Causal Counterfactuals and Impossible Worlds

Daniel Nolan

There seem to be tight connections between claims about what caused what, and many claims about what would have happened if things had been otherwise. A special, but important, case of these are the connections between causal structure and what would have been different had things been different. A lot of careful and ingenious work has gone into trying to articulate the connections between the two, though it is probably safe to say that no completely satisfactory account has yet emerged: or at the very least, those who are completely satisfied with an account of the connection between causation and counterfactuals are few, and disagree with each other about *which* is the completely satisfactory account.

This paper is not directly in the service of either of the ambitious analytic projects of analysing causation in terms of the holding of certain counterfactuals, nor of analysing counterfactuals in terms of causal matters. It focuses instead on one of the traditional puzzles that connect the two that arise almost whatever one takes the connection between counterfactuals and causation to be. The puzzle has no neat label that I am aware of, but it arises in its clearest form when we consider counterfactuals involving antecedent states that involve a difference from the actual course of events at a particular time, and a consequent, at least in part involving a state somewhat later than the time of the antecedent difference. In the possible worlds framework, the puzzle is often put in terms of what other differences there in the relevant possible worlds where the antecedent is true. Do those worlds match ours with respect to nearly all their pasts until the time relevant to the antecedent? Do they require 'small miracles' relative to the laws of the actual world? Is their causal structure the same except for an 'intervention' on a state associated with the antecedent? . . . and so on. Let me label this problem the 'deviation problem' to suggest that it concerns what deviations from actuality would be required for the antecedent to be true, in the class of counterfactuals of interest.

In this paper I will propose a novel solution to the deviation problem. This solution will have several signal advantages over a number of the better-known proposed solutions to this problem, though it also incurs some distinctive costs. I am not sure, then, whether something like it will turn out to be the best solution, though I am sure it deserves a run for its money alongside its better-known cousins. I will do so by presenting the puzzle and solution in a closest-worlds framework of the Stalnaker-Lewis variety: those familiar with alternative systems will likely be able to see easily enough how to fit the kind of positive proposal I offer into those approaches.

After some remarks about the problem my solution is intended to solve, and the kinds of resources I will deploy to construct my version of the solution, I will argue for a number of desiderata for a solution that traditional approaches compromise, before presenting my solution and displaying that it can satisfy those desiderata. Finally, I will discuss some of the vices of the particular solution I offer, and while I do not intend to suggest this solution is without costs, I will have some things to say about why those vices may not be as great drawbacks as they might initially appear.

1. The Target

The problem I wish to address in this paper is more specific than the general problem of the truth-conditions of counterfactuals. It the narrower problem of offering a story about the truth-conditions of what I am calling *causal counterfactuals*. So it would be good to begin by being a bit more specific about which counterfactuals I have in mind.

The counterfactuals I will pay attention to are those whose antecedents concern a specific one-off event or state, and their consequents deal with what happens after that event or state (or as a consequence of a failure of the antecedent event to influence the consequent). Furthermore, they are the non-*backtracking* counterfactuals of this sort. A *backtracker*, intuitively, is a counterfactual that invites us to consider what would have had to be different in the causal ancestry of an event if it were to have come about. How exactly to demarcate which counterfactuals are backtrackers is controversial, but see Lewis 1979 pp 33-34 for the *locus classicus* of a characterisation of back-tracking conditionals.

The causal counterfactuals that are my focus here all come with an *antecedent time*.[1] Some counterfactual conditionals have antecedents that are about a relatively particular event: ‘if I had set the fire alarm off at 11.00 this morning, the fire service would have been here by 11.30’. In such cases, we can talk about a particular time associated with the antecedent: roughly, the time at which the situation described by the antecedent would have obtained. In the example, that time would be

[1] Lewis 1979 introduces the idea of an antecedent time T_A, though he characterizes it as ‘the time the antecedent is about’, which is not quite how I will characterize the notion I wish to use.

some period around 11am. Many other antecedents suggest a time in a much less explicit way. 'If I had skipped breakfast, I would have had more to eat at lunch' suggests a time around my normal breakfast time, or perhaps my actual breakfast time. Not every counterfactual is associated in this way with a time, and it would be natural to extend the account given to cases where there are a number of salient times associated with an antecedent of a causal counterfactual, but I will restrict my discussion for tractability.

For the purposes I want to use the notion, the 'antecedent time' associated with a counterfactual will not always be the time the event associated with the antecedent would take place. It will often need to extend some relatively short time before that event. The motivating idea is that the antecedent time is the time at which a world where the antecedent occurs would have to rapidly become quite different from the actual world, so will typically involve difference from the actual world some time immediately before as the 'run up'. (Were I to have turned off on the previous exit ramp, it would not be by a last minute swerve or by teleportation, but rather may well have been by getting into the correct lane, signalling a turn, etc.) So it seems better to say the antecedent time can include some period before the time of the event explicitly invoked by the antecedent. A tricky and unsolved problem is exactly how much of the run-up to the relevant event is best to include in the 'antecedent time'. The puzzle about setting an antecedent time is unfinished business, but the details should not matter for current purposes.[2]

The target problem, then, is the deviation problem for causal counterfactuals. What would be different were the antecedent of a causal counterfactual true? Or, to put it in the world terminology introduced below, what are the most relevantly similar worlds where the antecedent of such a counterfactual is true?

The proposal to be offered will be straightforwardly generalisable in a number of ways, including to cases where antecedents have a number of natural antecedent times associated with them, to some counterfactuals with general antecedents, and even to counterfactuals where the states associated with the antecedent and consequent are non-causally related. But the interested reader can chart for herself how the proposal in this paper can be generalized, as even the relatively narrow range of counterfactuals I have picked out will give us quite enough fish to fry.

