

WILEY

A Causal Theory of Counterfactuals

Author(s): Eric Hiddleston

Source: *Noûs*, Vol. 39, No. 4 (Dec., 2005), pp. 632-657

Published by: Wiley

Stable URL: <http://www.jstor.org/stable/3506114>

Accessed: 03-09-2016 11:41 UTC

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
<http://about.jstor.org/terms>



Wiley is collaborating with JSTOR to digitize, preserve and extend access to *Noûs*

A Causal Theory of Counterfactuals

ERIC HIDDLESTON

Wayne State University

I develop an alternative to standard semantic theories of subjunctive or “counterfactual” conditionals using “causal models” of roughly the sort investigated by Judea Pearl (1999, 2000) and Spirtes et al. (2000). The currently standard approach, due primarily to Robert Stalnaker (1968, 1970) and David Lewis (1973, 1986b), appeals to relations of “similarity” of possible worlds. On this view, ‘If A had been, then C would have been’ is true iff C is true in all possible worlds in which A is true and which are otherwise “similar” to actuality.

Lewis suggests that similarity of possible worlds is a sort of global resemblance, something like the overall similarity of cities, faces, and philosophies (1973 94–5). Similarity of worlds can be specified without reference to causal relations, natural necessities, or other “nonHumean” rubbish. The idea is that we can then use counterfactuals to provide reductive analyses of causation, dispositions, and other causal concepts.

I argue that this approach gets the order of analysis backward: a proper understanding of counterfactuals requires reference to causal relations and practically nothing else. The general idea that I urge is nicely expressed by Igal Kvarn (1986 44):

In considering the contrary-to-fact assumption, we consider this change, and only this change, and contemplate the possible effects *of this change* against the background of the actual course of events, and only these effects. Consequently, in considering this change, we are not prompted to question those actual events which have nothing to do with it.

In evaluating ‘If A had been, then C would have been’, we start with a causal network of actual events. If A is actually false, we introduce a *minimal break* in that causal network that allows A to be true. Then we trace out the causal consequences of that break, holding fixed other matters which are not

influenced by the initial break or its consequences. ‘If A had been, then C would have been’ is true iff C is true in such situations. I will use “causal models” roughly like those of Judea Pearl (1999, 2000) and Spirtes et al. (2000) to represent causal networks and “minimal breaks” in them.

§1 presents Lewis’s theory of counterfactuals and raises some counterexamples. §2 offers a diagnosis. §3 introduces causal models and develops an alternative account of counterfactuals.

1. Lewis’s theory and some problem cases.

Lewis offers the following truth condition (1986b 41):

ANALYSIS 2: A counterfactual “If it were that A, then it would be that C” is (nonvacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than any world where A is true but C is false.

I will write such counterfactuals ‘ $A > C$ ’. Typically, Analysis 2 simplifies to the following:

ANALYSIS 2’: $A > C$ is true iff C is true in all worlds in which A is true and that are otherwise most similar to actuality.

Lewis says that Analysis 2 is a “skeleton”; “It must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context” (1986b 41). He does not tell us what “similarity” as such is, but he does offer a detailed account of the “standard resolution of vagueness” of overall similarity (1986b 47–8):

- (1) It is of the first importance to avoid big, widespread, diverse violations of law.
- (2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
- (3) It is of the third importance to avoid even small, localized, simple violations of law.
- (4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

This theory embodies the common idea that in evaluating $A > C$, we hold fixed the actual past of A, perhaps with some minor, prior divergence that allows A to come true (clauses (2)–(3)), suppose A true, and then let the future fall as actual law dictates or allows ((1)).

Here I raise two common sorts of cases in which Lewis’s theory recommends the wrong answer. The first sort of case (Examples #1–2) involves antecedents that are causally irrelevant to their consequents.

Example 1: Alice offers Ben a bet on a fair, indeterministic coin toss. Ben bets tails. Alice flips the coin, and it lands heads. Ben loses. Alice says to Ben, "If you had bet heads, you would have won". That seems true.¹

Example 2: There is a black box with a source that produces a photon once every ten seconds. Each photon travels toward a piece of glass in the box, and detectors register whether it is transmitted (chance .8) or reflected (chance .2). Lights on the box register the outcome: a green light for transmission, a red light for reflection. Alice and Ben have a policy that Ben always bets \$1 on reflection (when he bets at all), at fair odds of 4:1. Ben takes a break from the game and goes to get a soda from the machine down the hall. While he is gone, the red light (improbably) lights five consecutive times. When Ben returns, Alice informs him of this unlikely sequence and says, "If you had not left to get that soda, you would be \$20 richer now." That seems true again.

These cases are apparent counterexamples to Lewis's theory of counterfactuals. I will set aside Lewis's clause (4) ("approximate similarity") for the moment, and return to it below. In Example #1, Lewis's clause (2) recommends that similar worlds in which Ben bets "Heads" agree exactly with actual history until shortly before the time of Ben's bet. At such a point, clauses (2)–(3) together recommend that some minor events (presumably in Ben's brain) occur differently (by a "small miracle" if necessary, (3)). Those minor differences lead Ben to say "Heads" rather than "Tails". Clause (1) tells us that these worlds have the same laws as the actual world, and thus that the outcome of the coin toss is not determined by its prior causes. And that's all (setting aside (4) for the moment): from the point of the initial difference, all "similar" worlds proceed according to actual law, and all such worlds are "similar" in the relevant respects. In some such worlds, the coin lands tails. So, Lewis's theory recommends that Alice's counterfactual in Example #1 is false. But it seems straightforwardly true.

The second sort of counterexample involves antecedents which have some causal bearing on their consequents, but only of a positive sort (Kvart 1986 calls these cases of "pure positive influence", e.g., 39–40).

Example #3: Jane is bitten by a poisonous snake while hiking. There is an antidote to the snake's venom, but Jane does not have any of it with her. Improbably, Jane's body fights off the poison, and she survives. The antidote, while good, is not perfect: there remains some chance of death among people exposed to both the poison and the antidote. It seems true, 'If Jane had taken the antidote, then she would have lived'.

Lewis's theory (1)–(3) also counts this counterfactual false. In the most "similar" worlds according to (1)–(3) in which Jane takes the antidote, there is some chance that she dies anyway (the antidote is imperfect). So, it is not the case that Jane survives in *all* such worlds, on Lewis's account.

So far, I have set aside Lewis's clause (4): "It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly" (1986b 48). Lewis adds this clause, seemingly as an afterthought, precisely due to cases like Examples #1–2. He says in a parenthetical explanation of (4), "It is a good question whether approximate similarities of particular fact should have little weight or none. Different cases come out differently and I would like to know why" (1986b 48). He continues to mention three cases, one of which is the case of the coin toss (he describes it merely as "a case, reported in Slote"). He suggests that approximate similarity has some weight in the coin toss case and none in other cases.

But adding clause (4) to Lewis's theory changes many correct recommendations of clauses (1)–(3). Here is just one example.

Example #4: Alice offers Ben a bet on a coin toss, but this time Alice can influence its outcome. Ben bets tails. Hoping to win, Alice flips the coin so that it has a high chance (.8, say) of landing heads. It does land heads. She says to Ben, "If you had bet heads, then you would have won". That seems false.

Whatever it is exactly, we know that "approximate similarity" is supposed to apply to outcomes of coin tosses. If so, then Lewis's theory with clause (4) gives the wrong answer in Example #4.

In addition, clause (4) cannot explain other cases that embody the same phenomenon as Example #1.

Example #5: It is 1973. Nixon sits with his finger on the dreaded button, watching on closed circuit television as Alice and Ben bet on coin tosses. He tells his advisors that he will press the button should Ben win the next bet, and put it away otherwise. Ben bets tails, the coin lands heads, and Ben loses the bet. Nixon reluctantly puts the button back in the drawer. One of his advisors breathes a sigh of relief, thinking to himself, "If Ben had bet heads, then he would have won and Nixon would have destroyed us all." That seems true.

