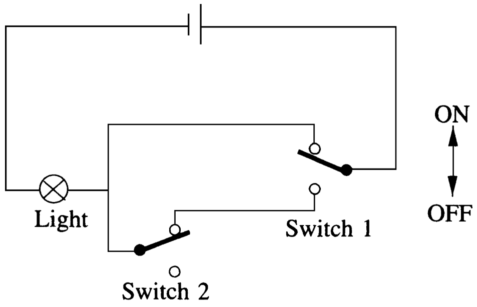

Figure 10.1 Switch 1 (and not switch 2) is perceived to be causing the light, though neither is necessary.

Thus, the distinction between type and token claims is a matter of degree in the structural account. The more episode-specific evidence we gather, the closer we come to the ideals of token claims and actual causes. The notions of PS and PN (the focus of Chapter 9) represent intermediate points along this spectrum. Probable sufficiency (PS) is close to a type-level claim because the actual scenario is not taken into account and is, in fact, excluded from consideration. Probable necessity (PN) makes some reference to the actual scenario, albeit a rudimentary one (i.e., that x and y are true). In this section we will attempt to come closer to the notion of actual cause by taking additional information into consideration.

10.1.2 Preemption and the Role of Structural Information

In Section 9.2, we alluded to the fact that both PN and PS are global (i.e., input–output) features of a causal model, depending only on the function $Y_x(u)$ and not on the structure of the process mediating between the cause (x) and the effect (y). That such structure plays a role in causal explanation is seen in the following example.

Consider an electrical circuit consisting of a light bulb and two switches, as shown in Figure 10.1. From the user's viewpoint, the light responds symmetrically to the two switches; either switch is sufficient to turn the light on. Internally, however, when switch 1 is on it not only activates the light but also disconnects switch 2 from the circuit, rendering it inoperative. Consequently, with both switches on, we would not hesitate to proclaim switch 1 as the "actual cause" of the current flowing in the light bulb, knowing as we do that switch 2 can have no effect whatsoever on the electric pathway in this particular state of affairs. There is nothing in PN and PS that could possibly account for this asymmetry; each is based on the response function $Y_x(u)$ and is therefore oblivious to the internal workings of the circuit.

This example is representative of a class of counterexamples, involving *preemption,* that were brought up against Lewis's counterfactual account of causation. It illustrates how an event (e.g., switch 1 being on) can be considered a cause although the effect persists in its absence. Lewis's (1986) answer to such counterexamples was to modify the counterfactual criterion and let x be a cause of y as long as there exists a *counterfactual dependence chain* of intermediate variables between x to y; that is, the output of every link in the chain is counterfactually dependent on its input. Such a chain does not exist for switch 2 because, given the current state of affairs (i.e., both switches being on), no

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

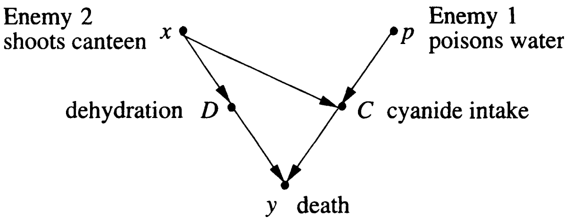

Figure 10.2 Causal relationships in the desert traveler example.

part of the circuit would be affected (electrically) by turning switch 2 on or off. This can be shown more clearly in the following example.

Example 10.1.1 (The Desert Traveler–after P. Suppes) A desert traveler T has two enemies. Enemy 1 poisons T's canteen, and enemy 2, unaware of enemy 1's action, shoots and empties the canteen. A week later, T is found dead and the two enemies confess to action and intention. A jury must decide whose action was the *actual cause* of T's death.

Let x and p be (respectively) the propositions "enemy 2 shot" and "enemy 1 poisoned the water," and let y denote "T is dead." In addition to these events we will also use the intermediate variable C (connoting cyanide) and D (connoting dehydration), as shown in Figure 10.2. The functions $f_i(pa_i, u)$ are not shown explicitly in Figure 10.2, but they are presumed to determine the value of each child variable from those of its parent variables in the graph, in accordance with the usual understanding of the story:[2]

$$\begin{aligned} c &= px', \\ d &= x, \qquad\qquad (10.1) \\ y &= c \vee d. \end{aligned}$$

When we substitute c and d into the expression for y, we obtain a simple disjunction

$$y = x \vee px' \equiv x \vee p, \qquad (10.2)$$

which is deceiving in its symmetry.

Here we see in vivid symbols the role played by structural information. Although it is true that $x \vee x'p$ is logically equivalent to $x \vee p$, the two are not structurally equivalent; $x \vee p$ is completely symmetric relative to exchanging x and p, whereas $x \vee x'p$ tells us that, when x is true, p has no effect whatsoever – not only on y, but also on any of the intermediate conditions that could potentially affect y. It is this asymmetry that makes us proclaim x and not p to be the cause of death.

According to Lewis, the difference between x and p lies in the nature of the chains that connect each of them to y. From x, there exists a causal chain $x \rightarrow d \rightarrow y$ such that every element is counterfactually dependent on its antecedent. Such a chain does not exist from p to y because, when x is true, the chain $p \rightarrow c \rightarrow y$ is *preempted* (at c);

[2] For simplicity, we drop the "$\wedge$" symbol in the rest of this chapter.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

that is, c is "stuck" at false regardless of p. Put another way, although x does not satisfy the counterfactual test for causing y, one of its consequences (d) does; given that x and p are true, y would be false were it not for d.

Lewis's chain criterion retains the connection between causation and counterfactuals, but it is rather ad hoc; after all, why should the existence of a counterfactual dependence chain be taken as a defining test for a concept as crucial as "actual cause," by which we decide the guilt or innocence of defendants in a court of law? The basic counterfactual criterion does embody a pragmatic rationale; we would not wish to punish a person for a damage that could not have been avoided, and we would like to encourage people to watch for circumstances where their actions could make a substantial difference. However, once the counterfactual dependence between the action and the consequence is destroyed by the presence of another cause, what good is it to insist on intermediate counterfactual dependencies along a chain that connects them?

10.1.3 Overdetermination and Quasi-Dependence

Another problem with Lewis's chain is its failure to capture cases of simultaneous disjunctive causes. For example, consider the firing squad in Figure 9.1, and assume that riflemen A and B shot together and killed the prisoner. Our intuition regards each of the riflemen as a *contributory* actual cause of the death, though neither rifleman passes the counterfactual test and neither supports a counterfactual dependence chain in the presence of the other.

This example is representative of a condition called *overdetermination,* which presents a tough challenge to the counterfactual account. Lewis answered this challenge by offering yet another repair of the counterfactual criterion. He proposed that chains of counterfactual dependence should be regarded as intrinsic to the process (e.g., the flight of the bullet from A to D) and that the disappearance of dependence due to peculiar surroundings (e.g., the flight of the bullet from B to D) should not be considered an intrinsic loss of dependence; we should still count such a process as *quasi-dependent* "if only the surroundings were different" (Lewis 1986, p. 206).