[2] A general account will also need to say when an antecedent time ends, and may want to derive the fact that we treat the past differently from the future in causal counterfactual contexts from some more basic principle.

2. Resources

The most salient commitment of the approach to be used is to an apparatus of worlds, possible and impossible. Relying on accounts of counterfactual conditionals in terms of *possible* worlds has, by now, a long history, with the important papers of Stalnaker 1968, Stalnaker and Thomason 1970, and Lewis's influential 1973 book helping to make this approach to counterfactuals close to orthodox. I will adopt some of the details of Lewis's specific proposal. For the counterfactual A[]->B to be true at a possible world *w*, B must be true at all the 'nearest' worlds to *w* where A is true (and A[]->B is false otherwise). Lewis explains nearness in terms of *similarity*, and Lewis holds that which aspects of similarity are relevant is set by the context in which the sentence expressing A[]->B is produced. (I will have more to say about context, below.)

In Lewis's framework a lot of the work in determining the truth-value of a counterfactual is done by the similarity measure on worlds. The dimensions of relevant similarity are not a matter of all-things-considered similarity (whatever that would be), but are rather a matter of similarity in relevant respects. *Which* are the relevant respects, and how they are weighted against each other, is then a central question for this account (and, given that the account says that relevant similarity is determined by context, a question that must be re-asked for each context of utterance). Lewis 1979 is his account of what he takes the relevant dimensions of similarity to be for causal counterfactuals, and while I will not be endorsing that account, it is an example of the *kind* of account that is needed. This paper will follow Lewis in appealing to comparisons of relevant similarity as part of the machinery for delivering truth-conditions of counterfactuals.

Both Lewis's and Stalnaker's systems were constructed so that when the antecedent of a counterfactual was impossible, the counterfactual was automatically true. However, it does seem very natural to not treat counterfactuals with impossible antecedents all in the same way: these so-called *counterpossible* counterfactual conditionals seem to be usefully employed in logic, mathematics, metaphysics, and in many other areas (Nolan 1997). For convenience, I will refer to these conditionals simply as 'counterpossibles', though that label often has a wider application to all sorts of conditionals with impossible antecedents (including indicative conditionals, for example).

The easiest way to incorporate counterpossibles into a worlds framework is to include impossible worlds as well as possible worlds in the account of truth-conditions of counterfactuals. Consider the conditional 'If 27x4 were 131, then 131 would be composite'. Plausibly, of the impossibilities where 27x4=131, the most relevantly similar to actuality are those where being the product of two whole numbers other than 1 or 0 is sufficient for being composite, and will treat that consequent as true. On the other hand, when evaluating 'If 27x4 were equal to 131, then 27x4 would be equal to 1310', then the same, or a very similar, 27x4=131 worlds seem most relevantly similar: and in none of those will 27x4=1310. That would be a gratuitous departure from actual mathematical truth, so the second counterpossible mentioned is false. A full account requires a story about what makes for relevant similarity between a possible and impossible world, of course.

The particular example above may not strike everyone as convincing. But what is important is to notice that treating some impossible worlds as more relevantly similar to actuality than others is a natural way to extend the basic Lewis/Stalnaker semantics for counterfactuals.

If we do so, we face a number of choices about how to understand these impossible worlds. One straightforward way to do so is to model them as sets of propositions: but to not insist that these sets are closed under logical consequence. Modelling worlds as arbitrary sets of propositions will give us possible worlds as well as impossible ones, but it should not matter exactly where we draw the possible/impossible line, except that I will assume that sets of propositions which are jointly inconsistent are associated with impossibilities.

One distinction is particularly important to keep in mind when employing impossible worlds: the distinction between what is true *according to* an impossible world and what is true *about* an impossible world. (Something like this distinction is sometimes characterized as the distinction between what is true *in* a world versus what is true *of* a world, but that terminology can be more confusing.) An impossible world may not have true *according to it* that what happens is impossible, for example: it might represent that everything that happens in it is possible. It may not have true *according to it* that a contradiction is true, even if the proposition that roses are red, and also the proposition that roses are not red, are both true according to it. Suppose we model impossible worlds with sets of propositions, as above. One set can contain the proposition that there is a round square cupola, while also containing the proposition that everything which exists is possible. One set can contain the following three propositions: that roses are red, that roses are not red, and that no contradictions are true. Even though the first set contains a proposition that there is a round square cupola, and so it is true *about* that set that it represents that there is an impossible object, it is not true *according to* the set. The second set contains a contradiction, so it is true *about* the set that it is contradictory, but the proposition that a contradiction obtains is not true *according* to the set, understood as an impossible world (or a fragment of one). Alternative theories of impossible worlds will offer different ways of accounting for the according to/about distinction, but the illustration should suffice to indicate the distinction I have in mind.

As well as worlds, possible and impossible, ordered by similarity in relevant respects, the account of causal counterfactuals I will develop shares with many such accounts a commitment to *laws of nature* rich enough to impose constraints on the evolution of the world in the respects we care about. I will try to stay relatively neutral on how to understand these laws of nature: in particular, as mentioned above, I will not take a stand here on whether some kind of Humean regularity account of laws of nature is sufficient, or whether something more metaphysically heavy-duty is required. I will presuppose for the rest of the paper that laws of nature are contingent, in the sense that they can vary from possible world to possible world. I do this in a concessive spirit, however: appeals to impossible worlds to evaluate causal counterfactuals are much more appealing if any rival laws of nature are *impossible*, and hold in no possible worlds at all.