Example #5 is a variation on a famous case of Kit Fine's (1975). Whatever exactly "approximate similarity" is, it must rate resemblance of worlds with respect to the nonoccurrence of a nuclear holocaust very significant, and more significant than resemblance with respect to the outcome of a coin toss. So, whatever it is, it is not operative in Example #5. But Example #5 is simply the coin toss case witnessed by Nixon. It is highly implausible that 'If Ben had bet heads, then he would have won' receives a different explanation in Examples #1 and #5. So, approximate similarity does not explain Example #1, either.

By contrast with Lewis's suggested account, there is an obvious similarity between the causal facts of Examples #1 and #5, and an obvious difference between those cases and Example #4. When the coin toss is fair, Ben's

betting tails is not a cause of the outcome of heads; whether Nixon watches is irrelevant to that. But when Alice cheats, Ben's betting heads is a cause of the outcome via Alice's intention to cheat and her flipping the coin so that it has a high chance of landing heads.

2. Diagnosis.

These cases have received some discussion in the literature, but most authors (like Lewis) set them aside as puzzles for another day. I think attending to them provides the basis for a different approach to counterfactuals generally. I have seen only two serious attempts to explain the cases, due to Igal Kvart (1986 and other places) and Stephen Barker (1999).² Here I give two more examples, and draw a moral.

The "obvious" difference between the causal facts of Examples #1 and #4 is mirrored in a difference in their probabilistic relations: in #1, Ben's bet is probabilistically irrelevant to the outcome, while in #4 it is not. Kvart offers accounts of the cases that make use of these probabilistic relations. Here is a further problem:

Example #6: Alice offers Ben a bet on a fair, indeterministic coin toss. Ben bets tails, the coin lands heads, and Ben loses the bet. (6a): 'If Ben had bet heads, and then grabbed the coin from Alice and flipped it himself, he would have won'. (6b): 'If Ben had bet heads, grabbed the coin in flight, and then flipped it himself, he would have won'. Both (6a) and (6b) seem false.

Ben's grabbing the coin does not affect the *probability* that it lands heads (/he wins). But as in the cheater case (#4), Ben's grabbing the coin does constitute interference with the actual outcome of heads and we do not hold it fixed. Kvart is aware of these problems, and offers complex probabilistic theories of "causal irrelevance" and "purely positive causal relevance" to avoid them (1986 chaps. 4, 5). Problems remain, however, and later papers attempt fixes (1991, 1994). Eventually, Kvart proposes a development of his probabilistic theory of counterfactuals as a theory of causation itself (1997). I do not think Kvart's theory of causation is successful, but that is another story. The current point is that to explain the examples, we need to appeal to *actual causes*. No probabilistic substitute will do, unless it is not a substitute but a successful analysis. The actual causes of the coin's landing heads include that Alice flipped it, that it traveled along a certain trajectory, etc. Interference with those causes constitutes interference with the outcome, whether or not it alters the outcome's probability.³

A second natural explanation of the cases appeals to John Pollock's notion of "undercutting": "To say that *P* undercuts *Q* is to say that *P* nomically implies the falsity of every causal history of *Q*" (1981 245). Pollock's idea is that minimally altered P-worlds are ones that *maximally*

overlap the actual world with respect to facts *not* undercut by P (246). This captures the similarity of Examples #1 and #5 (fair toss), and their difference from #4 and #6 (cheater and grabber cases). When the coin toss is fair but not when Alice cheats, Ben's actual bet of tails is not among the causes or causal history of the outcome; thus the antecedent that Ben bets heads does not "undercut" that outcome and we hold it fixed.

But a case of Barker's (1999), shows that this natural explanation is too simple.

Example #7: Jane considers betting \$1000 that snake-eyes will come up on a roll of two six-sided dice. She refrains from betting, the dice are thrown, and they land snake-eyes. Unbeknownst to Jane, there are electromagnets under the table which if activated prevent dice from landing snake-eyes. The magnets are controlled by Mandrake, who has a policy of interfering with high-stakes bets 5% of the time. Consider: 'If Jane had bet on snake-eyes [B] and Mandrake had not activated the magnets $[\neg M]$, then the dice would have landed snake-eyes [S]', $(B \ \& \ \neg M) > S$.⁴ That seems true.

The problem in Example #7 is that Jane's not betting ($\neg B$) is an actual cause of the dice landing snake-eyes (S). (Jane's not betting is a cause of Mandrake's not activating the magnet, and that is a cause of the dice landing snake-eyes.) So, B and anything that entails it, such as $(B \ \& \ \neg M)$, undercuts S. So, Pollock's natural idea recommends that $(B \ \& \ \neg M) > S$ is false. But it is true.

The problem in Barker's case is that the antecedent B *might* causally interfere with the outcome S, by leading Mandrake to activate the magnet, but also *might not*. So, we cannot hope to specify the requisite sort of irrelevance once and for all as a relation between antecedents and consequents: if B undercuts S, then $(B \ \& \ \neg M) > S$ is false (wrong answer); if it does not, then $B > S$ is true (wrong again).

Barker's own theory is inspired by Nelson Goodman's "cotenability" theory (as is mine). Goodman's general idea is that $A > C$ is true iff C follows lawfully from A together with a set of actual facts F, a "factual background" for A. Goodman ran into problems of *circularity*: it seemed that in order to specify F, he had to appeal to further counterfactuals. Barker embraces this seeming circularity: in order to evaluate $B > S$, we must also evaluate 'If B, then it might be that M' (e.g., 436–7). But we can avoid circularity because eventually the process *terminates*. Barker proposes an intensely complex algorithm for this evaluation procedure (444–8).

But if we accept the moral of Example #6 and appeal to the relation *x is an actual cause y*, then there is a fairly simple and noncircular solution. What is true in Example #7 is that the dice would land snake-eyes *unless* something interferes with them. That is, if Jane had bet, it would be that: the dice land snake-eyes if Mandrake does not activate the magnet. In effect, we

“hold fixed” the *material conditional*: if the actual, immediate causes of snake-eyes occur (the roll of the dice, the magnet’s being unactivated), then it occurs.

To begin to fix an idea, let’s say that a “causal break” from the actual course of events occurs in a hypothetical situation when the actual, immediate causes of E occur, but E does not. Then the moral of all of our Examples #1–7 is: “minimally altered” situations for antecedent P are ones that avoid causal breaks *to the extent that it is possible* to do so, consistently with P’s being true. For example, some causal break is required for Ben to bet heads in Example #1. But given that Ben’s actual bet is not a cause of the coin flip’s outcome (when the toss is fair), that initial break leaves the actual causes of the flip’s outcome intact (and their causes, etc.); so, it would require an extra, gratuitous break to change the outcome. Similarly, some causal break is required for Jane to bet on snake-eyes in Barker’s case. In those situations, either (a) Mandrake does not activate the magnet, or (b) he does. In (a), it would require an *extra* break for the dice not to land snake-eyes: the actual immediate causes of snake-eyes (the roll of the dice, the magnet’s being unactivated, etc.) occur. But in (b) no such break is required. And given Jane’s bet, no break is required for Mandrake to activate the magnet. Applying this idea generally, *all* differences from actuality in “minimally altered” P-situations must trace back to some initial break.

This will be the guiding idea of the theory I suggest. In order to spell it out, I turn to “causal models”.

3. Causal models and counterfactuals.

In this section, I offer an account of counterfactuals using stripped down versions of the “causal models” of Judea Pearl (1999, 2000) and Spirtes, Glymour, and Scheines (SGS 2000). These can be pictorially represented in graphs such as Fig. 1. Fig. 1(a) is a graph for Example #3, the poison and antidote; Fig. 1(b) is a graph for a case of a match’s lighting after being struck in the presence of oxygen. (These are “Directed Acyclic Graphs” or DAGs.) The arrows in the graphs represent causal relations. The *parents* of a node X, $pa(X)$, are the nodes that have arrows into it: in Fig. 1(a), $pa(\text{Survival}) = \{\text{Poison}, \text{Antidote}\}$.

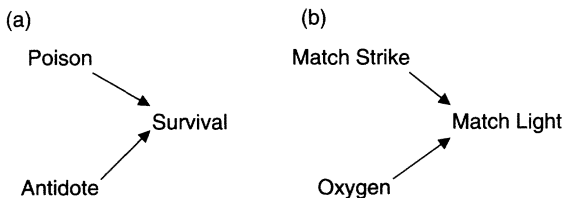


Figure 1. Two Examples.