Hall (2004) observed that the notion of quasi-dependence raises difficult questions: "First, what exactly is a process? Second, what does it mean to say that one process is 'just like' another process in its intrinsic character? Third, how exactly do we 'measure the variety of the surroundings'?" We will propose an answer to these questions using an object called a *causal beam* (Section 10.3.1), which can be regarded as a structural–semantic explication of the notion of a "process." We will return to chains and beams and to questions of preemption and overdetermination in Section 10.2, after a short excursion into Mackie's approach, which also deals with the problem of actual causation – though from a different perspective.

10.1.4 Mackie's INUS Condition

The problems we encountered in the previous section are typical of many attempts by philosophers to give a satisfactory logical explication to the notion of single-event causation (here, "actual causation"). These attempts seem to have started with Mill's observation that no cause is truly sufficient or necessary for its effect (Mill 1843, p. 398). The

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

numerous accounts subsequently proposed – based on more elaborate combinations of sufficiency and necessity conditions – all suffer from insurmountable difficulties (Sosa and Tooley 1993, pp. 1–8). Mackie's treatment (1965) appears to be the earliest attempt to offer a semiformal explication of "actual causation" within this logical framework; his solution, known as the INUS condition, became extremely popular.

The INUS condition states that an event C is perceived to be the cause of event E if C is "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result" (Mackie 1965).[3] Although attempts to give INUS precise formulation (including some by Mackie 1980) have not resulted in a coherent proposal (Sosa and Tooley 1993, pp. 1–8), the basic idea behind INUS is appealing: If we can think of $\{S_1, S_2, S_3, \ldots\}$ as a collection of every minimally sufficient set of conditions (for E), then event C is an INUS condition for E if it is a conjunct of some S_i. Furthermore, C is considered a *cause* of E if C is an INUS condition for E and if, under the circumstances, C was sufficient for one of those S_i. Thus, for example, if E can be written in disjunctive normal form as

$$E = AB \vee CD,$$

then C is an INUS condition by virtue of being a member of a disjunct, CD, which is minimal and sufficient for E. Thus C would be considered a cause of E if D were present on the occasion in question.[4]

This basic intuition is shared by researchers from many disciplines. Legal scholars, for example, have advocated a relation called NESS (Wright 1988), standing for "necessary element of sufficient set," which is a rephrasing of Mackie's INUS condition in a simpler mnemonic. In epidemiology, Rothman (1976) proposed a similar criterion – dubbed "sufficient components" – for recognizing when an exposure is said to cause a disease: "We say that the exposure E causes disease if a sufficient cause that contains E is the first sufficient cause to be completed" (Rothman and Greenland 1998, p. 53). Hoover (1990, p. 218) related the INUS condition to causality in econometrics: "Any variable that causes another in Simon's sense may be regarded as an INUS condition for that other variable."

However, all these proposals suffer from a basic flaw: the language of logical necessity and sufficiency is inadequate for explicating these intuitions (Kim 1971). Similar conclusions are implicit in the analysis of Cartwright (1989, pp. 25–34), who starts out enchanted with INUS's intuition and ends up having to correct INUS's mistakes.

The basic flaw of the logical account stems from the lack of a syntactic distinction between formulas that represent stable mechanisms (or "dispositional relations," to use Mackie's terminology) and those that represent circumstantial conditions. The simplest manifestation of this limitation can be seen in contraposition: "A implies B"

[3] The two negations and the two "buts" in this acronym make INUS one of the least helpful mnemonics in the philosophical literature. Simplified, it should read: "a necessary element in a sufficient set of conditions, NESS" (Wright 1988).

[4] Mackie (1965) also required that every disjunct of E that does not contain C as a conjunct be absent, but this would render Mackie's definition identical to the counterfactual test of Lewis. I use a broader definition here to allow for simultaneous causes and overdetermination; see Mackie (1980, pp. 43–7).

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

is logically equivalent to "not B implies not A," which renders not-B an INUS-cause of not-A. This is counterintuive; from "disease causes a symptom" we cannot infer that eliminating a symptom will cause the disappearance of the disease. The failure of contraposition further entails problems with transduction (inference through common causes): if a disease D causes two symptoms, A and B, then curing symptom A would entail (in the logical account of INUS) the disappearance of symptom B.

Another set of problems stems from syntax sensitivity. Suppose we apply Mackie's INUS condition to the firing squad story of Figure 9.1. If we write the conditions for the prisoner's death as:

$$D = A \vee B,$$

then A satisfies the INUS criterion, and we can plausibly conclude that A was a cause of D. However, substituting $A = C$, which is explicit in our model, we obtain

$$D = C \vee B,$$

and suddenly A no longer appears as a conjunct in the expression for D. Shall we conclude that A was not a cause of D? We can, of course, avoid this disappearance by forbidding substitutions and insisting that A remain in the disjunction together with B and C. But then a worse problems ensues: in circumstances where the captain gives a signal (C) and both riflemen fail to shoot, the prisoner will still be deemed dead. In short, the structural information conveying the flow of influences in the story cannot be encoded in standard logical syntax – the intuitions of Mackie, Rothman, and Wright must be reformulated.

Finally, let us consider the desert traveler example, where the traveler's death was expressed in (10.2) as

$$y = x \vee x'p.$$

This expression is not in minimal disjunctive normal form because it can be rewritten as

$$y = x \vee p,$$

from which one would obtain the counterintuitive result that x and p are equal partners in causing y. If, on the other hand, we permit nonminimal expressions like $y = x \vee x'p$, then we might as well permit the equivalent expression $y = xp' \vee p$, from which we would absurdly conclude that not poisoning the water (p') would be a cause for our traveler's misfortune, provided someone shoots the canteen (x).

We return now to structural analysis, in which such syntactic problems do not arise. Dispositional information is conveyed through structural or counterfactual expressions (e.g., $v_i = f_i(pa_i, u)$) in which u is generic, whereas circumstantial information is conveyed through prepositional expressions (e.g., $X(u) = x$)) that refer to one specific world $U = u$. Structural models do not permit arbitrary transformations and substitutions, even when truth values are preserved. For example, substituting the expression for c in $y = d \vee c$ would not be permitted if c (cyanide intake) is understood to be governed by a separate mechanism, independent of that which governs y.

Using structural analysis, we will now propose a formal setting that captures the intuitions of Mackie and Lewis. Our analysis will be based on an aspect of causation called

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

sustenance, which combines elements of sufficiency and necessity and also takes structural information into account.