A resource that will be lurking in the background is the view that counterfactuals exhibit a certain amount of *context sensitivity*, here implemented by allowing that which standard of relevant similarity can vary from one context of utterance to another. This context variability of counterfactuals also makes a difference to how I am conceiving of this project. It is not the task of providing the truth-conditions of all counterfactuals whatsoever, nor of isolating some semantic ambiguity in conditional locutions so that one disambiguation is the 'causal' one. It is to explain the truth-conditions of *some* uses of counterfactuals: perhaps many typical counterfactual utterances. I take it this context-sensitive approach to counterfactuals is in the spirit of Lewis's (see Lewis 1979 pp 33-35), but not in the spirit of everyone who offers closest-world analyses of counterfactual conditionals: Bennett 2003 does not allow for contextual variability, for example.

3. Desiderata for a Solution

Given a closest-world approach to counterfactual conditionals, a solution to the deviation problem will be the specification of some conditions on relevant similarity. This specification will ensure the most relevantly similar worlds when evaluating a causal counterfactual are ones that are not gratuitously different from ours, and line up with correct counterfactual judgements. That is, when the causal counterfactual ‘A []-> B’ is true, the most relevantly similar worlds according to which A is true are ones that have B true according to them as well; and when ‘A[]->B’ is false, this condition will not obtain.

There are three plausible desiderata for this solution that are each supported by plausible argument. They appear to be jointly inconsistent if we want to get the intuitive truth-values for ordinary causal counterfactuals, and so solutions to the deviation problem in the literature give up on one or more of them. In this section I will outline and defend each of these desiderata, before explaining in the next section how we can maintain all three in a solution to the deviation problem.

3.1 The Same Laws of Nature, and No Counterfactual Miracles

When considering a counterfactual situation, we initially assume the same fundamental principles of nature are at work. There are at least two reasons to think that we do this: one is that we freely employ stable generalisations about what leads to what when reasoning about whether things being a certain way at the antecedent time leads to the consequent obtaining. (Hmm, the weight would have been here, so the scales would have moved to here, so that would have tripped the lever….) The second reason is that, for the kind of antecedents found in causal counterfactuals, ‘if it had been that A then the laws of nature would have been different to the actual laws of nature’ rarely sounds like an appealing counterfactual. That suggests that for most of these antecedents, the relevantly similar worlds where they are true are not one where the laws of nature differ from the actual ones. For that matter, ‘Even if it had been that A, the laws of nature would have been the same’ also normally sounds fine, if a little odd to utter – we tend to take for granted that various differences would not have resulted in different laws of nature.

Perhaps despite these observations, the laws of nature at such counterfactual worlds need not be exactly the same. Even if they are not exactly the same, presumably they should not be too arbitrarily different – that would violate the motivating idea of similarity in relevant respects, at least insofar as we are concerned with laws that affect the influence (or not) of the antecedent event on the state associated with the consequent. Despite this initial appeal, a number of authors have endorsed the option that the nearby worlds where the antecedent obtains vary with respect to the laws. Terminology for these deviations from actual laws of nature introduced by Lewis labels them as 'miracles'.[3] Let me now say in some more detail why we should not be happy with a theory that invokes miracles for standard causal counterfactuals.

Had things been different in various ordinary ways, that would not have taken a miracle, or so we think. My having lunch at a different cafe, or Everest being found twenty metres closer to K2 than it in fact is, or there being half a tank of petrol in a car instead of a full tank, would not require violations of the actual laws of nature. Or so we ordinarily think. Of course, we could be radically wrong about the actual laws of nature: perhaps everything that happens, happens as a matter of nomic necessity, for example. And I do not mean to say that *no* counterfactual whose antecedent is about specific causal processes could have a consequent saying there were miracles and still be true: if I were to cause a miracle worker to perform miracles, there would be a miracle.

In 'Are We Free to Break the Laws?' (Lewis 1981) Lewis defends the view that some possible agents in deterministic worlds are free to perform acts such that, if they perform them, a miracle would occur. And in Lewis 1979, he defends the view that in deterministic worlds, often the nearest world in which standard antecedents for causal counterfactuals are true are ones with miracles: small miracles', that do not make for too much distance between worlds.

Lunching at a different cafe, Everest being a little closer to K2, cars having different levels of petrol in their tank, and so on are all compatible with the actual laws of nature. We are inclined to think that were any of these things the case, we would not need to have different laws of nature, but just differences in ordinary matters of fact. Nobody trying to change the level of petrol in a car ever tries to do this by changing the laws themselves. Very few people who regret they did something that was under their control blame the laws of nature rather than their particular actions. I do not think those led to suppose that leaving a chip uneaten, or putting on a different shirt in the morning, would have taken a miracle, do so because that strikes them as the pre-theoretically compelling view: rather, they seem pushed there by the apparent lack of a feasible theoretical alternative that has other features they want.

If quotidian things could be otherwise without requiring miracles, that would seem preferable. For

[3] Well, strictly speaking, the definition should make no mention of the actual world, an event in w_1 counts as a miracle according to a possible world w_2 provided the event is against the laws of w_2, and I intend this general definition to be the one in force: but the more general definition is unnecessary to get the idea.

one thing, miracles cannot happen. (That is, events that in fact violate our laws are nomically impossible – though they are possible in a more generous sense, and are nomically possible in the worlds where they occur.) But wearing a different shirt or not eating a chip are things that can happen, we ordinarily think. On the face of it, were they to happen, nothing impossible would need happen.

