Counterfactual Analyses of Causation: the Problem of Effects and Epiphenomena Revisited

0. Introduction

Nearly all recent discussion of counterfactual analyses of causation have focussed on redundant causation: various forms of preemption, chance lowering cause and trumping.[1] Few have attended to the problem of effects or the problem of epiphenomena. These problems are about ensuring that the counterfactual analysis does not certify that event e causes event c where either the causation is in the opposite direction – the problem of effects – or e and c have a common cause but neither causes the other – the problem of epiphenomena. It might seem that apparent cases of counterfactual dependency of causes on effects, or one epiphenomenal event on another, which certify bogus judgments can be explained away either by attending closely to the character of counterfactual dependence – see Lewis 1973, 1979 – or by being clear about what we mean by an event – see Lewis 1973. I disagree. This paper develops the view that the problem of effects and epiphenomena present the counterfactual analysis with a real dilemma. The problem in a nutshell is that removal of the event c from the actual world never occurs without some further disturbance of the actual world. These disturbances harbor the possibility of counterfactual dependencies that entail bogus causation. In eradicating the counterfactual dependencies that entail bogus causation the counterfactual analysis invariably undermines its capacities to explain real causation elsewhere. This dilemma is not specific to Lewis's theories (1973, 1986, 2000). It turns up across the board.[2]

1. Effects

The counterfactual analysis proposes that judgements of causation are informed by judgements of counterfactual dependency between events. The classical counterfactual analysis is in terms of would-counterfactuals – counterfactuals of the form: if P had been the case, Q would have been, abbreviated (P > Q). The classical analysis accepts the following thesis:

> **Sufficiency**: For all c and e where c and e occur, where it is the case that if c had not occurred, e would not have occurred, then c caused e.

Note, however, that the counterfactuals relevant to the causal judgments are of a specific kind, so called forwardtracking counterfactuals.[3] I say more about their semantics below.

Counterfactual dependency is not meant to be necessary for causation, but it is meant to be sufficient. There are, however, prima facie counterexamples to **Sufficiency**. I look now at several kinds of cases of this, which will enable us to more clearly grasp the issues at stake.

The first kind of case is Flichman's (1997) discussion of Lewis' (1973b) treatment of his own, Lewis', barometer example. The air pressure causes the barometer to have a certain reading and not vice versa. Lewis claims that the counterfactual 'If the barometer had not read 1000mb the pressure would not have been 1000mb' is false, because in the relevant closest-worlds the barometer is simply dysfunctional. But that seems to commit us to the following counterfactual, understood as a fowardtracker: 'If the barometer had not read 1000mb then it would not have been that the barometer worked'. By **Sufficiency**, the counterfactual, apparently, commits us to the barometer's reading 1000mb causing it to function properly, which is false. The right means of escape for Lewis is supplied by Helen Beebee (1997). She points out that for Lewis, events are predominantly intrinsic properties of spatio-temporal regions. 'The barometer works' does not denote any

event in this sense, since dispositional properties are too extrinsic to be included as components of events *qua* relata of causation. Thus barometer counterfactual is true but nothing follows about causation. The moral then is that a necessary condition on admissible events c and e in causation is:

[A] c and e are not essentially disjunctive conditions or dispositional states, or more generally, events whose essential descriptions involve overly extrinsic properties.

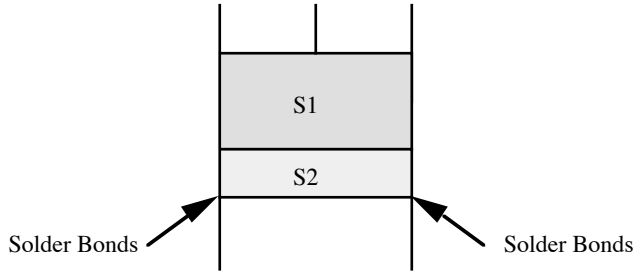
A second kind of case is illustrated by Bennett's (1984, pp. 79-81) discussion. Say Smith was elected in 1999 and is President in 2000. We do not think her being President in 2000 caused her being elected in 1999, but we have the counterfactual, 'If Smith had not been President in 2000, she would not have been elected in 1999', which looks like a true forwardtracker. In this case, the correct reply is not so much that the events mentioned are too extrinsic, but that they are not distinct from each other in the required sense – see Lewis (1986b: 212-213, 263-264). If c's essential characterization entails some component of e's then c and e are simply not candidates for causation, strictly understood. Being a president is partly defined by having been elected. So we are disinclined to say that Smith's being elected caused her to be President. Thus we can accept the truth of the counterfactual above on a forwardtracking reading without commitment, through **Sufficiency**, to backwards causation by holding that the events are not distinct. Thus, a necessary condition on admissible events c and e in causation is:

[B] c and e are distinct from each other.

One might have faith that all the apparent counterexamples to **Sufficiency** can be dealt with by pointing out that either [A] or [B] fail to be met. Not so. I now describe a case in which [A] and [B] are met, but there is still apparent commitment to bogus backwards causation. I argue that the problem here is with the semantic character of the forwardtracking reading of counterfactuals.