10.2 PRODUCTION, DEPENDENCE, AND SUSTENANCE

The probabilistic concept of causal sufficiency, PS (Definition 9.2.2), suggests a way of rescuing the counterfactual account of causation. Consider again the symmetric overdetermination in the firing-squad example. The shot of each rifleman features a PS value of unity (see (9.43)), because each shot would cause the prisoner's death in a state u' in which the prisoner is alive. This high PS value supports our intuition that each shot is an actual cause of death, despite a low PN value (PN = 0). Thus, it seems plausible to argue that our intuition gives some consideration to sufficiency, and that we could formulate an adequate criterion for actual causation using the right mixture of PN and PS components.

Similar expectations are expressed in Hall (2004). In analyzing problems faced by the counterfactual approach, Hall made the observation that there are two concepts of causation, only one of which is captured by the counterfactual account, and that failure to capture the second concept may well explain its clashes with intuition. Hall calls the first concept "dependence" and the second "production." In the firing-squad example, intuition considers each shot to be an equal "producer" of death. In contrast, the counterfactual account tests for "dependence" only, and it fails because the state of the prisoner does not "depend" on either shot alone.

The notions of dependence and production closely parallel those of necessity and sufficiency, respectively. Thus, our formulation of PS could well provide the formal basis for Hall's notion of production and serve as a step toward the formalization of actual causation. However, for this program to succeed, a basic hurdle must first be overcome: productive causation is oblivious to scenario-specific information (Pearl 1999), as can be seen from the following considerations.

The ***dependence*** aspect of causation appeals to the necessity of a cause x in maintaining the effect y in the face of certain contingencies, which otherwise will negate y (Definition 9.2.1):

$$X(u) = x, \quad Y(u) = y, \quad Y_{x'}(u) = y'. \tag{10.3}$$

The ***production*** aspect, on the other hand, appeals to the capacity of a cause (x) to bring about the effect (y) in a situation (u') where both are absent (Definition 9.2.2):

$$X(u') = x', \quad Y(u') = y', \quad Y_x(u') = y. \tag{10.4}$$

Comparing these two definitions, we note a peculiar feature of production: To test production, we must step outside our world momentarily, imagine a new world u' with x and y absent, apply x, and see if y sets in. Therefore, the sentence "x produced y" can be true only in worlds u' where x and y are false, and thus it appears (a) that nothing could possibly explain (by consideration of production) any events that did materialize in the actual world and (b) that evidence gathered about the actual world u could not be brought to bear on the hypothetical world u' in which production is defined.

To overcome this hurdle, we resort to an aspect of causation called sustenance, which enriches the notion of dependence with features of production while remaining in a world

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

u in which both x and y are true. Sustenance differs from dependence in the type of contingencies against which x is expected to protect y. Whereas the contingencies considered in (10.3) are "circumstantial" – that is, evolving from a set of circumstances $U = u$ that are specific to the scenario at hand – we now insist that x will maintain y against contingencies that evolve from *structural* modification of the model itself (Pearl 1998b).

Definition 10.2.1 (Sustenance)
Let W be a set of variables in V, and let w, w' be specific realizations of these variables. We say that x causally sustains y in u relative to contingencies in W if and only if

$$\begin{aligned} &\text{(i)} && X(u) = x; \\ &\text{(ii)} && Y(u) = y; && (10.5)\\ &\text{(iii)} && Y_{xw}(u) = y \textit{ for all } w; \textit{ and} \\ &\text{(iv)} && Y_{x'w'}(u) = y' \neq y \textit{ for some } x' \neq x \textit{ and some } w'. \end{aligned}$$

The sustenance feature of (10.5) is expressed in condition (iii), $Y_{xw}(u) = y$, which requires that x *alone* be sufficient for maintaining y. It reads: If we set X to its actual value (x) in u then, even if W is set to any value (w) that is different from the actual, Y will still retain its actual value (y) in u. Condition (iv), $Y_{x'w'}(u) = y'$, attributes to $X = x$ the "responsibility" for sustaining $Y = y$ under such adverse conditions; if we set X to some other value (x'), then Y will relinquish its current value (y) under at least one setting $W = w'$. Put together, (iii) and (iv) imply that there exists a setting $W = w'$ in which x is both necessary and sufficient for y.

Is sustenance a reasonable requirement to impose on an "actual cause"? Consider again the two bullets that caused the prisoner's death in Figure 9.1. We consider A to be an actual cause of D in this scenario, because A would have sustained D "alone," even in the absence of B. But how do we express, formally, the absence of B in our scenario $U = u$, given that B did in fact occur? If we wish to (hypothetically) suppress B within the context created by u, then we must use a structural contingency and imagine that B is made false by some external intervention (or "miracle") that violates the rule $B = C$; for example, that rifleman B was prevented from shooting by some mechanical failure. We know perfectly well that such failure did not occur, yet we are committed to contemplating such failures by the very act of representing our story in the form of a multistage causal model, as in Figure 9.1.

Recalling that every causal model stands not for just one but for a whole set of models, one for each possible state of the $do(\cdot)$ operator, contemplating interventional contingencies is an intrinsic feature of every such model. In other words, the autonomy of the mechanisms in the model means that each mechanism advertises its possible breakdown, and these breakdowns signal contingencies against which causal explanations should operate. It is reasonable, therefore, that we build such contingencies into the definition of actual causation, which is a form of explanation.

The choice of W in Definition 10.2.1 should be made with caution. We obviously cannot permit W to include all variables that mediate between X and Y, for this would preclude any x from ever sustaining y. More seriously, by not restricting W we run the risk of removing genuine preemptions and turning noncauses into causes. For example, by choosing $W = \{X\}$ and $w' = 0$ in the desert traveler story (Figure 10.2), we turn

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

enemy 1 into the actual cause of death, contrary to intuition and contrary to the actual scenario (which excludes cyanide intake). The notion of "causal beam" (Pearl 1998b) is devised to make the choice of W minimally disruptive to the actual scenario.[5]

10.3 CAUSAL BEAMS AND SUSTENANCE-BASED CAUSATION

10.3.1 Causal Beams: Definitions and Implications

We start by considering a causal model M, as defined in Section 7.1, and selecting a subset S of *sustaining* parent variables for each family and each u. Recall that the arguments of the functions $\{f_i\}$ in a causal model were assumed to be minimal in some sense, since we have pruned from each f_i all redundant arguments and retained only those called pa_i that render $f_i(pa_i, u)$ nontrivial (Definition 7.1.1). However, in that definition we were concerned with nontriviality relative to all possible u; further pruning is feasible when we are situated at a particular state $U = u$.