The case for requiring miracles in the nearby worlds relevant for ordinary causal counterfactuals often requires some other assumptions: that those worlds share a lot of other truths with ours, and that the laws tightly constrain how things can be at the antecedent time given those other truths, or at least that we should have an account flexible enough that it delivers similar results even when the laws restrict the possible outcomes in this way. Let us, then, turn to the other two desirable features the nearby (i.e. relevantly similar) worlds should have when evaluating causal counterfactuals.

3.2 Common Past

The first two constraints suggest a picture: that we consider the closest worlds where what happens at some time associated with the antecedent varies from what in fact happens, but is otherwise as similar as feasible, in respect of what happens at that time, to the actual world; and which obeys the actual laws of nature at all the times in that world (and perhaps shares the actual laws of nature to boot). One feature that this approach would have is that it may well result in some of the closest worlds having quite different pasts from the actual world. It is a familiar point that in a world that is deterministic and chaotic (in the technical sense of chaotic), very small changes at one time can result in very great differences in the future: one butterfly flaps its wings in one place, and a tornado that would not have otherwise happened happens on the other side of the world a year later.[4] In such deterministic and chaotic worlds, fixing the physical state at a time and determining what follows, given that state and the laws, for *earlier* times can produce equally extravagant differences. An extra butterfly flapping its wings now, plus the laws, might entail tornadoes in the years *before* that are absent from the actual world.

The oddness of these counterfactual worlds diverging in their pasts (and in some cases, more and more radically as times considered are earlier and earlier than the antecedent) may only seem a

[4] I am not aware of any meteorological model that is this sensitive to slight air movements, and the case that the weather is *this* sensitive has not been made, to my knowledge: but the illustration of the principle is useful enough, perhaps as a convenient fiction, even if actual weather systems do not behave like this.

curiosity so long as we consider causal counterfactuals with consequents only about the future. But they give incorrect results about a range of counterfactuals which have consequents that are partly about the future and partly about the past. Cases like this have been presented by Lewis [1979 p 33] and by Bennett 2003 (pp 202, 214). Let me present two, very ordinary, cases of this sort. Suppose I have just come home and looked through my bag, worried about my umbrella, and located it. I say 'if I had left my umbrella in the bar, that would have been the third umbrella I'd have lost this month': and let us suppose that counterfactual is true in the envisaged case. For that to be true, the nearest umbrella-in-the-bar worlds have to be worlds that do not just have to resemble ours with respect to laws or how the world unrolls from the incident in the bar, but must also agree with ours in how many umbrellas I had lost earlier in the month. Examples like this can be multiplied indefinitely: it is very natural to utter counterfactuals with consequents that require all sorts of matches with the actual past to be true. The natural thing to think, here, is that our practice of uttering causal counterfactuals takes for granted that the past before the antecedent time would be as it in fact is. Counterfactuals like these also count against theories that consider situations with *no* past before the antecedent time: see Paul and Hall 2013 pp 47-48 for a proposal of this variety.

3.3 Compatibility with Determinism, and Near Determinism

If determinism is true, then necessarily if the laws are as they actually are and some complete time-slice of the past is as it actually is, then every other event will be as it actually is. So given determinism, keeping both the laws and the past the same is a challenge if we are looking for a relevantly similar world W where some antecedent which is actually false is true according to W. A theory that is compatible with determinism is one that allows that even when an antecedent of a causal counterfactual is false at a deterministic world, still there will typically be worlds relevantly similar to that world in which that antecedent is true.

Indeed, the past and the laws will be able to rule out some antecedent times varying from what actually happened even if determinism is not true. Having indeterministic laws is not the same as saying that anything goes: it just means that there can be *some* variation in worlds that share *some* common past and the same laws. If our laws were indeterministic about some phenomena but not others, or some stages of the evolution of the universe but not others, they might still have the result that the past before an antecedent time, plus those laws, guarantee that the antecedent will not occur at the antecedent time. Call such cases cases of 'near-determinism'.

Some philosophers' reactions will be that this is so much the worse for determinism, and near-determinism. Some are inclined not to worry about how an account of counterfactuals deals with determinism and near-determinism, since the kind of indeterministic world we inhabit seems to be one where almost any event permitted by the laws at all could happen after almost any history, albeit often with a vanishingly small chance of in fact doing so. However, I think there are several reasons to prefer an account that can vindicate many of our ordinary causal counterfactuals even if determinism were true, and one reason to like the *style* of account that can accommodate determinism even if our focus is only to produce an account fit for the sort of indeterministic world many take us to in fact be in.

One reason to look for a theory compatible with determinism is our tendency to, at least sometimes, treat determinism as irrelevant for causal counterfactuals. 'Okay, I wouldn't have left my umbrella behind, given the past and the laws. But if I had, would I have been able to get it back?'. Since we are (presumptively) competent users of counterfactual expressions, that suggests that determinism would not by itself render all causal counterfactuals vacuous or otherwise defective. Another, related, reason is that apparently competent users did not eschew the construction when determinism was widely believed (at least among the Newtonian intelligentsia).

This paper will remain of some interest even for those unconvinced of these motives to have an account compatible with determinism. After all, the issue of whether determinism *would* have serious consequences for the truth-values of causal counterfactuals can be of interest (perhaps relatively academic interest) even to those convinced of indeterminism: and so the prospects of theories that yield the normal truth-values for causal counterfactuals even in deterministic worlds otherwise rather like our own (superficially, at least), should be of interest in weighing up different verdicts about what impact determinism *would* have on causal counterfactual propositions.