To do that we must briefly look at the forwardtracking reading. There is only one developed account of forwardtracking counterfactuals in the literature: Lewis' (1979) theory. For Lewis, a would-counterfactual (P > Q) is true just in case the P-worlds – worlds in which P is the case – that are closest (most similar) to the actual world @ are Q-worlds. The forwardtracking reading is fixed by a set of weighted aspects of comparative similarity – see Lewis (1979). Given these weightings, P-worlds that maximize perfect match but minimize miraculousness are deemed closest to @; we seek to minimize miracles even if at the expense of a little perfect matching of particular fact. I call this principle of balance: Min-Miracle and the P-worlds it produces Min-Miracle worlds.

With this semantic structure in mind, consider the following case. A lead cylindrical slab S1 is supported by a copper wire. S1 rests barely on an iron slab S2. Both are within a metal cylinder. S2 is supported by strong solder bonds at the bottom and the cylinder itself, as below:



At midnight S1's wire breaks, and bears down upon S2. The solder bonds break and S2 descends 20 cm as does S1. S's descending causes S2's descent. I note that the conditions [A] and [B] are met here; the events of S1's descending 20 cm and S2's descending 20cm are intrinsic properties of regions and are distinct from each other. Our acceptance that S1's descending caused S2's descent shows this. If so we can legitimately ask if the causation operates in the opposite direction. We judge that S2's descent does not cause S1's descent.[4] Nevertheless, a case can be made that (1) below is true:

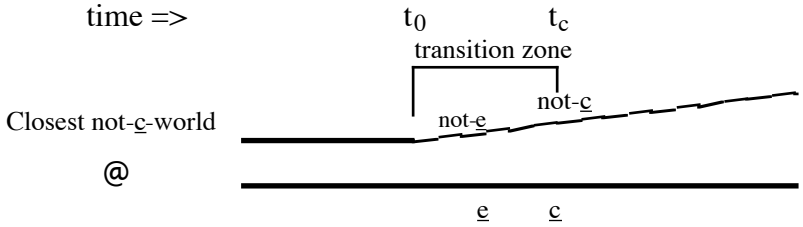
(1) If S2 had not descended 20 cm, S1 would not have.

(1) looks true if we assume that the closest (S2 does not descend)-worlds are Min-Miracle worlds. These are worlds, presumably, where a small miracle occurs in which the wire does not break, S1 remains in position and so does S2. What (S2 does not descend)-worlds are there, that are not (S2 does not descend)-worlds? One set are worlds where the wire breaks as it does in @, and S1 commences its move down, but instead of forcing S2 down, breaking the solder bonds as it does, it passes right through S1. This interpenetration-world has marginally more perfect match than the Min-Miracle worlds, but at the cost of a significant miracle, one involving millions of law violations. None of these miracles are big in Lewis' (1979) technical sense – call these BIG from now on. That is, they are not wide, diverse violations of law characteristic of *convergence miracles*, i.e., miracles that re-instate perfect match after an initial divergence point. Either way, these worlds are ruled out as closest by Min-Miracle since they involve significantly larger miracles but only marginally more perfect match. Are there any other (S2 does not descend)-worlds in which S1 descends? Aren't there worlds where S2 just catches on something and fails to descend as a result? Sure! There might be such worlds, but they are either ones where S1 does not descend or ones where S1 interpenetrates with S2. In sum, given the intimate relation between S1 and S2, the only way for S2 not to descend but for S1 to continue to do so is interpentration: the production of a massive miracle.

If this is right, by Min-Miracle, the closest (S2 does not descend)-worlds, are those in which S1 does not descend. (1) then is a true forwardtracker. Therefore, we have backwards causation where there is none by **Sufficiency**.[5]

How has the counterfactual analysis been brought to this impasse? Min-Miracle entails that for a given macroscopic event c, the closest (c does not occur)-worlds will involve divergence at a time t_0 a little before the reference time of c, t_c, where at most a small miracle will occur, with lawful development of the world thereafter and thus no subsequent perfect match.[6] Call the t_0-t_c time period, *the*

transition zone. The worry is that in this zone, there will be an events e that fails to occur in all the closest (c does not occur)-worlds. The situation I am describing is illustrated below:



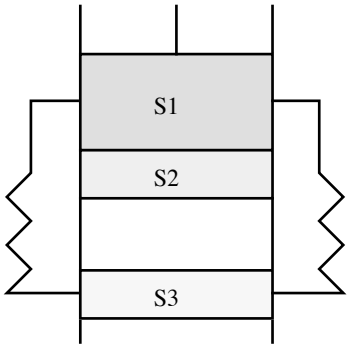
One might be tempted to think this will not occur for the following reason. That any macroscopic event c, c will always have a set F of causal factors in its immediate past, where F has at least two members and such members are counterfactually independent of each other. That the removal of no particular member of F will produce a (c-doesn't occur)-world that is a clear winner with respect to comparative closeness of (c-doesn't occur)-worlds to the actual. Tradeoffs between miracle size and perfect match will just preclude straight out winners.[7] But the example I have just given shows that this is not the case. The immediate causal factors in S2's descending are S1's descending and the laws of physics. We can either allow S1 to descend or violate the laws.

2. Epiphenomena

The question now is how to eradicate the unwanted commitment. A tempting response at this stage is to modify **Sufficiency**. A temporal restriction looks like the obvious way. Counterfactual dependency of an event e on c is sufficient for c's causing e just in case the time of c, t_c, is prior to t_e, otherwise not. If we restrict **Sufficiency** in this way, the truth of (1) brings with it no commitment to causation.