To illustrate, consider the function $f_i = ax_1 + bux_2$. Here $PA_i = \{X_1, X_2\}$, because there is always some value of u that would make f_i sensitive to changes in either x_1 or x_2. However, given that we are in a state for which $u = 0$, we can safely consider X_2 to be a trivial argument, replace f_i with $f_i^0 = ax_1$, and consider X_1 as the only *essential* argument of f_i^0. We shall call f_i^0 the *projection* of f_i on $u = 0$; more generally, we will consider the projection of the entire model M by replacing every function in $\{f_i\}$ with its projection relative to a specific u and a specific value of its nonessential part. This leads to a new model, which we call *causal beam*.

Definition 10.3.1 (Causal Beam)
For model $M = \langle U, V, \{f_i\}\rangle$ and state $U = u$, a causal beam *is a new model $M_u = \langle u, V, \{f_i^u\}\rangle$ in which the set of functions f_i^u is constructed from $\{f_i\}$ as follows.*

1. *For each variable $V_i \in V$, partition PA_i into two subsets, $PA_i = S \cup \overline{S}$, where S (connoting "sustaining") is any subset of PA_i satisfying*[6]

$$f_i(S(u), \bar{s}, u) = f_i(S(u), \bar{s}', u) \quad \textit{for all } \bar{s}'. \tag{10.6}$$

   In words, S is any set of PA_i sufficient to entail the actual value of $V_i(u)$, regardless of how we set the other members of PA_i.

2. *For each variable $V_i \in V$, find a subset W of $\overline{S}$ for which there exists some realization $W = w$ that renders the function $f_i(s, \overline{S}_w(u), u)$ nontrivial in s; that is,*

$$f_i(s', \overline{S}_w(u), u) \neq V_i(u) \quad \textit{for some } s'.$$

[5] Halpern and Pearl (1999) permit the choice of any set W such that its complement, $Z = V - W$, is sustained by x; that is, $Z_{xw}(u) = Z(u)$ for all w.

[6] Pearl (1998b) required that S be minimal, but this restriction is unnecessary for our purposes (though all our examples will invoke minimally sufficient sets). As usual, we use lowercase letters (e.g., s, $\bar{s}$) to denote specific realizations of the corresponding variables (e.g., S, $\overline{S}$) and use $S_x(u)$ to denote the realization of S under $U = u$ and $do(X = x)$. Of course, each parent set PA_i would have a distinct partition $PA_i = S_i \cup \overline{S}_i$, but we drop the i index for clarity.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

Here, $\overline{S}$ should not intersect the sustaining set of any other variable $V_j, j \neq i$. (Likewise, setting $W = w$ should not contradict any such setting elsewhere.)

3. *Replace $f_i(s, \overline{s}, u)$ by its projection $f_i^u(s)$, which is given by*

$$f_i^u(s) = f_i(s, \overline{S}_w(u), u). \tag{10.7}$$

Thus the new parent set of V_i becomes $PA_i^u = S$, and every f^u function is responsive to its new parent set S.

Definition 10.3.2 (Natural Beam)
A causal beam M_u is said to be natural if condition 2 of Definition 10.3.1 is satisfied with $W = \emptyset$ for all $V_i \in V$.

In words, a natural beam is formed by "freezing" all variables outside the sustaining set at their actual values, $\overline{S}(u)$, thus yielding the projection $f_i^u(s) = f_i(s, \overline{S}(u), u)$.

Definition 10.3.3 (Actual Cause)
We say that event $X = x$ was an actual cause of $Y = y$ in a state u (abbreviated "x caused y") if and only if there exists a natural beam M_u such that

$$Y_x = y \quad \text{in } M_u \tag{10.8}$$

and

$$Y_{x'} \neq y \text{ in } M_u \quad \text{for some } x' \neq x. \tag{10.9}$$

Note that (10.8) is equivalent to

$$Y_x(u) = y, \tag{10.10}$$

which is implied by $X(u) = x$ and $Y(u) = y$. But (10.9) ensures that, after "freezing the trivial surroundings" represented by $\overline{S}$, $Y = y$ would not be sustained by some value x' of X.

Definition 10.3.4 (Contributory Cause)
We say that x is a contributory cause of y in a state u if and only if there exists a causal beam, but no natural beam, that satisfies (10.8) *and* (10.9).

In summary, the causal beam can be interpreted as a theory that provides a sufficient and nontrivial explanation for each actual event $V_i(u) = v_i$ under a hypothetical freezing of some variables ($\overline{S}$) by the $do(\cdot)$ operator. Using this new theory, we subject the event $X = x$ to a counterfactual test and check whether Y would change if X were not x. If a change occurs in Y when freezing takes place at the actual values of $\overline{S}$ (i.e., $W = \emptyset$), we say that "x was an actual cause of y." If changes occur only under a freeze state that is removed from the actual state (i.e., $W \neq \emptyset$), we say that "x was a contributory cause of y."

> ***Remark:*** Although W was chosen to make V_i responsive to S, this does not guarantee that $S(u)$ is necessary and sufficient for $V_i(u)$ because local responsiveness

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

does not preclude the existence of another state $s'' \neq S(u)$ for which $f_i^u(s'') = V_i(u)$. Thus, (10.8) does not guarantee that x is both necessary and sufficient for y. That is the reason for the final counterfactual test in (10.9). It would be too restrictive to require that w render f^u nontrivial for every s of S; such a W may not exist. If (10.8)–(10.9) are satisfied, then $W = w$ represents some hypothetical modification of the world model under which x is both sufficient and necessary for y.

Remarks on Multivariate Events: Although Definitions 10.3.3 and 10.3.4 apply to univariate as well as multivariate causes and effects, some refinements are in order when X and Y consist of sets of variables.[7] If the effect considered, E, is any Boolean function of a set $Y = \{Y_1, \dots, Y_k\}$ of variables, then (10.8) should apply to every member Y_i of Y, and (10.9) should be modified to read $Y_{x'} \implies \neg E$ instead of $Y_{x'} \neq y$. Additionally, if X consists of several variables, then it is reasonable to demand that X be minimal – in other words, to demand that no subset of those variables passes the test of (10.8)–(10.9). This requirement strips X from irrelevant, overspecified details. For example, if drinking poison qualifies as the actual cause of Joe's death then, awkwardly, drinking poison and sneezing would also pass the test of (10.8)–(10.9) and qualify as the cause of Joe's death. Minimality removes "sneezing" from the causal event $X = x$.

Incorporating Probabilities and Evidence

Suppose that the state u is uncertain and that the uncertainty is characterized by the probability $P(u)$. If e is the evidence available in the case, then the probability that x caused y can be obtained by summing up the weight of evidence $P(u \mid e)$ over all states u in which the assertion "x caused y" is true.