One reason why it is instructive to consider options for counterfactuals under determinism is that some of the pressures on theories of counterfactuals under *indeterminism* are similar. Some indeterministic frameworks, including the ones that most plausibly describe the actual world, can produce variation from actuality at the antecedent time, while keeping the laws and past fixed, only at the cost of strange and mind-bogglingly unlikely chance events: quantum tunnelling, massive spontaneous decay, and the like. Call an incredibly unlikely quantum event that would mimic a Lewisian miracle a 'semi-miracle'. (Note I do not mean a 'quasi-miracle' in Lewis's sense (Lewis

1986 p 60), which is its own can of worms.) Solving the deviation problem by keeping the past and laws fixed and postulating a semi-miracle to bring about an antecedent looks only a little less bad to many than postulating Lewisian miracles.

If you think that semi-miracles should not be needed in an account of counterfactuals under the sort of indeterminism we probably have in this world, then you have a similar dilemma to the determinist. Just as same laws (and no miracles), same past, and a contrary-to-fact event at the antecedent time form an inconsistent triad, same laws, same past, a typical contrary-to-fact event at the antecedent time and *no semi-miracles* at the antecedent time will often be an inconsistent tetrad, when ‘semi-miracle’ is understood in the appropriate way. The solution I will offer those who wish to allow for determinism will be straightforwardly adaptable for those who wish to endorse typical causal counterfactuals under indeterminism without its being true according to relevantly similar worlds that semi-miracles occur.

3.4 Theorists Often Compromise On These Desiderata

Given the above desiderata, it is easy to see why many theorists have thought one or more of them has to go. There is apparently a tension. If world W is deterministic, then the past of W plus the laws of W entail all the other propositions true according to W. (All the other propositions about events subject to the laws of nature, in any case.) So if the antecedent of a counterfactual, A, is true according to a world with the same laws and same past as W, it looks like A would have to be entailed by *W’s* laws and the past. So when we want to consider some *counterfactual* world which disagrees with W - when we want to consider a counterfactual with an antecedent A that is *false* according to W – it seems that something has to give, out of the laws and the past.

Different attempts to solve the deviation problem give up one or more of the desiderata listed above. For example, Lewis 1979 offers two accounts that address the deviation problem, and settles on the second. (Both presuppose determinism for purposes of tractability, though he offers some remarks about how to extend them to indeterminism in Lewis 1986.) Lewis’s first account is not presented as a closest-worlds account, though it is easily paraphrasable into one. Both options Lewis considers violate one or more of the desiderata, above. The first permits violations of the actual laws during ‘a transition period beginning shortly before t_A’ (Lewis 1979 p 39) (in my jargon, violations of actual law during the ‘antecedent time’). Lewis’s second proposal in principle allows for violation of both the desideratum about laws and the one about a common past (and indeed for cases where both are violated): but in practice it lends itself to violations of the principle about laws in deterministic worlds. Typically, the relevantly similar worlds are ones that contain miracles relative to the deterministic worlds.

Jonathan Bennett (2003) talks of 'forks', which occur in relevantly similar worlds when the world first diverges from the actual world in a significant way. He talks of forks happening not long before what *he* calls the antecedent time T_A. In my usage, Bennett's whole fork happens within what I call the antecedent time. Bennett allows that forks can happen in one of three ways in the relevantly most similar worlds. One is through indeterministic variation. The second is through a miracle in Lewis's sense. The third way does not require any changes in laws, even in deterministic worlds, but only 'tiny imperceptible' differences in great stretches of the past (Bennett 2003, 217-8). This violates the 'common past' desideratum in a comparatively low-cost way, at least in worlds that have laws kind enough to permit it, since while worlds differing in this way do not have common pasts, they have pasts that are the same in a number of respects we care about.

Bennett wishes to allow that all three of these methods of forking can occur in the relevantly similar worlds (Bennett 2003, 218). Only the second two will be available in deterministic worlds (and in some indeterministic ones): so in effect, Bennett compromises *both* the desideratum about laws and the desideratum about the common past.

One increasingly popular style of analysis of causal counterfactuals is to invoke the sort of causal models popularized by Pearl 2000. Different theorists have used these models in different ways to offer semantics for causal counterfactuals, but these approaches either compromise on the desiderata listed above or fall silent about the semantics of relevant counterfactuals, or often both. Woodward 2003, for example, both adopts a story of 'interventions' for evaluating interventionist counterfactuals that, in effect, asks us to consider setups where there are breaches of actual laws (see Woodward p 136 where he compares his interventions to Lewis's 'small miracles'), and in effect only offers a way of understanding only a narrow range of counterfactuals: those whose antecedents and consequents only concern values of variables in causal models, or counterfactuals related to these. (Counterfactuals with consequents about law violations or miracles are not treated.) Menzies 2002 p 828 gives a closest-world account of the truth-conditions of a wide variety of 'causally relevant counterfactuals', albeit truth-conditions relative to causal models rather than guidance about which counterfactual claims are true or false *simpliciter*. I take it the best way to understand Menzies here is in light of Menzies 2004 where he describes his view as a contextualist one: context fixes one or more causal models, and a counterfactual in a context is true provided it is true relative to the contextually relevant model(s).

While Menzies's account is broad enough to evaluate counterfactuals with any propositions in their antecedents and consequents, (as a result of using complete worlds in the semantics), on the face of it this view also compromises the desideratum that there are no law violations. In systems that are comprehensive enough so that the initial conditions and the laws guarantee the falsehood of the antecedent, 'miracles', in the Lewisian sense, occur in the closest worlds where the antecedent obtains (Menzies 2003 p 164). Menzies also allows other law violations and differences in initial conditions for some counterfactuals, but says these are not the typical case for the sort of conditionals we are presently considering, as opposed e.g. to explicit counterlegals or backtrackers. (p 163-5).