One problem with this idea is that it just rules out by fiat the conceptual possibility of backwards causation, but there seems to be nothing conceptually wrong with backwards causation. The other problem is that the temporal restriction

does not filter out all the cases of bogus causation generated by counterfactual dependency, since there are cases where e is counterfactually dependent on c and t_c is prior to t_e, but c does not cause e. These are cases where c and e are effects of a common cause; c and e are epiphenomena. Take this variation of the Slab-case. S1 is attached by arms down through, another slab, S3, as below:



If S1 descends, S3 will descend, but due to springs in the rod, its descent is delayed. Suppose in fact, that S1 descends pushing down S2, and S3 a little later. S1's descent causes S2's descend and S3's descent, but S2's descent does not cause S3's descent. S2's and S3's descent have a common cause: S1's descent. But assume Min-Miracle and **Sufficiency** in its temporally restricted form. We have:

> (2) If S2 had not descended, S3 would not have descended a few moments after.

(2) is true, if (1) is. The nearest, (S2 does not descend)-worlds are ones where S1 does not descend and in which, consequently, S3 does not descend. (I note that (S2 does not descend)-worlds in which S1 does not descend, but S3 moves down spontaneously involve further miracles with no added perfect match.) (2) is a true forwardtracker. Given the temporally restricted **Sufficiency**, (2) entails that S2's descending caused S3's descend. That's false.

It seems that the counterfactual analysis is not salvageable by mere temporal restrictions on **Sufficiency**. Another idea is to deny that the counterfactual

dependency of an event e on c is ever sufficient for c's causing e. But it is clear enough that if we are to retain the classical counterfactual analysis, which is couched in terms of would-counterfactual dependency, we need to retain **Sufficiency** in some form.[8] I look at variations of the classical analysis, —probabilistic and agency accounts—below. For now, let us see what can be done with the classical analysis.

3. Max-Match

If neither restricting **Sufficiency** nor judicious application of [A] and [B] provides a solution, the way out would appear to be some kind of modification of the comparative similarity metric that delivers the result that the closest (S2 does not descend)-worlds in our examples include some (S1 descends)-worlds, which is to say that they include worlds in which there is interpenetration of S1 with S2. That means rejecting Min-Miracle and replacing it with another similarity metric. We saw above that the reason that the interpenetration worlds were not in the class of closest worlds, given Min-Miracle, was that they offered marginal perfect match increase at the expense of major miracles – though not BIG miracles, viz., miracles with great qualitative largeness and scatteredness. It seems we need nothing less than the following principle, which I call Max-Match:

> Max-Match: Increasing perfect match, even marginally, produces a world just as similar to @ as a Min-Miracle world as long as the miracles involved are not BIG.[9]

With Max-Match (1) and (2) are false, since worlds with increased perfect match, but significantly greater miracles, like interpenetration, are just as similar as worlds in which small miracles in the wire occur, but with a few instants less perfect match.

It is no objection to the counterfactual analysis that this idea of tradeoff between increased perfect match and miracle size captures no antecedent intuition we have about similarity of possible worlds. As Lewis makes clear, determination of

the similarity measure for forward trackers is an entirely empirical affair guided only by the goal of grounding the correct counterfactual, and ultimately causal, judgments that we make.

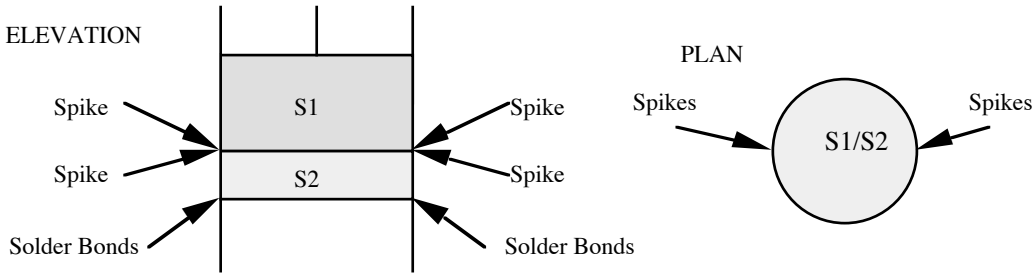
We can state fairly clearly the conditions under which we are forced to accept Max-Match to avoid bogus backwards causation judgments that c causes e. These are cases where e causes c and the only positive immediate antecedent for c is e. The other causal factors for c, leaving aside the operation of laws, are negative ones or, none at all – as seems to be the case in the Slab-example just discussed. Here the only way to secure not-c-worlds with e is to allow massive law violation but argue that marginal increase in perfect match compensates for it. (Matters hold similarly for bogus causation in the case of epiphenomena.)

Have we then provided a solution to the problem of effects and epiphenomena? We have not. This is so for several reasons. First, we have good evidence that Max-Match does not capture the forwardtracking reading. Accepting Max-Match, the rationale it provides for why we, ordinary speakers, reject (1) is, as a piece of reasoning, very suspect. That reasoning is:

> **Arg**: (1) is false, since the claim that if S2 had not descended, S1 would not have is false. If S2 had not descended, it might have been because S1 passed right through it.

Whatever our theory of the similarity metric governing counterfactuals is, it is reasonable to expect that the arguments it predicts to be sound, are ones to which we are willing to subscribe. But **Arg** is no such argument. Do we happily accept the might-conditional, which is the negation of the would-counterfactual (1), in **Arg**? I think not, if we hold firmly in our minds that interpenetration is physically impossible. Not only does **Arg** look suspect in itself, the counterfactual analysis must propose that our unhesitating claim that there is no backwards causation in the example is based on our willingness to produce **Arg**. That looks very implausible.