Definition 10.3.5 (Probability of Actual Causation)
Let U_{xy} be the set of states in which the assertion "x is an actual cause of y" is true (Definition 10.3.2), and let U_e be the set of states compatible with the evidence e. The probability that x caused y in light of evidence e, denoted $P(caused(x, y \mid e))$, is given by the expression

$$P(\text{caused } (x, y \mid e)) = \frac{P(U_{x\,y} \cap U_e)}{P(U_e)}. \tag{10.11}$$

10.3.2 Examples: From Disjunction to General Formulas

Overdetermination and Contributory Causes

Contributory causation is typified by cases where two actions concur to bring about an event yet either action, operating alone, would still have brought about the event. In such cases the model consists of just one mechanism, which connects the effect E to the two

[7] These were formulated by Joseph Halpern in the context of the definition presented in Halpern and Pearl (1999).

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

actions through a simple disjunction: $E = A_1 \vee A_2$. There exists no natural beam to qualify either A_1 or A$_2$ as an actual cause of E. If we fix either A_1 or A_2 at its current value (namely, true), then E will become a trivial function of the other action. However, if we deviate from the current state of affairs and set A_2 to false (i.e., forming a beam with $W = \{A_2\}$ and setting W to false), then E would then become responsive to A_1 and so pass the counterfactual test of (10.9).

This example illustrates the sense in which the beam criterion encapsulates Lewis's notion of quasi-dependence. Event E can be considered quasi-dependent on A_1 if we agree to test such dependence in a hypothetical submodel created by the $do(A_2 = \text{false})$ operator. In Section 10.2 we argued that such a hypothetical test – though it conflicts with the current scenario u – is implicitly written into the charter of every causal model. A causal beam may thus be considered a formal explication of Lewis's notion of a quasi-dependent process, and the combined sets W represent the "peculiar surroundings" of the process that (when properly modified) renders $X = x$ necessary for $Y = y$.

Disjunctive Normal Form

Consider a single mechanism characterized by the Boolean function

$$y = f(x, z, r, h, t, u) = xz \vee rh \vee t,$$

where (for simplicity) the variables X, Z, R, H, T are assumed to be causally independent of each other (i.e., none is a descendant of another in the causal graph $G(M)$). We next illustrate conditions under which x would qualify as a contributory or an actual cause for y.

First, consider a state $U = u$ where all variables are true:

$$X(u) = Z(u) = R(u) = H(u) = T(u) = Y(u) = \text{ true}.$$

In this state, every disjunct represents a minimal set of sustaining variables. In particular, taking $S = \{X, Z,\}$ we find that the projection $f^u(x, z) = f(x, z, R(u), H(u), T(u))$ becomes trivially true. Thus, there is no natural beam M_u, and x could not be the actual cause of y. Feasible causal beams can be obtained by using $w = \{r', t'\}$ or $w = \{h', t'\}$, where primes denote complementation. Each of these two choices yields the projection $f^u(x, z) = xz$. Clearly, M_u meets the conditions of (10.8) and (10.9), thus certifying x as a contributory cause of y.

Using the same argument, it is easy to see that, at a state u' for which

$$X(u') = Z(u') = \text{true} \quad \text{and} \quad R(u') = T(u') = \text{false},$$

a natural beam exists; that is, a nontrivial projection $f^{u'}(x, z) = xz$ is realized by setting the redundant ($\bar{S}$) variables R, H, and T to their actual values in u'. Hence, x qualifies as an actual cause of y.

This example illustrates how Mackie's intuition for the INUS condition can be explicated in the structural framework. It also illustrates the precise roles played by structural (or "dispositional") knowledge (e.g., $f_i(pa_i, u)$) and circumstantial knowledge ($X(u) =$ true), which were not clearly distinguished by the strictly logical account.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

The next example illustrates how the INUS condition generalizes to arbitrary Boolean functions, especially those having several minimal disjunctive normal forms.

Single Mechanism in General Boolean Form

Consider the function

$$y = f(x, z, h, u) = xz' \vee x'z \vee xh', \tag{10.12}$$

which has the equivalent form

$$y = f(x, z, h, u) = xz' \vee x'z \vee zh'. \tag{10.13}$$

Assume, as before, that (a) we consider a state u in which X, Z, and H are true and (b) we inquire as to whether the event $x : X =$ true caused the event $y : Y =$ false. In this state, the only sustaining set is $S = \{X, Z, H\}$, because no choice of two variables (valued at this u) would entail $Y =$ false regardless of the third. Since $\overline{S}$ is empty, the choice of beam is unique: $M_u = M$, for which $y = f^u(x, z, h) = xz' \vee x'z \vee xh'$. This M_u passes the counterfactual test of (10.9), because $f^u(x', z, h) =$ true; we therefore conclude that x was an actual cause of y. Similarly, we can see that the event $H =$ true was an actual cause of $Y =$ false. This follows directly from the counterfactual test

$$Y_h(u) = \text{false} \quad \text{and} \quad Y_{h'}(u) = \text{true}.$$

Because Definitions 10.3.3 and 10.3.4 rest on semantical considerations, identical conclusions would be obtained from any logically equivalent form of f (not necessarily in minimal disjunctive form) – as long as f represents a single mechanism. In simple, single-mechanism models, the beam criterion can therefore be considered the semantical basis behind the INUS intuition. The structure-sensitive aspects of the beam criterion will surface in the next two examples, where models of several layers are considered.

10.3.3 Beams, Preemption, and the Probability of Single-Event Causation

In this section we apply the beam criterion to a probabilistic version of the desert traveler example. This will illustrate (i) how structural information is utilized in problems involving preemption and (ii) how we can compute the probability that one event "was the actual cause of another," given a set of observations.

Consider a modification of the desert traveler example in which we do not know whether the traveler managed to drink any of the poisoned water before the canteen was emptied. To model this uncertainty, we add a bivalued variable U that indicates whether poison was drunk ($u = 0$) or not ($u = 1$). Since U affects both D and C, we obtain the structure shown in Figure 10.3. To complete the specification of the model, we need to assign functions $f_i(pa_i, u)$ to the families in the diagram and a probability distribution $P(u)$. To formally complete the model, we introduce the dummy background variables U_X and U_P, which represent the factors behind the enemies' actions.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

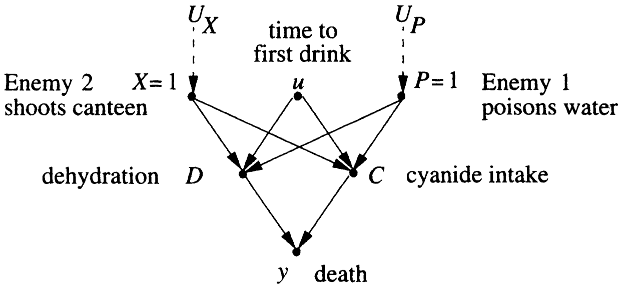

Figure 10.3 Causal relationships for the probabilistic desert traveler.