The nodes of these graphs do not represent events, but *variables* for events (or properties⁵). In Fig. 1(a), the node Poison does not represent *that* Jane is poisoned by the snake, but *whether* she is. A causal model includes specifications of values for these variables: Poison = 0 (no poison) or Poison = 1 (poison). This might seem unnatural, but it allows us to describe relations among the variables with *causal laws* (or mini-laws) such as: if Poison = 1 and Antidote = 0, then $p(\text{Survival} = 1) = .05$. In the cases that Pearl and SGS focus on, these causal relations can be represented by “structural equations” which specify the value of a variable as a function of the values of its parents:

$$\text{Match Light} = \text{Match Strike} \times \text{Oxygen}.$$

This equation means: Match Light = 1 iff both Match Strike = 1 and Oxygen = 1.

The bulk of the work that SGS and Pearl do with causal models depends upon the assumption that the relevant causal relations are deterministic, and that probabilities are only measures of our ignorance. Their “structural equations” are fully deterministic: they specify the value of X as a *function* of the values of $pa(X)$. In addition, Pearl and SGS employ assumptions about causal relations and probabilities (the “Causal Markov Condition” and “Faithfulness”) which are inconsistent with some sorts of indeterminism.⁶ These may be good assumptions for some purposes, but they are probably false taken universally (they are on standard interpretations of quantum mechanics, at least). And commonsense seems to me to presuppose a sort of *quasi-indeterminism*: many processes such as coin flips and die rolls behave as if they were indeterministic, and so we treat them as such. This quasi-indeterminism may be due to either determinism or indeterminism at the fundamental level, and commonsense is not committed either way. But an adequate representation of ordinary causal and counterfactual reasoning (including Examples #1–7) must allow for it.

So, I will use causal models that are much stripped down from those of Pearl and SGS. I do not assume either “Markov” or “Faithfulness”, and I allow for indeterministic laws. These have the form

$$(Y_1 = y_1 \ \& \ \dots \ \& \ Y_n = y_n) \Rightarrow p(X = x) = z$$

where ‘ $\Rightarrow$ ’ is a strict conditional, the Y_i are all of the parents of X , the y_i are their specific values, and z is an “objective” probability that X has the value x .

A *causal model schema* is a pair $\langle G, E \rangle$. G is a graph (a DAG) consisting of a set of variables V and arrows between some of them. E is a set of equations relating the values (or probabilities of values) of each variable X

in G to the values of its parents $pa(X)$ in G . (Strictly speaking, the graph is redundant given the equations, but it is useful to keep in mind.)

A *causal model* is a triple $\langle G, E, A \rangle$, where $\langle G, E \rangle$ is a model schema, and A is an assignment of values to the variables in G which is *possible* given E . That is, no variable X gets a value $A(X)$ which the equations say has probability 0 given the values A assigns to X 's parents: for any X , $p(A(X)|A(pa(X))) > 0$.

So, for a simple example, our model for the match lighting would be the graph G of Fig. 1(b), the "law" E that $\text{Match Light} = \text{Match Strike} \times \text{Oxygen}$, and the assignment A of values $\text{Match Strike} = 1$, $\text{Oxygen} = 1$, $\text{Match Light} = 1$. In the case of the poison, the laws would be more complex, specifying the probabilities of $\text{Survival} = 1$ given different assignments of values to Poison and Antidote .

The case of the poison illustrates that the arrows in models do not correspond exactly to the relation *event x is an actual cause of event y* . There is an arrow from Antidote to Survival , but $\text{Antidote} = 0$ does not *cause* $\text{Survival} = 1$. Instead, Jane's not taking the antidote is a negative influence, a potential preventer of survival. The arrows represent relations of *causal influence*, and these are more general than the relation *x is an actual cause of y* . To a first approximation, causes are *positive* influences, and preventers are *negative* influences. The relation between variables $X \rightarrow Y$ is instead: the value of X is causally relevant to the probabilities of Y 's values.

The moral of the grabber case of the coin toss is that we need the relation *x is an actual cause of y* in our theory of counterfactuals. For practical reasons we should avoid tying our theory of counterfactuals to the full details of a theory of actual causes. But we also need a concept that is precise enough to give us definite predictions about the truth values of counterfactuals. So, I will use the substitute concept of *direct positive influence*. (In §5, I argue that this is a good substitute.)

The plan is this: I offer an account of "direct positive influence", use that to characterize a "causal break" in a model, and then offer an account of a "minimally altered" model. The idea is that a counterfactual $\phi > \psi$ is true in a model M when ψ is true in ϕ -models that are minimally altered from M .

In what follows, I will restrict my attention to models that share a model schema $\langle G, E \rangle$, and differ on the values assigned to the variables: the causal laws will always be "held fixed".⁷ So, suppose $M = \langle G, E, A \rangle$ is our "actual" model, and $A_1, A_2, \dots$, are the alternative possible valuations for $\langle G, E \rangle$. Let $M_i = \langle G, E, A_i \rangle$.

Suppose X is a parent of Y in M , $X = x$, $Y = y$, Y 's other parents are Z^- , and these parents have the values z^- .

Direct positive influence (first approximation⁸):

$X = x$ has *direct positive influence* on $Y = y$ in M iff

$$p(Y = y | X = x \ \& \ Z^- = z^-) > p(Y = y | X \neq x \ \& \ Z^- = z^-).$$

Let the *postive parents* of Y in M be $ppa(Y)_M = \{X: X = x \text{ has direct positive influence on } Y = y \text{ in } M\}$.

That is, $X = x$ has direct positive influence on $Y = y$ when the parent X 's having value x *raises the probability* that the child Y has value y (given the actual values of Y 's other parents). $ppa(Y)_M$ is the set of *positive parents* of Y in M . For example, in the case of the match from Fig. 1(b), the positive parents of Match Light = 1 are Match Strike = 1 and Oxygen = 1. But in the case of the antidote from Fig. 1(a), the set of positive parents of Survival = 1 is empty: both Poison = 1 and Antidote = 0 are negative influences.

The next step is to characterize "causal breaks":

Causal Break:

A *causal break* in model M_i relative to M is a variable Y such that $A_i(Y) \neq A(Y)$, and for each $X \in ppa_M(Y)$, $A_i(X) = A(X)$.

$Break(M_i, M) = \{Y: Y \text{ is a causal break in } M_i \text{ from } M\}$.

That is, a causal break (in M_i from M) occurs at a variable Y that has a nonactual value in M_i despite the fact that all of Y 's positive parents have their actual values in M_i (I illustrate in a moment). Note that if a variable Y has no parents, then differences in Y 's value count as breaks: same values of all positive parents (namely none), but different values for Y . $Break(M_i, M)$ is the set of breaks in M_i from M .

We will also need to represent cases in which models *preserve* causal relations among variables.

$Intact(M_i, M) = \{Y: A_i(Y) = A(Y) \text{ and for each } X \in ppa(Y)_M, A_i(X) = A(X)\}$.

Intact is the set of variables Y for which M_i gives actual values *both* to Y itself *and* to Y 's positive parents.

To illustrate these concepts, consider two models for a case of Alice and Ben betting on photons, illustrated in Fig. 2. In model 2(a), Ben bets on transmission (Bet = t), the photon is reflected, and he loses. In model 2(b), Ben bets on reflection (Bet = r) and wins. 2(b) contains a causal break from 2(a) at Bet: same values of positive parents of Bet as 2(a), namely none, different value of Bet. This is the only break from 2(a), however, so $Break(M_{2(b)}, M_{2(a)}) = \{Bet\}$. The parents of Emit (none) have the same values in 2(a) and 2(b), and so does Emit. Similarly for Reflect. $Intact(M_{2(b)}, M_{2(a)}) = \{Emit, Reflect\}$. Note that even though the value of Win differs in these two models, Win does *not* count as a causal break: a positive parent of Win (Bet) has a different value in 2(b) than in 2(a), so Win is not a candidate for inclusion in *Break* or *Intact*. The value of Win is "up for grabs", as we will say.