The second reason is that accepting Max-Match undermines our ability to explain the causal judgements we actually make. Add to the slab-case the following twist. As above, a lead cylindrical slab S1 is supported by a copper wire. S1 rests barely on an iron slab S2. Both are within a metal cylinder. S2 is supported by strong solder bonds at the bottom and the cylinder itself. S2 has two little sharp steel protuberances at its top which, if S2 is lowered, will scratch the inside of the cylinder. S1 has the same little steel protuberances at its bottom in exactly the same position. The set up is represented below:



An instant before 2:00 am, the wire, due to tension, breaks at its weakest point where it joins S1. S1 then bears down on S2 causing S2 to descend at 2:00 am and the solder bonds to break. With S2's descending the spikes on S2 scratch the cylinder's inside. S1 and S2 descend and the cylinder is scratched. We accept then (3) but reject (4):

(3) S2's descending caused the cylinder to be scratched.

(4) S2's descending caused S1 to descend.

The dilemma for the counterfactual analysis is that it wants to explain both judgments by appeal to our grasp of conditions of counterfactual dependency. But it appears it has no way of explaining our judgment (3) without committing us erroneously to the negation of (4). First, note that conditions [A] and [B] on events hold here with respect to the three events denoted in (3) and (4). We deny (4) not because either of [A] or [B] fail to be met, but because the first event failed to cause the second.

If we accept Min-Miracle we can explain our judgement (3) as being one based on the truth of the counterfactual (5),

(5) If S2 had not descended, the cylinder would not have been scratched.

and our acceptance of **Sufficiency**. But we are also committed to the bogus judgement (4) by **Sufficiency** and (1), which, as we saw above, is true given Min-Miracle:

(1) If S2 had not descended, S1 would not have descended.

On the other hand, accepting Max-Match to cast off commitment to (1) and (4) leaves us, I submit, with no explanation of why we make the correct judgement (3). This takes some explaining, but first note that (5) is no longer true, given Max-Match, and so the explanation for our judgement (3) cannot be that (5) is true and **Sufficiency** is correct. (5) is not true, since, given Max-match, the closest (S2 does not descend)-worlds also include worlds like this:

w1: There is perfect match with @ up to the wire's breaking and S1 begins to descend. But an instant after, a significant miracle occurs. S2 does not descend because S1 passes right through S2. S1 descends but, as it does so, it scratches the cylinder in exactly the same places it was scratched in by S2 in @.

w2: Same as w1 but no scratching.

Worlds w1 and w2 tie with respect to closeness to @. If, say, the miracle of interpenetration involved the atomic forces in S2 failing to operate, S1 would have scratched the cylinder. If it involved the atomic forces in S1 failing to operator, then perhaps it would not have. Or perhaps, in this second case, once S1's spikes cleared S2's, it would then have functioned as steel and scratched the cylinder. And so on.

If w1 and w2 are included amongst the most similar worlds, then both (5) is false. What then is the basis of our judgement that (3) is the case? I am not suggesting in posing this question that counterfactual dependency is necessary for causation. It is rather that causal judgements are meant to be based on conditions

expressed in terms of counterfactual dependencies of some kind, and that if (5) is not the dependency that grounds (3) there is no other. My justification for this claim is that a search for potential grounds for (3) reveal (5) as the only one.

If we assume Max-Match, then the present example looks like a case of preemption. If S2 had not descended then the pre-empted cause, S1's descending, might have scratched the cylinder. So what solution to preemption can be brought in to solve the problem? Lewis's 1973 theory affirms that cases of preemption, the causal judgement is based on recognition that there is a chain of dependencies involving events intermediate between cause and effect. But in the present case, there is no such chain of intermediate events.

Another type of analysis for preemption is Lauri Paul's (1998) suggestion that time delay is a sign of cause. Thus, c causes e if and only if were c not to have occurred, e would not have or would have occurred later. We might argue that this condition is met under the circumstances since the scratching by S1 would have been later than it did in the actual world, albeit, by an instant:

> (6) If S2 had not descended, the cylinder either would not have been scratched or would have been scratched later.

But (6) is false, since the cylinder might not have been scratched; the worlds w1 and w2 are equally close.[10]

Lewis's (2000) causation-as-influence theory can be seen as a generalization of Paul's idea. Roughly, an event c influences and event e just in case, were c to have occurred earlier, later, in a different manner, etc., e would have occurred earlier, later or in a different manner, or not at all. In short, dependence of when, how and whether an event e occurs on when, how and whether another event c occurs is sufficient for c's causing e. Assuming Max-Match the contention is that the basis of (3) is not whether on whether influence, there being none, but some other kind of

influence. One might think that there is an influence of time. For times x of an appropriate domain prior to 2:00 am we have:

(7) If S2 had descended at x, the cylinder would have been scratched at x.

The instance of (7) for x = 2:00 am is vacuously true – given true antecedents and consequents. Assuming Max-Match, the other cases are non-vacuously true. Thus, as things stand, the influence proposal seems to work as a response to the Slab-case as described. The basis of the causal judgment (3), that S2's descending caused the cylinder to be scratched, must be temporal influence.