The usual understanding of the story yields the following functional relationships:

$$c = p(u' \vee x'),$$
$$d = x(u \vee p'),$$
$$y = c \vee d,$$

together with the evidential information

$$X(u_X) = 1, \qquad P(u_P) = 1.$$

(We assume that T will not survive with an empty canteen (x) even after drinking unpoisoned water before the shot ($p'u'$).)

In order to construct the causal beam M_u, we examine each of the three functions and form their respective projections on u. For example, for $u = 1$ we obtain the functions shown in (10.1), for which the (minimal) sustaining parent sets are: X (for C), X (for D), and D (for Y). The projected functions become

$$c = x', \quad d = x, \quad y = d, \tag{10.14}$$

and the beam model $M_{u=1}$ is natural; its structure is depicted in Figure 10.4. To test whether x (or p) was the cause of y, we apply (10.8)–(10.9) and obtain

$$\begin{aligned} &Y_x = 1 \text{ and } Y_{x'} = 0 \text{ in } M_{u=1}, \\ &Y_p = 1 \text{ and } Y_{p'} = 1 \text{ in } M_{u=1}. \end{aligned} \tag{10.15}$$

Thus, enemy 2 shooting at the container (x) is classified as the actual cause of T's death (y), whereas enemy 1 poisoning the water (p) was not the actual cause of y.

Next, consider the state $u = 0$, which denotes the event that our traveler reached for a drink before enemy 2 shot at the canteen. The graph corresponding to $M_{u=0}$ is shown in Figure 10.5 and gives

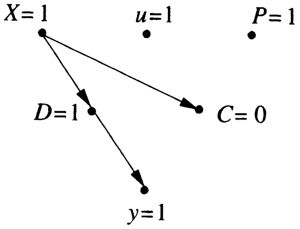

Figure 10.4 Natural causal beam representing the state $u = 1$.

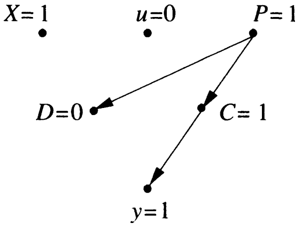

Figure 10.5 Natural causal beam representing the state $u = 0$.

$$\begin{aligned} &Y_x = 1 \text{ and } Y_{x'} = 1 \quad \text{in} \quad M_{u=0}, \\ &Y_p = 1 \text{ and } Y_{p'} = 0 \quad \text{in} \quad M_{u=0}. \end{aligned} \tag{10.16}$$

Thus, in this state of affairs we classify enemy 1's action to be the actual cause of T's death, while enemy 2's action is not considered the cause of death.

If we do not know which state prevailed, $u = 1$ or $u = 0$, then we must settle for the *probability* that x caused y. Likewise, if we observe some evidence e reflecting on the probability $P(u)$, such evidence would yield (see (10.11))

$$P(\text{caused}(x, y \mid e)) = P(u = 1 \mid e)$$

and

$$P(\text{caused}(p, y \mid e)) = P(u = 0 \mid e).$$

For example, a forensic report confirming "no cyanide in the body" would rule out state $u = 0$ in favor of $u = 1$, and the probability of x being the cause of y becomes 100%. More elaborate probabilistic models are analyzed in Pearl (1999).

10.3.4 Path-Switching Causation

Example 10.3.6 Let x be the state of a two-position switch. In position 1 ($x = 1$), the switch turns on a lamp ($z = 1$) and turns off a flashlight ($w = 0$). In position 0 ($x = 0$), the switch turns on the flashlight ($w = 1$) and turns off the lamp ($z = 0$). Let $Y = 1$ be the proposition that the room is lighted.

The causal beams M_u and $M_{u'}$ associated with the states in which the switch is in position 1 and 2 (respectively) are shown in the graphs of Figure 10.6. Once again, M_u entails

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

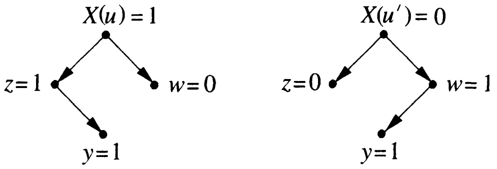

Figure 10.6 Natural beams that represent path switching in Example 10.3.6.

$Y_x = 1$ and $Y_{x'} = 0$. Likewise $M_{u'}$ entails $Y_x = 1$ and $Y_{x'} = 0$. Thus "switch in position 1" and "switch in position 2" are *both* considered actual causes for "room is lighted," although neither is a necessary cause.

This example further highlights the subtlety of the notion of "actual cause"; changing X from 1 to 0 merely changes the course of the causal pathway while keeping its source and destination the same. Should the current switch position ($X = 1$) be considered the actual cause of (or an "explanation of") the light in the room? Although $X = 1$ enables the passage of electric current through the lamp and is in fact the only mechanism currently sustaining light, one may argue that it does not deserve the title "cause" in ordinary conversation. It would be odd to say, for instance, that $X = 1$ was the cause of spoiling an attempted burglary. However, recalling that causal explanations earn their value in the abnormal circumstances created by structural contingencies, the possibility of a malfunctioning flashlight should enter our mind whenever we designate it as a separate mechanism in the model. Keeping this contingency in mind, it should not be too odd to name the switch position as a cause of spoiling the burglary.

10.3.5 Temporal Preemption

Consider the example mentioned in the preface of this chapter, in which two fires are advancing toward a house. If fire A burned the house before fire B, then we would consider fire A "the actual cause" of the damage, even though fire B would have done the same were it not for A. If we simply write the structural model as

$$H = A \vee B,$$

where H stands for "house burns down," then the beam method would classify each fire as an equally contributory cause, which is counterintuitive – fire B is not regarded as having made any contribution to H.

This example is similar to yet differs from the desert traveler; here, the way in which one cause preempts the other is more subtle in that the second cause becomes ineffective only because the effect has already happened. Hall (2004) regards this sort of preemption as equivalent to ordinary preemption, and he models it by a causal diagram in which H, once activated, inhibits its own parents. Such inhibitory feedback loops lead to irreversible behavior, contrary to the unique-solution assumption of Definition 7.1.1.