The next step is to characterize a “minimally altered” model. Let a ϕ -model be a model in which proposition ϕ is true. *Break* and *Intact* are sets of variables, and they can be larger or smaller or simply incomparable with other sets of variables. Let’s say that $Break(M_i, M)$ is *minimal* among ϕ -models iff there is no ϕ -model M_k such that $Break(M_i, M) \supset Break(M_k, M)$. That is, $Break(M_i, M)$ is minimal (among ϕ -models) when no other ϕ -model includes only a “smaller” set of breaks in the causal relations of M . “Smaller” here is represented by set-inclusion rather than by counting numbers of breaks.

Similarly, let’s say that $Intact(M_i, M)$ is *maximal* among ϕ -models iff there is no ϕ -model M_k such that $Intact(M_i, M) \subset Intact(M_k, M)$. That is, $Intact(M_i, M)$ is maximal (among ϕ -models) when there is no other ϕ -model that shares a “larger” set of intact causal relations with M .

These definitions allow that two distinct $Intact(M_i, M)$ and $Intact(M_k, M)$ could each be maximal (and similarly with minimal *Breaks*): each could include some variables that the other does not. Such cases are commonplace, and will be important later.

Suppose for the following definition that ϕ is either an atomic proposition, $X = x$, or a negation or conjunction of them.⁹

ϕ -Minimal Model:

Model M_i and $Break(M_i, M)$ are ϕ -minimal relative to M iff

- (i) M_i is a ϕ -model,
- (ii) for Z the set of variables in G that are not descendants of ϕ , $Intact(M_i, M) \cap Z$ is maximal among ϕ -models, and
- (iii) $Break(M_i, M)$ is minimal among ϕ -models.

The ϕ -minimal models are our substitute for the most “similar” ϕ -worlds. Note that if ϕ is true in M , then M is the unique ϕ -minimal model relative to M and $\{\}$ is the unique minimal *Break*.

In effect, clauses (ii) and (iii) specify two components of a “factual background” for ϕ . Clause (ii) says roughly that a factual background for ϕ includes a maximal set of actual events that are not causally influenced by ϕ (i.e., actual values for a maximal subset of ϕ ’s nondescendants Z).¹⁰

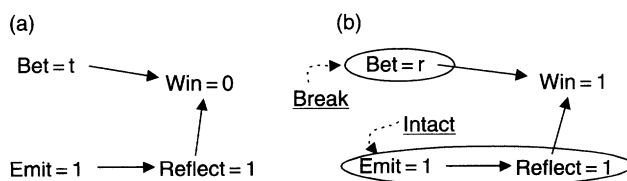


Figure 2. Betting on photons.

Changes to such matters must be as minimal as possible, consistently with ϕ 's being true. This forces any causal break to be as minor and late as is lawfully possible. Clause (iii) says in effect that a factual background also includes a maximal set of *material conditionals* of the sort: if the positive parents of X have their actual values, then X has its actual value. Clause (iii) tells us that (to the extent it is possible) actual cause-effect relations would obtain if ϕ were true, *unless* something nonactual prevents them from obtaining. Clause (iii) does not suffice on its own because it is consistent with (iii) that all variables have different values in M and M_i : one way to minimize "causal breaks" from the actual course of events is to have different events from the beginning of time (e.g., one break at the Big Bang is pretty minimal). And clause (ii) alone does not explain Barker's case. Roughly, clauses (i)–(iii) together have the effect that $\phi > \psi$ is true iff ψ follows lawfully from: (i) ϕ , (ii) events not causally influenced by ϕ , and (iii) actual causal relations that ϕ does not prevent from obtaining. Less roughly:

TCM (Theory of Counterfactuals in a Model):

$(\phi > \psi)$ is true in a model M and a context C iff ψ is true in every ϕ -minimal model M_i for which *Break*(M_i , M) is relevant in C.

I include the qualification about "relevant" *Breaks* to capture some facts about context-dependence. I will ignore those for the time being, however, and return to them in the next section. For current purposes, TCM says that $(\phi > \psi)$ is true in M iff ψ is true in all models that are ϕ -minimal relative to M.

To illustrate TCM, recall the case of the photon (Fig. 2). Suppose that model 2(a) represents the actual case. 2(b) gives the unique minimal model for the antecedent that Ben bets on reflection ($\text{Bet} = r$). *Break*($M_{2(b)}$, $M_{2(a)}$) = {Bet}, and there is no smaller *Break* that is consistent with $\text{Bet} = r$. Similarly, *Intact*($M_{2(b)}$, $M_{2(a)}$) = {Emit, Reflect} is maximal in ($\text{Bet} = r$)-models (among the nondescendants of Bet). For the antecedent $\text{Bet} = r$, we "hold fixed" the actual values of Emit and Reflect: they would retain their actual values whatever the prior probabilities. We add a minimal break from model 2(a) with $\text{Bet} = r$, and see what follows. The value of Win is "up for grabs", but it follows from the laws of the models that $\text{Win} = 1$. So, it is true in model 2(a) that Ben would win if he had bet on reflection.

Suppose Alice can cheat on the photon reflector (Fig. 3). She has a foot switch that changes the angle of the reflective glass inside the box, making reflection of the photon more likely (changing its chance from .2 to .8, say). Ben bets on transmission, Alice hits the foot switch, and the photon is reflected. Now it seems straightforwardly false, 'If Ben had bet on reflection, then he would have won'.

Suppose the laws of this model M include $\text{Switch} = \text{Bet}$. TCM recommends that ' $\text{Bet} = r > \text{Win} = 1$ ' is false in M. Again, any model in which

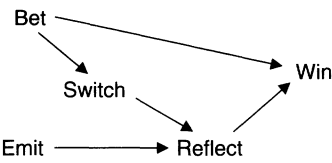


Figure 3. Cheating on photons.

$Bet = r$ includes a break from M at Bet . But the nonactual value of Bet then propagates through the system: $Switch$ is not a candidate for inclusion in *Break* or *Intact* because its positive parent Bet has a nonactual value. Similarly for $Reflect$ and Win . The unique maximal $Intact(M_i, M)$ is $\{Emit\}$, and the unique minimal $Break(M_i, M)$ is $\{Bet\}$. The values of the other variables are “up for grabs”, and this time the laws do not uniquely determine their values. So, ‘ $Bet = r > p(Win) = .2$ ’ is true, but it is false both that Ben would win and that he would lose.

Fig. 4 represents a case of a ceremonial cannon which explodes, creating a flash and a bang. Suppose the variables are binary and all have value 1 in M . Suppose the relevant laws of M are:

- (a) $(Fuse\ Lit = 1) \Rightarrow p(Explosion = 1) = .95$,
- (b) $Flash = Explosion$
- (c) $Bang = Explosion$

That is, the fuse is imperfect, but once there has been an explosion (in the actual absence of baffles, mufflers, etc.) there will definitely be both a flash and a bang. TCM counts it true, ‘If the flash had not occurred, then the bang would not have occurred’, ‘ $Flash = 0 > Bang = 0$ ’. $Intact(M_i, M) = \{Fuse\ Lit\}$ is maximal for $Flash = 0$ (among nondescendants of $Flash$) because there is no possible assignment on which $Flash = 0$ and $Explosion = 1$. By law (b) $Flash = 0$ necessitates $Explosion = 0$, and by (c) that necessitates $Bang = 0$. Lewis and Pearl would count as relevant a model M' in which $Flash = 0$ and $Bang = 1$. But if the laws really are as we stated, then this is an impossible situation, and “causal breaks” must be possible. I take it that $Flash = 0 > Bang = 0$ is intuitively true in this case.

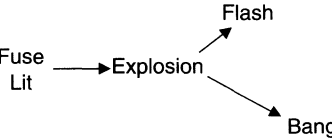


Figure 4. A cannon.

Lewis would claim that this answer is correct only in special “contexts” that allow “backtracking”, but it seems to me that our answer is uniquely correct given the facts of this case.

‘Flash = 0 > Fuse Lit = 1’ turns out true by TCM: the fuse would still have been lit if the flash had not occurred. TCM allows and requires “backtracking”, but only when the causes of an event lawfully necessitate it. Note though that “backtrackers” are true (when they are) for the same reasons as any other counterfactuals. That is, TCM gives a unified semantic account of “backtrackers” and others, while Lewis is forced to appeal to “nonstandard” similarity relations (Bennett 1984 emphasizes the importance of this unity).