Menzies's version of a causal models theory of counterfactuals does have some signal advantages over others in dealing with the present puzzle. The 'laws' of one of Menzies's causal models do not need to be the really-and-truly laws of nature, even when a causal model is one of the appropriate ones given a context. When an 'interfering factor' is not ruled out by the laws of a causal model, the nearest worlds where the laws of the model and the initial conditions obtain may have interferers, even if the model is a deterministic one. Menzies also does not insist that a contextually salient causal model need be actually instantiated: the actual world may contain interferers the model is silent about. So the laws of a contextually appropriate causal model need not even be true universal generalisations. While Menzies himself is willing to employ miracles, a Menzies-style view that used enough of these other resources could be developed to preserve all of the desiderata I mentioned in most cases. In some special contexts and worlds, however, when the contextually salient causal models are (i) instantiated at the world of utterance, (ii) comprehensive enough to rule out interferers and (iii) deterministic, even a Menzies-style view will have few options but to violate one of our desiderata.

There are, of course, many revisionist options for assigning truth-conditionals to counterfactuals, including taking their truth conditions to be those of the material conditional, offering a non-truth-apt account, or endorsing the claim that they are nearly all false (for the last see Hájek unpublished).

I will assume, for the purposes of this paper at least, that we should prefer a theory of the truth-conditions of causal counterfactuals that more closely tracks our reflective assent and dissent: classifying counterfactuals as true when we take them to be true in good conditions, and false when we take them to be false in good conditions. (And where our guide to which conditions are 'good', in this sense, is our ordinary standards for counterfactual evaluation, not our standards 'enlightened' by much philosophical theorising.)

4. The New Option

The option I wish to propose satisfies all of the desiderata discussed above. For a causal counterfactual to be true at a world W, it must be that all the nearest worlds according to which the antecedent is true, and which meet certain other conditions, are worlds according to which the consequent is true as well. The conditions on the worlds W* nearest to W are: first, that the laws of nature true at W are true according to W*. Second, that for the antecedent time t, claims entirely about the goings on in W* before t are true according to W* if, and only if, they are true according to W. Third, no proposition is true according to W* which says that a violation of the laws of nature occurs.[5] Let us in addition require that there will be some worlds among the W*s where the antecedent is true, and the nearest of the worlds in W* where the antecedent is true are worlds where the time after *t* is constrained to evolve by the goings-on at *t* and the laws of nature in the natural way (however it is best to spell that out).

What this proposal does not insist on, however, is that the closest worlds to actuality where the two conditions are satisfied and the antecedent is true are *possible* worlds. It might well be that a certain past, together with things being a certain way at an 'antecedent time', is strictly inconsistent with a certain body of laws of nature.

There is an important respect in which these worlds do not give rise to miracles: even if there is a proposition true according to them about what happens that *in fact* is inconsistent with the propositions true according to them about the laws, or indeed about what mere universal generalisations hold in these worlds, still the world may not have true *according to it* that there are any law violations, and may have no contradictions true *according to* these worlds. What is true according to possible worlds is closed under logical consequence, but this is not in general the case

[5] When considering the indeterministic case, we may also wish to insist that no proposition to the effect that a semi-miracle occurs at W* occurs, or something more sophisticated along those lines.

for impossible worlds. This respect in which they do not give rise to miracles is important, because the truth-value of counterfactuals such as ‘if I had skipped lunch, something ruled out by the actual laws of nature would have occurred’ depends on whether ‘something ruled out by the actual laws of nature occurred’ is true *according to* the relevantly most similar lunch-skipping world. And even if the conjunction of the actual past, the actual laws, and my skipping lunch is inconsistent, *impossible* worlds representing all three need not have true according to them that there is a law violation, either of their own laws or the actual laws.

There would be more to say about relevant similarity in a complete account. For example, I have not tried to spell out explicitly what conjunctions of information about the past before the antecedent time plus the future of the antecedent time are true according to such closest worlds. There must be plenty of the usual ones, on pain of not vindicating the right counterfactuals that have consequents partly about the past and partly about the future, like the examples in section 3.2. It is not necessary to spell out all of the details exactly in order to evaluate this style of proposal, however: nor, indeed, would an *exact* specification of any sort be very plausible, since there is likely to be some semantic indeterminacy as to exactly which counterfactual conditionals are true in which contexts.

Still, enough has been said to make clear why, once we appeal to impossible worlds, we can retain the laws and past of W together with the antecedent, while holding on to many of the ordinary verdicts about the truth-values of different causal counterfactuals: and can do this even if determinism, or something like it, is true in our world, or a world like ours. The compromises explored by Lewis, Bennett, and others are not needed after all. Furthermore, we can see why the quick argument in Lewis 1981 fails: he suggests that we can discount worlds where contradictory propositions are each are true as relevantly nearest, ‘for if I had raised my hand, there would still have been no true contradictions’ (Lewis 1981 p 292). But ‘there still are no true contradictions’ (or however else we should extract the consequent) need not be true according to any of the impossible worlds that are relevantly nearest in such a case: Lewis’s objection fails to touch the proposal in this paper.