A more direct way of expressing the fact that a house, once burned, will remain burned even when the causes of fire disappear is to resort to dynamic causal models (as in Figure 3.3), in which variables are time-indexed. Indeed, it is impossible to capture temporal relationships such as "arriving first" by using the static causal models defined in Section 7.1; instead, dynamic models must be invoked.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

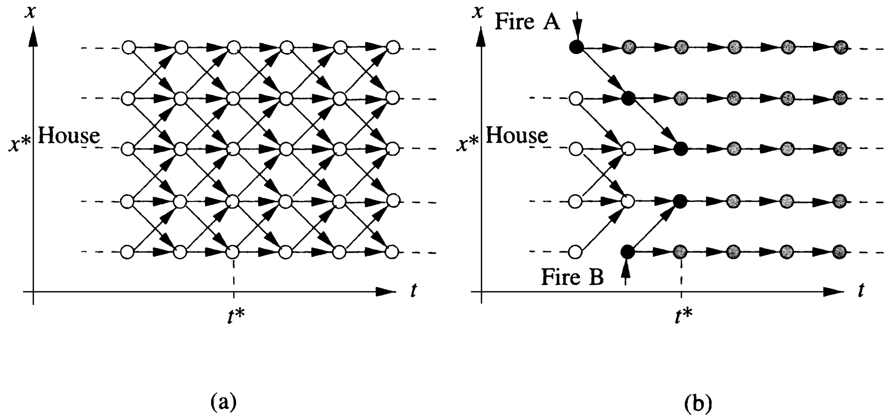

Figure 10.7 (a) Causal diagram associated with the dynamic model of (10.17). (b) Causal beam associated with starting fire A and fire B at different times, showing no connection between fire B and the state of the house at $x = x^*$.

Let the state of the fire $V(x, t)$ at location x and time t take on three values: g (for green), f (for on fire), and b (for burned). The dynamic structural equations characterizing the propagation of fire can then be written (in simplified form) as:

$$V(x,t) = \begin{cases} f & \text{if } V(x, t-1) = g \text{ and } V(x-1, t-1) = f, \\ f & \text{if } V(x, t-1) = g \text{ and } V(x+1, t-1) = f, \\ b & \text{if } V(x, t-1) = b \text{ or } V(x, t-1) = f, \\ g & \text{otherwise.} \end{cases} \qquad (10.17)$$

The causal diagram associated with this model is illustrated in Figure 10.7(a), designating three parents for each variable $V(x, t)$: the previous state $V(x + 1, t - 1)$ of its northern neighbor, the previous state $V(x - 1, t - 1)$ of its southern neighbor, and the previous state $V(x, t - 1)$ at location x. The scenario emanating from starting fire A and fire B one time unit apart (corresponding to actions $do(V(x^* + 2, t^* - 2) = f)$ and $do(V(x^* - 2, t^* - 1) = f))$ is shown in Figure 10.7(b). Black and grey bullets represent, respectively, space–time regions in states f (on fire) and b (burned). This beam is both natural and unique, as can be seen from (10.17). The arrows in Figure 10.7(b) represent a natural beam constructed from the (unique) minimally sufficient sets S at each family. The state of the parent set S that this beam assigns to each variable constitutes an event that is both necessary and sufficient for the actual state of that variable (assuming variables in $\bar{S}$ are frozen at their actual values).

Applying the test of (10.9) to this beam, we find that a counterfactual dependence exists between the event $V(x^* - 2, t^* - 2) = f$ (representing the start of fire A) and the sequence $V(x^*, t)$, $t > t^*$ (representing the state of the house through time). No such dependence exists for fire B. On that basis, we classify fire A as the actual cause of the house fire. Remarkably, the common intuition of attributing causation to an event that hastens the occurrence of the effect is seen to be a corollary of the beam test in the spatiotemporal representation of the story. However, this intuition cannot serve as the

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

defining principle for actual causation, as suggested by Paul (1998). In our story, for example, each fire alone did not hasten (or delay, or change any property of) the following event: E = the owner of the house did not enjoy breakfast the next day. Yet we still consider fire A, not B, to be the actual cause of E, as predicted by the beam criterion.

The conceptual basis of this criterion can be illuminated by examining the construction of the minimal beam shown in Figure 10.7(b). The pivotal step in this construction lies in the space–time region (x^*, t^*), which represents the house at the arrival of fire. The variable representing the state of the house at that time, $V(x^*, t^*)$, has a two-parent sustaining set, $S = \{V(x^* + 1, t^* - 1)$ and $V(x^*, t^* - 1)\}$, with values f and g, respectively. Using (10.17), we see that the south parent $V(x^* - 1, t^* - 1)$ is redundant, because the value of $V(x^*, t^*)$ is determined (at f) by the current values of the other two parents. Hence, this parent can be excluded from the beam, rendering $V(x^*, t^*)$ dependent on fire A. Moreover, since the value of the south parent is g, that parent cannot be part of any minimally sustaining set, thus ensuring that $V(x^*, t^*)$ is independent of fire B. (We could, of course, add this parent to S, but $V(x^*, t^*)$ would remain independent of fire B.) The next variable to examine is $V(x^*, t^* + 1)$, with parents $V(x^* + 1, t^*)$, $V(x^*, t^*)$, and $V(x^* - 1, t^*)$ valued at b, f, and f, respectively. From (10.17), the value f of the middle parent is sufficient to ensure the value b for the child variable; hence this parent qualifies as a singleton sustaining set, $S = \{V(x^*, t^*)\}$, which permits us to exclude the other two parents from the beam and so render the child dependent on fire A (through S) but not on fire B. The north and south parents are not, in themselves, sufficient for sustaining the current value (b) of the child node (fires at neighboring regions can cause the house to catch fire but not to become immediately "burned"); hence we must keep the middle parent in S and, in so doing, we render all variables $V(x^*, t)$, $t > t^*$, independent of fire B.

We see that sustenance considerations lead to the intuitive results through two crucial steps: (1) permitting the exclusion (from the beam) of the south parent of every variable $V(x^*, t)$, $t > t^*$, thus maintaining the dependence of $V(x^*, t)$ on fire A; and (2) requiring the inclusion (in any beam) of the middle parent of every variable $V(x^*, t)$, $t > t^*$, thus preventing the dependence of $V(x^*, t)$ on fire B. Step (1) corresponds to selecting the intrinsic process from cause to effect and then suppressing the influence of its nonintrinsic surrounding. Step (2) prevents the growth of causal processes beyond their intrinsic boundaries.

10.4 CONCLUSIONS

We have seen that the property of sustenance (Definition 10.2.1), as embodied in the beam test (Definition 10.3.3), is the key to explicating the notion of actual causation (or "cause in fact," in legal terminology); this property should replace the "but for" test in cases involving multistage scenarios with several potential causes. Sustenance captures the capacity of the putative cause to maintain the value of the effect in the face of structural contingencies and includes the counterfactual test of necessity as a special case, with structural contingencies suppressed (i.e., $W = \emptyset$). We have argued that (a) it is the structural rather than circumstantial contingencies that convey the true meaning of

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

causal claims and (b) these structural contingencies should therefore serve as the basis for causal explanation. We further demonstrated how explanations based on such contingencies resolve difficulties that have plagued the counterfactual account of single-event causation – primarily difficulties associated with preemption, overdetermination, temporal preemption, and switching causation.