The reader can verify that TCM recommends that $(B \ \& \ \neg M) > S$ is true in Barker’s case,¹¹ and that ‘If Jane had taken the antidote, she would have survived’ is true in Example #3.

So, TCM appears to be on the right track. It explains “backtrackers”, all of our problematic Examples #1–7 (barring the complication to come about contrastive influences), and also captures the successes of the common idea to “hold fixed the past” and let the future evolve according to actual law.

4. Context-dependence and comparison to Lewis.

Here I turn to the qualification in TCM concerning contextually relevant *Breaks*, and argue that our account of counterfactuals has some important general advantages over Lewis’s.

Here is a standard example of context-dependence:

Xenon: If you were to jump out of the window, then you would be injured.

Yvonne: No, I wouldn’t. If I were to jump, I would have put up a safety net first, and it would break my fall.

I take it that there is nothing to choose between Xenon’s and Yvonne’s favored counterfactuals. Only context and our momentary purposes can decide which if either we should prefer on a given occasion.

Fig. 5 illustrates a natural model for the case. Yvonne jumps only if either Safety Net = 1 or Rational = 0. In the actual case, Safety Net = 0 and Rational = 1, she does not jump (Jump = 0) and is not injured (Injury = 0). The qualification in TCM about contextually relevant *Breaks* is supposed to capture this case: Xenon treats {Rational} as a relevant minimal *Break*, while Yvonne treats {Safety Net} as relevant. Each of these are (Jump = 1)-minimal breaks, and they are *incomparable*: each gives up something the other keeps. The circumstances of the utterance lead Xenon and Yvonne to make different choices (in ways that are easier to represent than explain).

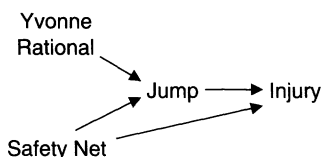


Figure 5. Context dependence

TCM is a version of Nelson Goodman's older "cotenability" approach to counterfactuals (1946), rather than Lewis's "similarity" approach. More recent versions of this approach are due to Pollock (1981) and Angelika Kratzer (1981). Lewis (1981) helpfully calls these theories "premise semantics" in contrast to his own "ordering semantics". The general idea of premise semantics is that $A > C$ is true iff C follows lawfully from A and a set of actual facts F , a "factual background" for A . These two sorts of theories are not that different in the end: roughly, we can let the "similar" A -worlds be those in which F holds. But some differences remain, and on those counts "premise semantics" is preferable. Here are two.

First, whatever it is exactly, Lewis's "similarity of worlds" is "global" and is (sometimes? often?) "approximate". By contrast, "minimality" is both local and "exact": only agreement of models on a factual background matters. So, according to Lewis's theory but not mine, approximate resemblance of worlds with respect to events not necessitated by the factual background (sometimes? often?) makes a difference for the truth conditions of counterfactuals.

But it does not. Suppose that Nixon sat watching Jane with his finger on the button, resolved to push it if she should flip the coin in her hand and it should land heads. She spent the coin without flipping it, and Nixon reluctantly put away the button. If there is *any* context in which "global" and "approximate" similarity of possible worlds counts for *anything*, there should be *some* context in which it is true 'If Jane had flipped the coin, then it would have landed tails'. But there is no such context.¹² More generally, according to TCM, there is no way that context can force a nonactual outcome of a nonactual chance process, while this should be fairly common on Lewis's view.

Second, my account of ϕ -minimal models allows for widespread *incomparability*. By contrast, Lewis's relation of "similarity" is an *ordering* of possible worlds, and is "connected": any two worlds are either more or less or equally similar to actuality (e.g., 1973 48). Barry Loewer (1979) identifies connectedness as the main purely formal difference between Goodman-style "cotenability" theories (i.e., premise semantics) and "similarity" theories (ordering semantics).

On this score, also, my account is preferable. Fig. 6 illustrates an example of John Pollock's (1981 254–5). The L_i are lights, and the S_i are switches. In

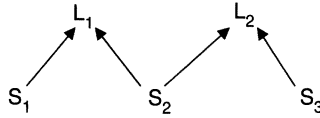


Figure 6. Pollock's case.

the actual case, the switches are open ($\neg S_i$) and the lights are off ($\neg L_i$). The switches work as follows:

$$\begin{aligned} L_1 &= S_1 \vee S_2 \\ L_2 &= S_2 \vee S_3. \end{aligned}$$

That is, a light is on iff at least one of the switches connected to it is closed.

Lewis's theory validates the following principle:

Limited Antecedent Strengthening (LAS):

- (i) $P > Q$ and
- (ii) $\neg(P > \neg R)$ entail
- (iii) $(P \ \& \ R) > Q$.

Suppose (i) and (ii) are true, and consider the most similar P -worlds. From (i), they are all Q -worlds. From (ii), some of them are R -worlds. (iii) asks us to look to the most similar $(P \ \& \ R)$ -worlds; by connectedness and (ii), those must be the subset of the most similar P -worlds from (i) in which R holds, too. But Q holds throughout the most similar P -worlds. So, (iii) is true on Lewis's account.

LAS relies on connectedness, but it is intuitively invalid. In the case of the switches, the following counterfactuals are intuitively correct.

- (i') $L_2 > \neg S_1$
- (ii') $\neg(L_2 > \neg L_1)$
- (iii') $\neg[(L_2 \ \& \ L_1) > \neg S_1]$.

(iii') is the negation of an instance of (iii), for $Q = \neg S_1$.

TCM counts each of (i')–(iii') true, and does not validate LAS. The reason is that a break at S_2 is incomparable with a break at S_1 and S_3 : if $L_1 \ \& \ L_2$ had been, then that might have been because S_2 , but might equally have been because S_1 and S_3 . Each of $\{S_2\}$ and $\{S_1, S_3\}$ are $(L_1 \ \& \ L_2)$ -minimal *Breaks*: each gives up some actual fact that the other keeps. So, TCM says it is not the case that if $(L_1 \ \& \ L_2)$, then it would have been that $\neg S_1$; but it is true that it *might* have been that $\neg S_1$.¹³

I would be happy to reject wholesale the idea that any noncausal concept of "similarity" has any bearing on counterfactuals, but it may have some minor

role to play. For example, I once heard a sports announcer say of an athlete with an average height but especially solid build, “If he were a truck, he would be a cement truck”. As far as I can tell, this sort of “similarity” comes into play only with impossible and highly fanciful antecedents. For my purposes these are marginal cases, and I set them aside (though see Daniel Nolan 1997 for a discussion of impossible antecedents). Truth conditions for these, if they have any nontrivial ones at all, rely heavily (perhaps exclusively) on metaphorical and epistemic considerations. To the extent that counterfactuals are philosophically interesting devices for reporting facts about the actual world, or analyzing causation, or doing anything else philosophers want to do with them, we should set aside such considerations.

5. Good models.

TCM is a theory of counterfactuals in causal models. My general account of counterfactuals is that they are true *simpliciter* when they are true in good models. Some complications arise in saying what a “good” model is, however.

Let’s say that a “case” is a set of events. I treat an event as an object’s (/n-tuple’s) having a property (/relation) at a time or over a duration of time. I assume that the variables in models represent properties in the first instance (property variables, really), and the equations represent lawful relations among those properties.

Good Model:

M is a good model for a case C iff

- (i) the properties represented in M are instantiated by objects in C,
- (ii) the laws of M are accurate of C, or accurate enough, and
- (iii) M is complete enough to accurately represent the causal relations among the events of C that appear in M.

This suggestion is supposed to function as follows:

TC (Theory of Counterfactuals):

$A > B$ is true of case C iff there is a good model M for C, and $A > B$ is true in M.

Truth-in-a-model is given by TCM.

Clause (i) of “Good Model” requires merely (for example) that if the model has a variable for whether a coin is flipped and it takes value 1, then the case includes a coin flip. Clauses (ii) and (iii) require further explanation. I take them in reverse order.