There is at least one intuitive motivation for appealing to impossible worlds as a *principled* way out of our puzzle, and not just a technical fix. The thought is that when we evaluate a causal counterfactual, we neglect some features of counterfactual scenarios, including, to some extent, the gritty detail of how the goings-on specified in the antecedent emerged from the past of the world under consideration. (Considerations of the gritty details tends to put us more in a frame of mind to evaluate backtracking counterfactuals: had A happened, how would it have come about?) Furthermore, this neglect is not just due to our ignorance about the exact processes that our past and laws permit. It may be that the rules of the practice itself are insensitive to some of the details here. If the principles of the practice are somewhat insensitive to these details, then even if the practice is primarily concerned with imposing constraints that it is individually possible to satisfy (sameness of laws, sameness of past, sameness of the kind of dependence the future displays on the laws plus the arrangement of the antecedent time), philosophers may well be over-idealising if they look for scenarios that are possible all-things-considered, rather than just matching possibilities in a number of respects, respects which conflict in an area with which the practice is not particularly concerned. If this picture is the right way of thinking about the counterfactual scenarios relevant to causal counterfactuals, then the discovery that these scenarios are all-things-considered impossible no longer seems particularly objectionable, nor even particularly surprising once we see how the competing demands of actual laws, actual past, and counterfactual antecedent can be inconsistent given determinism, or something close to determinism.

The new option I have proposed satisfies all the desiderata listed above, and if that is all we were concerned about it might at this point look ideal. Before reaching that conclusion, however, it would be well to look at the apparent costs of the theory. Others may disagree with me about what weight we should give each of these considerations, but I hope at least to touch on the ones most likely to occur to critics.

5. Counting the Costs

The first feature that some will take to be a cost is that the account requires impossible worlds as well as possible ones: and even those happy with possible worlds sometimes balk at ways things *couldn't* be. (See e.g. Stalnaker 2002.) Positing impossible worlds, and using them in the theory of counterfactuals, seems to me eminently worth doing for reasons entirely independent of puzzles about causal counterfactuals: see Nolan 1997 for arguments to this effect. Employing impossible worlds is not a reason *per se* to be suspicious of this account.

A second potentially objectionable feature of this theory is that it yields the result that many counterfactuals with possibly true antecedents and possibly true consequents nevertheless require evaluations of those propositions at impossible worlds to yield their truth values. Or, to put it a different way, the theory holds that the *nearest* worlds where certain possible antecedents are true are impossible worlds: they manage to be closer than all of the *possible* worlds where those antecedents obtain. This violates a plausible principle I have labelled the 'Strangeness of Impossibility Condition' (SIC) (Nolan 1997 550): the principle that every possible world is closer to every other possible world than any impossible world is to any possible world. This principle is plausible for at least two reasons. One is that it captures the idea that if something A, could obtain, its obtaining wouldn't be impossible: a natural way to try to capture this thought formally would be that, for possible A and impossible I, no case of 'if A then I' is true. Another is that if SIC is true, then something like the usual possible-worlds semantics for counterfactuals can hold in the vast range of ordinary cases when the antecedents of the conditionals that concern us are possible. Impossible worlds need only make a difference to truth conditions when impossible antecedents are in play.

Despite these advantages of the SIC, I think there are good enough independent reasons to reject it. I had already expressed some reservations about it in Nolan 1997 p 550, 569 ftnt 21, and more counterexamples can be found in Vander Laan 2004. In some contexts at least, relevant similarity seems to require that we count some impossible worlds as closer than some possible worlds.

Here is a relatively everyday example. Suppose I have been playing a game with a boy called Oliver, where we arrange balls into a square grid, then tip the balls into a bag, then count the balls that come out. Sometimes the total is 4, sometimes 9, sometimes 49.. and so on. While playing it, I introduce Oliver to the idea of a square number, in the obvious way. On a particular occasion, we come up with a count of 63 balls from the bag. 'If the bag had 63 balls in it, 63 would have been a square number' seems like an appropriate thing for me to say in explaining why I think we miscounted: and I think a true thing to say, in that context. But of course it is possible for that bag to contain 63 balls, and impossible for 63 to be square. So at least in some contexts of utterance of counterfactuals, SIC fails.

Even if we reject the SIC in full generality, however, there remains a problem. Many counterexamples to SIC involve relatively unusual cases, or at least unusual contexts of utterance. But the violations of SIC suggested by the theory in this paper are potentially far more widespread, and infect many more everyday conditionals. In deterministic worlds, almost all false antecedents in causal counterfactuals will invoke impossible worlds: and in indeterministic worlds where the laws put strict limits on what is permitted in the future given the past, many false antecedents will produce SIC violations. If we are to use the strategy of this paper to avoid semi-miracles, even cases where the past and the laws are co-possible with the antecedent may still be ones where the relevantly nearest world may be an impossible one that lacks semi-miracles rather than possible worlds containing semi-miracles. Even those suspicious of the SIC may be inclined to raise their eyebrows at the claim that goings-on in impossible worlds are relevant to the truth-value of mundane causal counterfactuals.

There are some ways to sugar this pill. If we think of people putting together beliefs about how things typically work, with beliefs about what happened before the relevant time, with beliefs about ordinary ways for the antecedent to be true at the antecedent time, and then roll the situation forward in thought, it might not be so surprising after all that the first three parts of that process might answer to standards that can come into conflict. It is agreed on all hands that we lack an a priori guarantee that the actual laws and the actual past will permit the antecedent to be the case, or permit it to be the case in a non-miracle-like way. Finding some possible situation that meets the demands well enough in a non-obvious way is a sensible thing to try: but when we discover that these worlds are unsatisfactory in various respects (containing miracles, containing different pasts, etc.), then involving an impossible situation that meets all the constraints our practice answers to does, after all, seem like a option that is not so unintuitive.