Sustenance, however, does not totally replace production, the second component of sufficiency – that is, the capacity of the putative cause to produce the effect in situations where the effect is absent. In the match–oxygen example (see Section 9.5), for instance, oxygen and a lit match each satisfy the sustenance test of Definition 10.3.3 (with $W = \emptyset$ and $\overline{S} = \emptyset$); hence, each factor would qualify as an actual cause of the observed fire. What makes oxygen an awkward explanation in this case is not its ineptness at sustaining fire against contingencies (the contingency set W is empty) but rather its inability to produce fire in the most common circumstance that we encounter, $U = u'$, in which a match is not struck (and a fire does not break out).

This argument still does not tell us why we should consider such hypothetical circumstances ($U = u'$) in the match–oxygen story and not, say, in any of the examples considered in this chapter, where sustenance ruled triumphantly. With all due respect to the regularity and commonality of worlds $U = u'$ in which a match is not struck, those are nevertheless contrary-to-fact worlds, since a fire did break out. Why, then, should one travel to such a would-be world when issuing an explanation for events (fire) in the actual world?

The answer, I believe, lies in the pragmatics of the explanation sought. The tacit target of explanation in the match–oxygen story is the question: “How could the fire have been prevented?” In view of this target, we have no choice but abandon the actual world (in which fire broke out) and travel to one ($U = u'$) in which agents are still capable of preventing this fire.[8]

A different pragmatics motivates the causal explanation in the switch–light story of Example 10.3.6. Here one might be more concerned with keeping the room lit, and the target question is: “How can we ensure that the room remains lit in the face of unforeseen contingencies?” Given this target, we might as well remain in the comfort of our factual world, $U = u$, and apply the criterion of sustenance rather than production.

It appears that pragmatic issues surrounding our quest for explanation are the key to deciding which facet of causation should be used, and that the mathematical formulation of this pragmatics is a key step toward the automatic generation of adequate explanations. Unfortunately, I must now leave this task for future investigation.

Acknowledgments

My interest in the topic of actual causation was kindled by Don Michie, who spent many e-mail messages trying to convince me that (1) the problem is not trivial and (2) Good’s

[8] Herbert Simon has related to me that a common criterion in accident liability cases, often applied to railroad crossing accidents, is the “last clear chance” doctrine: the person liable for a collision is the one who had the last clear chance of avoiding it.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

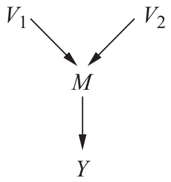

Figure 10.8 An example showing the need for beam refinement.

(1961, 1962) measures of causal tendency can be extended to handle individual events. He succeeded with regard to (1), and this chapter is based on a seminar given at UCLA (in the spring of 1998) in which "actual causation" was the main topic. I thank the seminar participants, Ray Golish, Andrew Lister, Eitan Mendelowitz, Peyman Meshkat, Igor Roizen, and Jin Tian for knocking down two earlier attempts at beams and sustenance and for stimulating discussions leading to the current proposal. Discussions with Clark Glymour, Igal Kvart, Jim Woodward, Ned Hall, Herbert Simon, Gary Schwartz, and Richard Baldwin sharpened my understanding of the philosophical and legal issues involved. Subsequent collaboration with Joseph Halpern helped to polish these ideas further and led to the more general and declarative definition of actual cause reported in Halpern and Pearl (2000).

Postscript for the Second Edition

Halpern and Pearl (2001a,b) discovered a need to refine the causal beam definition of Section 10.3.3. They retained the idea of defining actual causation by counterfactual dependency in a world perturbed by contingencies, but permitted a wider set of contingencies.

To see that the causal beam definition requires refinement, consider the following example.

> **Example 10.4.1** A vote takes place, involving two people. The measure Y is passed if at least one of them votes in favor. In fact, both of them vote in favor, and the measure passes.

This version of the story is identical to the disjunctive scenario discussed in Section 10.1.3, where we wish to proclaim each favorable vote, $V_1 = 1$ and $V_2 = 1$, a contributing cause of $Y = 1$.

However, suppose there is a voting machine that tabulates the votes. Let M represent the total number of votes recorded by the machine. Clearly $M = V_1 + V_2$ and $Y = 1$ iff $M \geq 1$. Figure 10.8 represents this more refined version of the story.

In this scenario, the beam criterion no longer qualifies $V_1 = 1$ as a contributing cause of $Y = 1$, because V_2 cannot be labeled "inactive" relative to M, hence we are not at liberty to set the contingency $V_2 = 0$ and test the counterfactual dependency of Y on V_1 as we did in the simple disjunctive case.

A refinement that properly handles such counterexamples was proposed in Halpern and Pearl (2001a,b) but, unfortunately, Hopkins and Pearl (2002) showed that the

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press

constraints on the contingencies were too liberal. This led to a further refinement (Halpern and Pearl 2005a,b) and to the definition given below:

Definition 10.4.2 (Actual Causation) (Halpern and Pearl 2005)
$X = x$ is an actual cause of $Y = y$ in a world $U = u$ if the following three conditions hold:

AC1. $X(u) = x$, $Y(u) = y$

AC2. *There is a partition of V into two subsets, Z and W, with $X \subseteq Z$ and a setting x' and w of the variables in X and W, respectively, such that if $Z(u) = z^*$, then both of the following conditions hold*:

(a) $Y_{x',w} \neq y$.

(b) $Y_{x,w,z^*} = y$ *for all subsets W' of W and all subsets Z' of Z, with the setting w of W' and z^* of Z' equal to the setting of those variables in $W = w$ and $Z = z^*$, respectively.*

AC3. *W is minimal; no subset of X satisfies conditions AC1 and AC2.*

The assignment $W = w$ acts as a contingency against which $X = x$ is given the counterfactual test, as expressed in AC2(a).

AC2 (b) limits the choice of contingencies. Roughly speaking, it says that if the variables in X are reset to their original values, then $Y = y$ must hold, even under the contingency $W = w$ and even if some variables in Z are given their original values (i.e., the values in z^*).

In the case of the voting machine, if we identify $W = w$ with $V_2 = 0$, and $Z = z^*$ with $V_1 = 1$, we see that $V_i = 1$ qualifies as a cause under AC2; we no longer require that M remains invariant to the contingency $V_2 = 0$; the invariance of $Y = 1$ suffices.

This definition, though it correctly solves most problems posed in the literature (Hiddleston 2005; Hall 2007; Hitchcock 2007, 2008), still suffers from one deficiency; it must rule out certain contingencies as unreasonable. Halpern (2008) has offered a solution to this problem by appealing to the notion of "normality" in default logic (Spohn 1988; Kraus et al. 1990; Pearl 1990b); only those contingencies should be considered which are at the same level of "normality" as their counterparts in the actual world.

Halpern and Hitchcock (2010) summarize the state of the art of the structural approach to actual causation, and discuss its sensitivity to choice of variables.

https://doi.org/10.1017/CBO9780511803161.012 Published online by Cambridge University Press