Clause (iii) is my attempt to give the theory of counterfactuals while avoiding the details of a full theory of causation. We did not appeal to the

concept of an “actual cause”, but instead used the substitute concept of “direct positive influence”. There are well-known cases in which causes and positive influences come apart. These pose problems for my theory of counterfactuals, and they pose problems precisely because of the divergence of my substitute concept from that of an actual cause.

Fig. 7 illustrates a case of “preemption”. Two assassins conspire to kill a king: the Master Assassin and the Bumbling Assassin. The Master Assassin hides in the closet, dagger in cloak. The Bumbling Assassin mixes poison in the king’s soup, and carries it from the kitchen. But the Bumbling Assassin really is a bumbling assassin, and he bumps the closet door as he passes by, locking the Master Assassin inside. Then he runs into a chair. He stumbles and the soup sloshes in the bowl, but he recovers, and sets the soup in front of the king, who eats it and dies shortly afterward.

TCM applied to the model of Fig. 7 gets the right answers for the case. For example, $BA = 0 > KD = 1$ is false. But we might also represent the case as in Fig. 8. Given laws similar to those we assumed for the original case, MA counts as a “direct positive influence” on KD in a model for Fig. 8, while BA does not. Thus, TCM treats the case as similar to that of the poison and antidote, Fig. 1(a), and we get the wrong answer that $BA = 0 > KD = 1$ is true.

This example illustrates both that positive influences (MA) can fail to be actual causes, and that negative influences (BA) can be. Given a model such as Fig. 8, *direct* positive influences can fail to be causes, and *direct* negative influences can be. However, it seems to me that this is possible only given models that are incomplete in crucial respects. Fig. 7 is clearly a better representation of the case in question, and in it the relations of direct influence do match the causal facts. I suggest that all cases of negative influences that are causes are cases in which the cause interacts with other influences on the effect. Those interactions occur in causally intervening processes. I suggest also that all cases of positive influences that are not actual causes are cases in which some intervening causal process does not go to completion. (Unfortunately, to argue these points would take us into the details of theory of causation.¹⁴)

Clause (iii) says, “M is complete enough to accurately represent the causal relations among the events of C that appear in M”. “Complete enough” means that the model represents cases of mixed influence as such (e.g., the multiple paths from BA to KD in Fig. 7), and includes enough variables to capture whether intervening causal processes go to completion (e.g., whether MA is locked in the closet). This tells us to use Fig. 7 rather than 8 for the case of the assassins. I suggest that these requirements suffice to solve the problems that arise from our use of the substitute concept of direct positive influence.

Clause (ii) of “Good Model” says, “(ii) the laws of M are accurate of C, or accurate enough”. This is the really difficult requirement. The issue

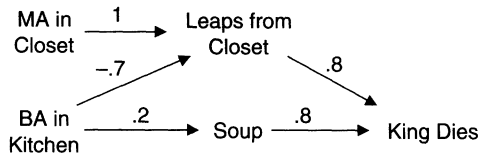


Figure 7. Preemption.

here boils down to that of saying what makes special science generalizations good (or good enough) despite the fact that they hold only under restricted boundary conditions, describe idealized cases, or are accurate only to some degrees of approximation. That is, an account here is an account of “approximate truth” or “approximate accuracy” of laws. I will not offer such an account, and as far as I know, there are no good accounts out there. It seems to me that such an account is needed by a broadly realistic view of the special sciences, whatever we think of counterfactuals; it should be no surprise that it is also needed in the theory of counterfactuals.

6. Contrastive causal influences.

Here I turn very briefly to the more complex matter, postponed from §3 (fn. 8), of providing an account of “positive parents” that is sensitive to the details of contrastive influences. The account I provide plugs into the theory TC already developed, and readers not interested in the details may wish to skip this section.

Suppose we have a simple model $P \rightarrow D$. P is dosage of poison, and it takes values 0 (no poison), 1 (low dosage), and 2 (high dosage). D is death. Suppose for the sake of illustration that the effect of the poison is (at least quasi-)indeterministic. $p(D|P = 0) = .01$, $p(D|P = 1) = .5$, $p(D|P = 2) = .99$. Fluffy the rat is given the low dosage, and survives. I take it that ‘If Fluffy had not been given any poison, then she would have survived’ is

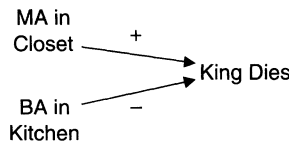


Figure 8. Preemption.

correct even though there would have been a .01 chance that she die—this is the same phenomenon as the case of the snakebite. But these are both false: ‘If Fluffy had been given the high dosage, then she would have survived’, ‘If Fluffy had been given the high dosage, then she would have died’.

We need to provide an account of positive influences that captures these contrastive facts. Suppose X is a parent of Y in M , Y ’s other parents are Z^+ , and $Z^- = z^-$ in M .

Positive influence:

$X = x$ in comparison to $X = x'$ has direct positive influence on $Y = y$ in M iff $p(Y = y|X = x \ \& \ Z^+ = z^+) > p(Y = y|X = x' \ \& \ Z^+ = z^+)$.

Let $ppa(Y)_{M, M_i} = \{X: A(X) \text{ in comparison to } A_i(X) \text{ has direct positive influence on } Y = y \text{ in } M\}$.

This account relativizes being a positive parent to a pair of models.

We replace $ppa(Y)_M$ with $ppa(Y)_{M, M_i}$ in the earlier definitions of *Break* and *Intact*, and proceed as before.

Finally we can explain the coin toss!

Fig. 9 illustrates a model M for the case. Bet takes values 0 (no bet), 1 (tails), and 2 (heads). Outcome takes similar values 0 (no outcome), 1 (tails), 2 (heads). The laws are the following:

Flip = 1 iff Bet $\neq$ 0

Outcome = 0 iff Flip = 0

Flip = 1 $\Rightarrow$ $p(\text{Outcome} = x) = .5$, for $x \in \{1, 2\}$

Win = 1 iff Outcome = Bet and Bet $\neq$ 0.

In the actual model M , Bet = 1, Flip = 1, Outcome = 2, Win = 0.

It turns out true in M that if Ben had bet heads, he would have won. The unique (Bet = 2)-minimal break is $Break(M_i, M) = \{\text{Bet}\}$. Bet = 2 does *not* have positive influence on Flip in comparison to Bet = 1, though it does in comparison to Bet = 0. So, Bet is not a positive parent of Flip in M in comparison to M_i . $\{\text{Flip}, \text{Outcome}\}$ is a maximal $Intact(M_i, M) \cap Z$.

7. Conclusion.

My aim has been to provide a reductive and actualistic account of the truth conditions of counterfactuals. The theory TC is preferable to other theories in its ability to explain cases and to provide a unified account of “back-trackers” (§3). In addition, I argued that there are general reasons to prefer the “premise semantics” approach to the “similarity” approach (§4).

The philosophical moral of TC is that reductive counterfactual analyses of causation (and other causal concepts) are mistaken. There is something

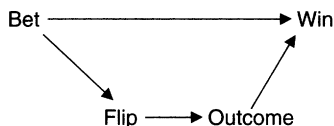


Figure 9. The coin toss.

intuitively natural about such theories and TC explains what it is: among the standard truth makers for a counterfactual of the sort ‘If C had not occurred, then E would not have occurred’ is the fact that C caused E.

I conclude by considering two objections, one to TC, and one to the philosophical moral. The first reaction of seemingly all followers of Lewis (as well as an anonymous referee) is that the *Examples #1–7* involve some nonstandard context. Something like TC may be correct for that context, they often concede, but Lewis’s theory remains correct for the standard context. This response is obviously *ad hoc*, and it also multiplies contexts beyond necessity. TC and Lewis’s theory agree on an extremely wide range of cases. Lewis’s guiding idea, which is also the guiding idea of most theories, is: hold fixed the past of the antecedent A, add some minor divergence that allows A to come true, and let the future unfold as the actual laws allow; $A > C$ is true iff C follows. My guiding idea is similar, except that we hold fixed not the *past* of the antecedent, but those actual facts that are *causally independent* of the antecedent. Given that backward causation is rare or nonexistent, my idea entails that Lewis’s will work pretty well most of the time. The two theories agree on their predictions in many cases, and where they disagree, TC is preferable. So, adding contexts in which Lewis’s theory is correct, but also agrees with TC, is simply gratuitous.