In fact, this influence of time cannot be the basis of the causal judgment (3) since it is an accidental feature of the slab-example. We can provide a slight variation of the case in which there is no such temporal influence but (3) is still true. Say that the spikes protruding from S1 and S2 are normally sheathed within the slabs; they do not protrude sufficiently to scratch the cylinder. But in fact they are designed to unsheathe when the wire breaks. In the actual world, the wire breaks driving down S1 and S2 and the spikes unsheathe and so the cylinder is scratched as a result of S2's descending. But now consider the relevant instances of (7). For any t prior to the time at which the wire breaks, the closest (S2 descends at t)-worlds will include worlds which preserve more perfect match with the actual by not having the wire break leading to S2 descending later, but by simply having the solder bonds that hold up S2 break. S2 descends but there is no scratching because there is no wire breaking. Similarly for times t after the actual wire breaking time, we can assume that some (S2 descends at t)-worlds will involve wire breaking and some not. Either way, instances of (7) do not come out true, and so there is no temporal influence, but (3) is true. Is there some other form of influence in the slab-example? There is not. This last variation, just confirms the suspicion that influence is too extrinsic a phenomenon to count grounds for causation.

So we are still looking for the basis of the judgment (4). I suggest, though I cannot prove this, that the only ground we are going to produce is (5), i.e., a counterfactual ruled out by Max-Match.[11] If so, the counterfactual analysis faces a dilemma. To explain our judgement (3) we need Min-Miracle, but that commits us, by **Sufficiency**, to (4), that is, bogus backwards causation. To dissolve commitment to (4), we need Max-Match, but that leaves us unable to explain our judgement (3).[12] I conclude that adopting Max-Match is not way out of the problem.

4. Probability

The state of play is that the counterfactual analysis of causation is unable to dispense with **Sufficiency** or modify the similarity metric. So where does the counterfactual analysis go from here? What I now consider are versions of the counterfactual analysis that diverge from the classical analysis in terms of would-counterfactual dependency. I examine two approaches: one that invokes the concept of objective probability and the other that invokes the notion of agency. I argue that both these approaches fail to provide us with any solution to the transition zone problems.

In this section, I first show how chances, or single case objective probabilities, have been introduced to cope with certain kinds of indeterministic causation, and then see if the resulting probabilistic account helps the counterfactual analysis with bogus forms of causation. What I show is that even before we get to the cases discussed above, the probabilistic analysis already faces bogus backwards causation commitments.

Suppose that I struck a match at 12:00 am and it lit, but that matches have at each moment of their existence a low level residual objective probability of spontaneously lighting in the next instant. It is then false that if I had not struck the match at 12:00 am, it would not have lit a moment later since it might have spontaneously lit. Would-dependency, even the dependencies of influence, do not hold here. However, chance raising comes to the rescue. In the actual world, the

chance of the match's lighting at 12:00 am is given by two components, the component based upon my striking it and the component based on potential for spontaneous lighting. The overall chance in the actual world at 12:00 am of the match lighting is higher than the chance, at 12:00 am, of lighting in the nearest possible worlds in which I do not strike the match, where this latter chance is based only on the background chance of spontaneous lighting. Thus, the actual chance is higher than the counterfactual chance. This chance raising – so the thought goes – is the ground for the causal judgment.

I have presented the probabilistic analysis as dealing with counterfactual probabilities defined in terms of what would be the chance of the effect e, had the cause c not occurred. But in general, this is what it cannot do. There are many cases where there is no definite answer to what the probability of e would have been had c not occurred, since in the closest not-c-worlds the chances at t_c of e are slightly different across worlds. Instead of considering the chance of e at t_c in all the closest not-c-worlds, we need to inquire after the chance in some of the closest not-c-worlds. That means specifying counterfactual chances in terms of might-conditionals, counterfactuals about what the chance might have been. Taking on this idea then, the basic sufficiency condition in the probabilistic analysis is the following – more or less as Lewis (1979 and postscripts) fashions it:

> **Prob-Sufficiency**: For all occurring events c and e, c caused e if [a]-[c] hold:
> [a] the chance at t_c of e is ω,
> [b] the highest value that the chance at t_c of e reaches in the closest not-c-world is θ,
> [c] ω is significantly larger than θ.

I have referred in [b] to the highest value that the chance reaches, but we could also develop an account which concerned itself with the lowest value that the chance reaches. I consider both options below.

I now give an example illustrating why indeed we need to move from would-counterfactuals to might-counterfactuals as **Prob-Sufficiency** proposes, and then I show how this is fatal to the probabilistic analysis. Let us assume Min-Miracle, as defined in §1. (In what follows, nothing crucial depends upon this assumption.) Say I strike a match M at 12:00 am and it has a residual chance of lighting. But suppose also that the following, somewhat artificial conditions hold. In striking matches I tend not to be in an antecedent brain state Π. Suppose that in the nearest (I do not strike)-worlds, some of the orderly transitions to my not striking the match at 12:00 am are ones where I am in Π. Suppose further that in some of these cases, the obtaining of Π leads to a change in the background conditions f for the residual probability of M spontaneously lighting. Say that Π is causally linked to a machine that increases temperature in the room. If so, there is no unique chance of M's lighting that holds at all the closest (not struck M)-worlds. However, we can still speak of the highest, or alternatively, the lowest, value that this probability gets to across the closest such worlds.