A second objection suggests that even if TC (or something like it) is correct as an account of counterfactuals as judged by the folk, for technical purposes we can introduce other types of counterfactuals for which Lewis’s (or some other) noncausal theory is correct. These counterfactuals can be used without circularity in a philosophical analysis of causation. So, even if TC is correct, this objection claims, the philosophical moral is not. I argue that this project is misguided.

Lewis introduces his counterfactual theory of causation to solve three problems for “regularity” theories (i.e., Covering Law theories, which may or may not appeal to a regularity theory of laws). The general problem, he complains, is that “regularity analyses tend to confuse causation itself with various other causal relations” (1986a 160). These include the problems of “effects” and “epiphenomena”. The problem of effects is that causes and effects are mutually associated: if A is necessary for B, then B is sufficient for A; if A is probabilistically relevant to B, then B is to A. The problem of

epiphenomena is that A and B might be associated not as cause and effect, but as joint effects of a common cause (e.g., early spots and later development of full-blown measles).

These are instances of the *fundamental problem* for reductive theories of causation: what is the difference between causal associations and others? People such as Hempel (1965) and the early Salmon (1971) struggled mightily to solve such problems (in the context of a theory of scientific explanation rather than causation).

In his 1973 paper "Causation" (1986a), Lewis proposes that counterfactuals provide the answer (160–1, 170–1). His idea is that (for example) while the appearance of spots is probabilistically associated with measles, it is *false* that if the spots *had not* appeared, then measles *would not* have occurred. Relations of similarity among possible worlds will solve the fundamental problem for reductive theories of causation.

Intuitively speaking, counterfactuals like the measles one are sometimes true but sometimes false: fill in appropriate circumstances and it seems true, 'If the building had remained standing, then the explosion would not have occurred', 'If the flash had not occurred, then the cannon would not have fired and the bang would not have occurred either'. So, in order to solve the fundamental problem, Lewis appeals to a distinction between "backtracking" counterfactuals and others. He says, "A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution. If so, call it a *back-tracking counterfactual*" (1986b 34). He continues (35) that backtrackers are true only given "nonstandard" contexts, and are false on his "standard" theory (1)–(4).

This claim is doubly mistaken: Lewis's theory (1)–(4) counts many backtrackers true, and the better theory TC counts them entirely standard. Recall the case of Xenon and Yvonne: If Yvonne had jumped, that would have been due either to her having a momentary lapse of reason or to her having set up a safety net earlier. Lewis's theory prefers the lapse of reason interpretation: it preserves a larger spatiotemporal region of perfect match with actuality. But the point is that in the case we face a choice concerning *which* backtrackers are true, and not a choice between some backtracking v. none. Lewis's theory recommends that it is true, 'If Yvonne had jumped (at t), then she would have had a lapse of reason (shortly before t)'. So, on a counterfactual theory of causation, Yvonne's not jumping at t (maybe = her standing stationary at t) causes her to be rational shortly beforehand.

Lewis and other authors standardly attempt to avoid this sort of problem by stipulating that backtrackers cannot do causal or explanatory work, even though some may be true. But this is just sleight of hand: Lewis *defined* backtrackers as ones that are false on his "standard resolution of vagueness" (1986b 34). The counterfactual about Yvonne is *not* a backtracker on that

characterization. We need to know what “backtrackers” *are* in order to set them aside, and we have been given precious little idea. We cannot give a causal characterization of “backtrackers” (e.g., consequent is a cause of antecedent, or they are effects of a common cause). This is the characterization that I have implicitly assumed, and these causal backtrackers are the ones that a counterfactual theory of causation needs to rule out as a class. But we cannot rule them out *as such*, by saying, “No causal backtrackers”. That would make the theory of causation straightforwardly circular. We might try a temporal characterization of “backtrackers” (antecedent *later than* consequent), but in the current context that rules out backward causation *a priori*, which Lewis says on multiple occasions that he does not want to do (1986a 170, 1986b Postscript D 62–3), and doesn’t help with “epiphenomena” anyway (spots precede measles, flash precedes bang). And we must give *some* account of backtrackers (*contra* a suggestion of McDermott’s¹⁵): the solution to the fundamental problem for theories of causation hangs in the balance.

In short, the appeal to counterfactuals has made no progress toward solving the fundamental problem that motivated it in the first place. Some illusion of progress can be engendered by the stipulation to set aside “backtrackers”; but in reality we need nothing short of a theory of causation to tell us what we are supposed to set aside in the first place.

The alternative diagnosis I would urge is that associations are *evidence* for causal relations. We should give up on the Humean project of reducing causation to our evidence for it. If so, then counterfactual relations are *entirely* epiphenomenal with respect to causal ones—merely hypothetical associations cannot serve as evidence. Instead, the associations are actual frequencies and conditional probabilities estimated from them. These associations provide evidence for the causal laws of our models, and those laws in turn make true whatever counterfactuals are true.

I have not said anything about indicative conditionals, but we might hope for a unified account. Brian Weatherson (2001) and Daniel Nolan (2003) defend closest-world accounts of indicatives. They assume Lewis’s account of subjunctives, but their general approach is compatible with my theory. It seems likely to me that causal models will be helpful in the account of indicatives, too.¹⁶

Notes

¹ This case is given by Michael Slote, who credits Sydney Morgenbesser (Slote 1978 27 fn. 33). The example is also mentioned in passing by Lewis (1986b 48) and discussed briefly by Jonathan Bennett (1984 76) and Michael McDermott (1999 304–5, 309, 332). Dorothy Edgington (1995 257–8) raises this case as a problem for Lewis, though it also appears to pose a problem for her own view (315): $p_t(\text{Win}|\text{Bet heads at } t) = .5$.

² In addition to Lewis, Jonathan Bennett and Michael McDermott make half-hearted attempts at explaining the case (Bennett 1984 76, McDermott 1999 310–3, 332). None of these authors provides an explanation for the difference between Examples #1 and #4. Bennett counts the counterfactual true in both cases. McDermott relies on unexplained “brute modal facts” to distinguish them (311–2), and he relies on our intuitions about counterfactuals to tell us what those “brute” facts are in any given case.

³ An anonymous referee suggests that despite my arguments here, some noncausal account of the *Examples* may still remain open. This is true: I cannot canvas all possible theories. I claim that my diagnosis is the only plausible one currently on offer.

⁴ Barker focuses on the following counterfactual rather than the one I mention (1999 432): “If Jane had bet on snake-eyes [B], it would have been 95% likely, given how things in fact turned out, that she would have won [W]”. Read literally, this turns out false on the account I will offer: there is no relevant situation in which the probability of winning is .95. Unfortunately, I cannot address the probabilities of conditionals in any detail here. However, it will turn out that the probability of W conditional on B and the “factual background” for B is .95. We might be able to exploit this fact to give some explanation of the plausibility of Barker’s conditional.

⁵ I take the variables to represent properties in the first instance. But I treat an event as an object’s having a property at a time, so the variables can equally represent a particular case (events) or a repeatable set-up (properties). The variables *must* represent properties in the first instance if the graphs are capable of representing multiple situations: individual events (such as Jane’s being bitten by the snake at a given time and place) occur only once.

⁶ The Causal Markov Condition (CMC) is that if the values of variable Y depend probabilistically on the values of X while holding fixed X’s parents, then X causally influences Y (e.g., SGS 11). That is, if $p(Y|pa(X)) \neq p(Y|X \& pa(X))$, then X influences Y. (Faithfulness is the converse.) CMC is inconsistent with some sorts of indeterminism. See, e.g., Nancy Cartwright 1998 for more discussion. [Added in proof: Daniel Steel argues in a recent paper (“Indeterminism and the Causal Markov Condition”, *British Journal for the Philosophy of Science* 56 (2005), 3–26) that CMC is perfectly compatible with determinism. I suspect that the issue here involves one’s conception of probability, though much more discussion would be needed to justify my claim in the text.]

⁷ That is, on the account to come, there are no “miracles”, and could be no “miracles”. Also, I will not be concerned with “counterlegals” or other impossible antecedents (though see Daniel Nolan 1997 for a discussion of these). Such counterfactuals will all turn out trivially true (i.e., true in all minimal ϕ -models, namely none).

⁸ The account here is only a first approximation because relations of positive/negative influence are *contrastive*. This matters to the case of the fair coin toss. Suppose Alice will flip the coin only if Ben bets something or other. Then Ben’s betting tails *rather than betting heads* does not influence the outcome, while his betting tails *rather than not betting at all* does. The first approximation account works for simple cases, and I give a more complex account in §6 that plugs into the theory developed here.

⁹ Disjunctive antecedents pose problems for all theories of counterfactuals, and I do not have anything to add concerning them. See, e.g., Barry Loewer 1976.

¹⁰ (ii) is a bit more complicated than my “rough” statement indicates, and the complication is required. It does not work simply to maximize *agreement of values* among variables not influenced by ϕ . Consider a simple model $A \leftarrow B \rightarrow C$. All variables have value 1. The equations include (1) $A = B$, (2) $(B = 0) \Rightarrow p(C = 1) = .05$, and (3) $(B = 1) \Rightarrow p(C = 1) = .95$. Consider the counterfactual antecedent that $A = 0$. (1) requires that B would be 0. (2) then *allows* that C have its actual value 1. So, a requirement to maximize agreement of values among nondescendants of A would recommend that $A = 0 > C = 1$ is true. But it is not. C is a nondescendant of A, but C is *not* included in any *Intact* set among A-models. In models in which $B = 0$, the positive parent B of C has a nonactual value; so, whatever C’s

value, it is not in any *Intact* set. Unlike my rough statement in the text, clause (ii) recommends that $A = 0 > C = 1$ is false.

¹¹ In addition, in Barker's case $p(S|B \& F) = .95$, for F the "factual background". This fact might give us some explanation for Barker's intuition about "If Jane had bet on snake-eyes, it would have been 95% likely . . . that she would have won". .95 is not the probability of the conditional (according to TCM), but a closely related conditional probability.

¹² Of course we can force the selection of worlds in which the coin lands tails by explicit stipulation: 'If Jane had flipped the coin *and it had landed tails*, then it would have landed tails'. That is not a different *context*, but a different *antecedent*, and one that positively entails that the coin lands tails. Also, we could make the original counterfactual true by changing the facts of the case (e.g., Mandrake was watching with his magnets, dedicated to ensuring that any coin flip lands tails). Again, that is not a different *context*, but a different *case*, and a case in which the outcome would not be the result of a chance process.

¹³ Context could potentially select a break at S_2 as especially relevant, and thereby make true $(L_1 \& L_2) > \neg S_1$; but we have been given no reason in the case to make such a selection.

¹⁴ For example, an anonymous referee notes that Jonathan Schaffer's "trumping preemption" (2000) appears to violate my suggestions in the text. I do not think the case is a genuine counterexample, though I cannot address the issue here.

¹⁵ McDermott 1995 considers a pair of counterfactuals, one of which is a backtracker $A > C$, and says, "We do not need to endorse all that [i.e., Lewis's account of backtrackers]. All we need accept is that a counterfactual like $[A > C]$ has two reasonable interpretations, one on which it is true, and one on which it is false. We can then stipulate that the reading . . . used in the analysis of causal statements is the latter" (1995 524 fn. 4). McDermott seems to think we can identify backtrackers by ostension (interpretations like the one that makes $A > C$ false). Of course, if TC is correct, there are no semantically different interpretations to pick out in this way. But in any case McDermott seems to think there is no need for an explicit account.

¹⁶ I would like to thank Richard Boyd, Daniel Nolan, Sydney Shoemaker, Jason Stanley, and Zoltan Gendler-Szabo for comments and discussion of previous drafts.

References

- Barker, Stephen (1999). "Counterfactuals, Probabilistic Counterfactuals, and Causation". *Mind* 108, 427–69.
- Bennett, Jonathan (1984). "Counterfactuals and Temporal Direction". *Philosophical Review* 93, 57–91.
- Cartwright, Nancy (1998). "Causality: Independence and Determinism". In A. Gammerman ed., *Causal Models and Intelligent Data Management*. Berlin: Springer. 51–63.
- Edgington, Dorothy (1995). "On Conditionals". *Mind* 104, 235–329.
- Fine, Kit (1975). Review of Lewis (1973). *Mind* 84, 451–8.
- Giere, Ronald (1985). "Constructive Realism". In Paul Churchland and Clifford Hooker, *Images of Science*. Chicago: Chicago University Press. 75–98.
- Goodman, Nelson (1946). "The Problem of Counterfactual Conditionals". Reprinted in Goodman, *Fact, Fiction, and Forecast*, 4th ed. Cambridge: Harvard, 1979.
- Hempel, Carl G. (1965). "Aspects of Scientific Explanation". In Hempel, *Aspects of Scientific Explanation*. New York: The Free Press. 331–496.
- Hitchcock, Christopher (2001). "The Intransitivity of Causation Revealed in Equations and Graphs". *Journal of Philosophy* 98, 273–99.
- Kvart, Igal (1986). *A Theory of Counterfactuals*. Indianapolis: Hackett.
- Kvart, Igal (1991). "Transitivity and Preemption of Causal Relevance". *Philosophical Studies* 64, 125–60.
- Kvart, Igal (1994). "Causal Independence". *Philosophy of Science* 61, 96–114.

- Kvart, Igal (1997). "Cause and Some Positive Causal Impact". In James Tomberlin, ed., *Philosophical Perspectives* 11, 401–32.
- Lewis, David (1973). *Counterfactuals*. Cambridge: Harvard University Press.
- Lewis, David (1981). "Ordering Semantics and Premise Semantics for Counterfactuals". *Journal of Philosophical Logic* 10, 217–34.
- Lewis, David (1986). *Philosophical Papers*, Volume II. Oxford: Oxford University Press.
- Lewis, David (1986a). "Causation". In Lewis 1986, 159–72.
- Lewis, David (1986b). "Counterfactual Dependence and Time's Arrow". In Lewis 1986, 32–52.
- Loewer, Barry M. (1976). "Counterfactuals with Disjunctive Antecedents". *Journal of Philosophy* 73, 531–37.
- Loewer, Barry M. (1979). "Cotenability and Counterfactual Logics". *Journal of Philosophical Logic* 8, 99–115.
- McDermott, Michael (1995). "Redundant Causation". *British Journal for the Philosophy of Science* XLVI, 523–44.
- McDermott, Michael (1999). "Counterfactuals and Access Points". *Mind* 108, 291–334.
- Nolan, Daniel (1997). "Impossible Worlds: A Modest Approach". *Notre Dame Journal of Formal Logic* 38, 525–72.
- Nolan, Daniel (2003). "Defending a Possible-Worlds Account of Indicative Conditionals". *Philosophical Studies* 116, 215–69.
- Pearl, Judea (1999). "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification". *Synthese* 121, 93–149.
- Pearl, Judea (2000). *Causality*. Cambridge: Cambridge University Press.
- Pollock, John (1981). "A Refined Theory of Counterfactuals". *Journal of Philosophical Logic* 10, 239–266.
- Salmon, Wesley (1971). "Statistical Explanation". In Salmon, *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, 1971. 29–87.
- Schaffer, Jonathan (2000). 'Trumping Preemption'. *The Journal of Philosophy*, XCVII, 165–81.
- Slote, Michael (1978). "Time in Counterfactuals". *Philosophical Review* 87, 3–27.
- Spiro, Peter, Clark Glymour, and Richard Scheines (2000). *Causation, Prediction, and Search*, 2nd ed., Cambridge: MIT Press.
- Stalnaker, Robert (1968). "A Theory of Conditionals". In N. Rescher ed., *Studies in Logical Theory*. Oxford: Basil Blackwell, 98–112.
- Stalnaker, Robert, and Thomason, Richmond (1970). "A Semantic Analysis of Conditional Logic". *Theoria* 36, 23–42.
- Weatherston, Brian (2001). "Indicatives and Subjunctives". *Philosophical Quarterly* 51, 200–216.