Unfortunately, admitting the need to introduce might-conditionals rather than would-conditionals into the probabilistic analysis is fatal for the account. Suppose that the basic sufficiency condition for causation is given as in **Prob-Sufficiency**; we are concerned with the highest probability of e in the closest not-c-worlds. Suppose now, as may very well be possible, the following probabilities hold whilst it remains the case that my striking M at 12:00 am caused it to light:

(i) In the actual world, the probability at 12:00 of M's lighting by striking is 60% and the residual probability of spontaneous lighting is 5%. So the actual world chance at 12:00 am of (M lights) is $(1 - ((1 - 60\%) \times (1 - 5\%))) = 62\%$,

(ii) The residual probability at 12:00 am of M's spontaneous lighting in one of the (M is not struck)-worlds due to the presence of Π is 90%. This is the highest counterfactual probability at 12:00 am of M's lighting.

Given (i) and (ii), **Prob-Sufficiency** does not certify this as a case of causation, since the worldly chance is lower than the highest counterfactual chance. Nevertheless, it is true that my striking M caused it to light. What then is the basis for the causal judgment? It is not any chance raising counterfactual or, indeed, any would-counterfactual.

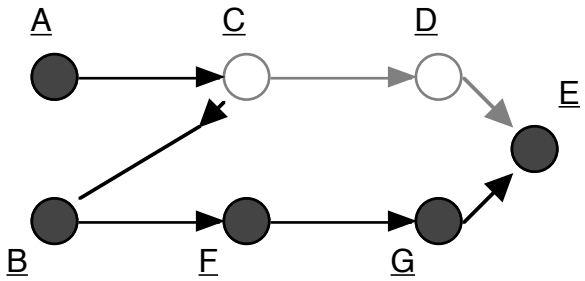
Perhaps it will be thought that the problem here is that we have expressed **Prob-Sufficiency** in terms of the highest chances of e in the closest not-c-worlds instead of the lowest. Some of the closest (M is not struck)-worlds include worlds in which Π does not obtain. In these worlds the residual probability of M's lighting is as in the actual world, low, 5%. So formulating **Prob-Sufficiency** in terms of lowest chance values will deliver the causal judgment we want: striking M caused it to light. But formulating **Prob-Sufficiency** in this way will not work in general. Say that my clicking my fingers at 12:00 am had nothing to do with M's lighting a little later. Instead it was struck by Fred at 12:00 am and had a very high probability, 99%, at 12:00 am, of lighting a little later. But say that in some of the closest (I did not click my fingers)-worlds the orderly transition to my not striking the match is one in which I am in brain state Π some instants before 12:00 am. Suppose that being in Π causes, on this particular occasion, by some circuitous but highly reliable route, a sprinkler system to spray M at 12:00 am, making the probability of lighting very low. In this case the actual world chance at 12:00 am is 99%, but the lowest, (I do not click my fingers)-worlds chance is close to zero percent. If so, by **Prob-Sufficiency**, expressed in terms of lowest otherworldly chances, my clicking my fingers caused M to light, which is false.

What I have described in these last cases are unwanted probabilistic dependencies emanating from the transition zone: the temporal region t_0-t_c described in §1. A non-actual event that occurs in some of the transitions to the (c does not occur)-worlds is able to influence the chances of the effect in those worlds. I have assumed so far Min-Miracle. Adopting Max-Match does nothing to alleviate the

difficulties since the class of closest antecedent worlds, given Max-Match, still includes, as a sub-class, the Min-Miracle worlds. If so, the problems described for **Prob-Sufficiency** occur whether we accept Max-Match or Min-Miracle. This is basically because the analysis is dealing with might-conditionals rather than would-conditionals.

We might be tempted to view these probabilistic transition zone problems as analogous to instances of probabilistic preemption. In probabilistic preemption, an event c causes e, but c lowers the chance of e by preempting a more reliable process leading to e. The cases above are not instances of this, since in these cases, c does itself preempt a more reliable process leading to e. Rather, some causal antecedent of c does that. Although our cases are not quite preemption, we might still ask if solutions to probabilistic preemption can be applied to them. They cannot.

One typical example of probabilistic preemption is the following from Menzies (1989, p. 647, 653). Take the diagram below representing a set of connected neurons. Forwards arrows represent stimulatory connections, backwards arrows represent inhibatory connections:



The causal process from B's firing to E's firing is faster but less reliable than that from A's firing to E's firing when allowed to run its course. In this case, both A and B fire at 12-00 am, but the B-E process occurs and the A-E process is inhibited. B's firing causes E to fire. A's firing does not cause E's firing. However, A's firing raises

the chance, at 12:00 am, of E's firing, whereas C's firing does not. So by **Prob-Sufficiency**, B's firing causes E's firing.

The solution to this sort of case, proposed by Menzies (1989), is that where there is some temporal gap between two events c and e, c causes e just in case for any finite set of times falling within the gap between c and e, events at those times, $f_1..f_n$, form a chain, c, $f_1..f_n$, e, for which there is chance raising between successive events. So in the Menzies-case above there is always for whatever selection of intermediate times such a chain linking A's firing and E's firing, but this condition does not hold for B's firing and E's firing, therefore B's firing is ruled out as a cause, as desired. Unfortunately, this sort of solution does not work for the cases of unwanted probabilistic effects from the transition zone I have just described. That is because the events being considered for causation have no temporal gap between them and so there is no place for intermediate events. For example, my striking the match M, or clicking my fingers, to take the other case, overlap temporally with M's lighting. So there is no way of introducing a chain of intermediate events.

There are other types of solutions to such cases of preemption, but I cannot see how they deal with the transition zone problems.[13] That is because solutions to probabilistic preemption deal with processes that occur temporally after the cause, which result in chance lowering of the effect. They do not deal with processes that occur before the cause, which due to the transition zone lower the chances of the effect. Given that the probabilistic analysis has its own transition problems, there is not much hope of applying it to the transition zone problems of §1 and §2.

5. Agency

What then of agency? Price (1991, 1992) argues that agency is a concept with which a counterfactual analysis might combine to produce an account peculiarly capable of explaining causal asymmetry and thus dealing with the problem of effects. Price (1992, p.516) presents the agency theory of causation as superior to Lewis' account

only in that it explains micro level causation whereas Lewis' does not. Price thinks that Lewis' account fails on the micro level because the type of de facto physical asymmetry that Lewis (1979) invokes to explain the asymmetry of counterfactual dependence, the asymmetry of overdetermination, does not obtain on the micro level. Nevertheless, Price's account depends upon Lewis' account working on the macro level. This is because Price suggests that agency is explained in terms of macro causation, which is reducible, on Lewis' account, to macro level asymmetry of the kind Lewis invokes. Price then proposes, roughly, that c cause e just in case an agent's bringing about c would be a means of achieving e's occurrence. He then applies the agency idea to both macro and micro levels to provide a more general account than Lewis'. The problem is that Lewis fails to explain macro causation; counterfactual dependence on the macro level fails to track macro causal dependence. But that renders dubious Price's counterfactual treatment of agency qua instance of macro causation.

We can see this most clearly in a simple case, of my striking the match M and lighting it, discussed above in §4. My striking M is my means of lighting it. But on the particular occasion in question the counterfactual, if I had not struck M, it would not have lit is false. Moreover, my striking M does not counterfactually raise its chance of lighting. In short, counterfactual dependencies fails to explain the agency in this case. For that reason, Price's agency theory, in this counterfactual form, is not likely to be right.[14]

6. Conclusions

We have seen that the counterfactual analysis of causation is dogged at every turn by unwanted dependencies from the transition zone. I think we have grounds for the conclusion that counterfactual analyses are unlikely to provide adequate accounts of causation. It is certainly true that there is some close, indeed, interesting connection between counterfactuals and causation. We can use counterfactuals to

communicate, if the context is right, causal information and also to work out causal paths in complicated cases. But no class of counterfactuals track causation, whether we take that as a reductive analytic claim or one of conceptual connection. The reason is due to the structural feature described in the paper, the transition zone and its unwanted counterfactual dependencies.

[1] See the special issue of the Journal of Philosophy April 2000 devoted to these problems.

[2] See Price (1992) for a discussion of problems for counterfactual analyses of causation in explaining causal asymmetry on the micro level. My concern is purely with macroscopic events and the failure of counterfactual analyses to explain macro level causation.

[3] See Lewis (1979). There are also *backtracking* counterfactuals, which do not track causation at all.

[4] Support for the claim that there is backwards causation here might seem to be provided by the observation that somehow a condition of the S1's descending was S2's descending. One might express that idea through a counterfactual, presumably not on a forwardtracking reading: 'If S2 had not descended, S1 would/could not have descended'. That is confusion. We surely cannot be committed to S2's descent causing S1's descent. Thinking we are depends upon conflating two events: the S2's not resisting and S2's descending. A cause of S1's moving down is the failure to resist by the solder bonds and friction at each temporal point; at each point the gravitational force downward is greater than the forces upward of friction. Each failure to resist leads to further descent; it involves the simultaneous descent of both S1 and S2.

[5] There are other cases of bogus backwards causation. Take Hausman's (1996, p. 60) example. Say that George is standing on a bridge over a river. He jumps. Assume that there is no way of stopping his fall once he jumps. A moment later George plunges into the water. Given Min-Miracle, the following looks like a true forwardtracker:

(i) If George had not plunged, he would not have jumped

The Min-Miracle (no plunge)-worlds look like those worlds where George just decides not to jump, a little miracle occurs in his brain leading to his not making the decision to jump. Other kinds of worlds, ones in which he jumps but spinning in mid-air manages to grasp the side of the bridge before he falls too far. This involves more law violation than the first given that George propels himself with some force outward. Or he falls all the way and just before he hits the surface a raft spontaneously appears and halts his fall, or a rope attaching him to the bridge comes into existence, etc. None of these miracles are BIG. But they involve significantly larger miracles but only marginally more perfect match. By Min-Miracle (i) is true and by **Sufficiency** we have backwards causation.

[6] For antecedent propositions P about microscopic events involving few parts, e.g., the movements of a single particle, then, maybe we can get to P without a temporal gap between t_0 and t_P and with only a very small miracle or with no miracle at all, assuming indeterminism.

[7] Lewis expresses the hope (1986b, p. 40) that there will be no 'detailed and definite dependence' of the kind discussed here.

[8] There are cases where the only sufficient condition for causation is whether-on-whether dependency, and that is the end of the matter. Say that Nixon presses the button at 12 am and causes a circuit to be closed in the Doomsday system and the world is destroyed thereafter. Say that a CIA agent, unbeknowst to Nixon, is watching him intending to deactivate the bomb system if Nixon pressed at any other time but 12 am. By shear luck, Nixon got this time precisely. If seems then that Nixon's pressing the button caused the circuit to close. But the only basis for the judgment seems to be a whether-on-whether dependence. If Nixon had pressed the button earlier it is not the case that it would have been closed earlier or later. If he had pressed the button in a slightly different manner, it is not the case that the closing of the circuit would have occurred in a different way.

[9] Max-Match is not obviously inconsistent with the thrust of Lewis position. It is not denying that there is a transition zone in some sense, that is, it is not holding that the closest not-c-worlds will match perfectly right up to the antecedent time, which Lewis (1986b p. 39-40) explicitly rules out. The principle holds something weaker, namely, that some of the closest not-c-worlds might have this character.

[10] Could we weaken Paul's conditions for causation to read: c causes e if and only if were c not to have occurred, e would not have or might have occurred later? We can't. Say S1's descending delayed some event f. Then if S2 had not descended, that event might of occurred later than it actually did. But S2's descent was not a cause of that f.

[11] For example, Ramachandran's (1997) M-set analysis will not help us find an alternative explanation of the truth of (5). This account is based on the idea of an M-set defined in the following way: a set of events $\{c_1, c_2, ...\}$ is an M-set for an event e just in case, if none of the members of $\{c_1, c_2, ...\}$ had occurred, *e* would not have occurred and there is no subset of $\{c_1, c_2, ...\}$ for which the same counterfactual dependency condition holds with respect to e. In outline the M-set analysis proposes, (Ramachandran (1997, p. 273-4):

For any actual events c and e, c causes e if and only if (a) and (b) hold:

(a) c belongs to an M-set for e;

(b) There are no M-sets for e, M and N, such that M contains c and N differs only in that it has one or more non-actual events in place of c.

This is modified in various ways in later developments, given in Ramachandran (1998), but these need not concern us here. If we assume Min-Miracle, then (4) comes out true on the M-set analysis since (5) is true. On the other hand, if we assume Max-Match, we are still unable to explain (3). There is no M-set for the event of the cylinder's being scratched. Thus for example, the most obvious candidate, {S2 descended, S1 descended} is not an M-set for the scratching of the cylinder because

although the following counterfactual is true – if S1 had not descended and S2 had not descended, the cylinder would not have been scratched – the counterfactual – if S1 had not descended, the cylinder would not have been scratched – is true as well.

[12] A last ditch idea is that forwardtrackers are not fixed by one particular similarity metric, but by two: Min-Miracle and Max-Match. Which similarity metric is in play is determined somehow by the consequent of the counterfactual. So although (P > Q) and (P > R) evaluated at @ have the same antecedents the nearest P-worlds in each case differ because we are evaluating with respect to different consequents. Thus (4) is evaluated with respect to Min-Miracle, and (5) with respect to Max-Match. The problem with this idea is that it is entirely unclear what factors about consequents would determine the change in metric.

[13] For example, Noordhof's (1999) more complicated probabilistic account entails the correctness of something very like **Prob-Sufficiency** on the least value reading; the difference is simply that Noordhof requires that the chances, both in the actual world and counterfactual worlds, are relativized to a time just before the putative effect. This slight variation is immaterial in the cases of unwanted transition zone dependencies, since the putative causes and effects have no temporal gap between them.

[14] Price does not in fact commit himself to understanding agency, or the means-end relation, in counterfactual terms. He also thinks that subjective probabilities might do the trick. See Price (1991). But that is no longer a counterfactual analysis.

References

Beebee, H., 1997. 'Counterfactual Dependence and Broken Barometers: A Response to
Flichman's Argument'. Crítica 24, pp. 107-117.

Bennett, J. 1984. 'Counterfactuals and temporal direction'. Philosophical Review 93, pp. 57-91.

Flichman, E., 1989, 'The Causalist Program, Rational or Irrational Persistence?', Crítica 21 pp. 29-53.

Hausman, D., 1996, 'Causation and Counterfactual Dependence Reconsidered', Nous 30 pp. 55-74.

Lewis, D. K., 1973a, Counterfactuals. Cambridge Massachusetts: Harvard University Press.

Lewis, D.K., 1973b, 'Causation' (plus postscripts), in Lewis 1986b, pp. 159-240. Originally published in 1973 in Journal of Philosophy 70.

Lewis, D. K., 1979, 'Counterfactual Dependence and Time's Arrow' (plus postscripts), in Lewis 1986b, pp. 32-66. Originally published in 1979 in Nous 13.

Lewis, D. K., 1980, 'Events' (plus postscripts), in Lewis 1986b, pp. 32-66. Originally published in 1979 in Nous 13.

Lewis, D.K., 1986a, 'Events' (plus postscripts), in Lewis 1986b, pp. 159-240.

Lewis, D. 1986b. Philosophical Papers, Volume 2 (Oxford: Oxford University Press).

Lewis, D. K., 2000, 'Causation as Influence', Journal of Philosophy 97 pp. 182-197.

Menzies, P., 1989, 'Probabilistic Causation and Causal Processes: A Critique of Lewis'. Philosophy of Science 56 pp. 85-116.

Noordhof, P., 1999, 'Probabilistic Causation, Preemption and Counterfactuals'. Mind 108, pp. 95-125

Paul, L, 1998, 'Keeping Track of the Time: Emending the Counterfactual Analysis of Causation', Analysis 58.3, pp. 191-198.

Price, H., 1991, 'Agency and Probabilistic Causality', British Journal for the Philosophy of Science, 42, pp. 157-76.

Price, H., 1992, 'Agency and Causal Symmetry', Mind 101, pp. 501-520.

Ramachandran, M., 1997, 'A Counterfactual Analysis of Causation'. Mind 106, pp. 263-277.