A third dubious commitment of this theory is that it requires a particular account of which worlds are most similar, in relevant respects, to the world at which the causal counterfactuals are being evaluated. Some will find the judgement of relevant similarity straightforwardly implausible. Let us suppose that a world is deterministic, and a strong man is standing, holding a hammer, next to a thin glass vase - and at that world he stands still. Why, apart from the demands of the theory, should we suppose that the most relevantly similar world where that man (or his counterpart) swings the hammer at the vase (or its counterpart) is a strange, inconsistent world, where even though the laws plus the past dictate that he stands still he nevertheless swings the hammer and the glass flies into pieces, and furthermore that according to this world, everything happens in accordance with its laws? We may be able to make some surrealist sense of such an impossibility, but surely a possibility where the laws are slightly different, or the past was different enough to produce a man ready to swing, should be counted as a more similar situation to the world we began with?

It is important to remember that not any old similarity judgement can be relied upon when evaluating these theories, and even though it is easy to provoke similarity judgements according to which the impossible worlds invoked in this theory are very dissimilar from worlds like ours, when we focus on the desiderata as giving us our standard, there will not always be a relevantly similar possible worlds that meets *that* standard, whatever else might be said for possible worlds that compromise those standards.

A fourth and final feature of this account which will make it unattractive to some is that it might sit uncomfortably with analyses of causation in terms of counterfactuals. It seems odd to hold that what causes what depends on how things are in inconsistent and impossible worlds: though which causal counterfactuals does seem to so depend. Of course, many who favour counterfactual analyses of causation have made their peace with the charge that their view seems to commit them to thinking that what actually causes what depends on the goings-on in different *possible* worlds, so many of the strategies for dealing with that challenge will be available to this view as well. It can also be protested that, in the important sense of dependence, the causal facts do *not* depend on the facts about the contents of other possible worlds. Perhaps, instead, both the causal facts and the facts about worlds depend on some deeper actual matter, or it can be pointed out that *closeness* of another possible world supervenes on what this world is like, so causal matters depend ultimately on actual matters of fact. (David Lewis's theory has this structure, for example.) Or perhaps the analysis is not one that shows that the truth of one side of the analysis *depends* on the truth of the other side: it could be claimed that it is a non-reductive analysis, or an important equivalence without one side being 'deeper' than the other.

My own view is that counterfactual analyses of causation, and other nomic phenomena like laws, chance, dispositions, difference making, and the rest are unlikely to succeed, and not just because of the technical troubles they have often faced. Still, insofar as they are plausible at all, they would seem to still be in contention even if the suggestion of this paper about how to account for causal conditionals were adopted.

6. Conclusion

We learn very easily how to use and evaluate causal counterfactuals, but it has been frustratingly difficult to capture the truth-conditions of these counterfactuals in other terms. While there is something very appealing in closest-world analyses of these conditionals, none of the standard attempts to account for them in that framework have been entirely satisfactory. The attempt presented in this paper is unlikely to bring our disagreements to a conclusion either. At best, it will be added as yet another option with distinctive benefits, but also distinctive costs. If my diagnosis of the problem we face is correct, the desiderata for a theory of causal counterfactuals appear to be in tension if we restrict ourselves only to possible worlds, so perhaps no ideal solution will be forthcoming. If there is no ideal solution in the offing, then the option set out in this paper deserves to be in the running for being the best of a group of less-than-ideal options.[6]

References

Bennett, J. (2003). A Philosophical Guide to Conditionals. Oxford: Oxford University Press.

Hájek, A. (unpublished). 'Most Counterfactuals are False'.

Lewis, D. (1973). Counterfactuals. Oxford: Basil Blackwell.

Lewis, D. (1979). 'Counterfactual Dependence and Time's Arrow', Noûs, 13: 455-76. Reprinted in Lewis 1986 pp 32-52, page numbers are for Lewis 1986.

Lewis, D. (1981). 'Are We Free To Break the Laws?', Theoria, 47: 113-21. Reprinted in Lewis 1986 pp 291-298, page numbers are for Lewis 1986.

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

Menzies, P. (2002). 'Causal Models, Token Causation, and Processes', Philosophy of Science, 71.5: 820-832

Menzies, P. (2004). 'Difference making in Context' in J. Collins, N. Hall, and L.A. Paul (eds), Causation and Counterfactuals. Cambridge MA: MIT Press, 139-180

Nolan, D. (1997). 'Impossible Worlds: A Modest Approach'. Notre Dame Journal of Formal Logic, 38.4: 535-572

Paul, L.A. and Hall, N. (2013). Causation: A User's Guide. Oxford: Oxford University Press.

[6] This research was supported under the Australian Research Council's *Discovery Projects* funding scheme (project number DP130104665). Thanks to the audience at the Doing Philosophy: From Metaphysics to Ethics Workshop at Bundanoon, NSW and to Arif Ahmed, Helen Beebee, Rachael Briggs, Adam Elga, Alan Hájek and Peter Menzies for discussion. Especial thanks also to the late Peter Menzies for his role in introducing me to the topics I have been grappling with in this paper: he will be sorely missed.

Pearl, J. (2000). Causality. Cambridge: Cambridge University Press.

Stalnaker, R. (1968). 'A Theory of Conditionals' in N. Rescher (ed), Studies in Logical Theory. American Philosophical Monograph Series, no. 2. Oxford: Basil Blackwell, 98-112.

Stalnaker, R. (2002). 'Impossibilities' in R. Stalnaker, Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays. Oxford: Oxford University Press, 55-67

Stalnaker, R. and Thomason, R. (1970). 'A Semantic Analysis of Conditional Logic'. Theoria, 36(1): 23-42

Vander Laan, D. (2004). 'Counterpossibles and Similarity' in F. Jackson and G. Priest (eds) Lewisian Themes. Oxford: Oxford University Press, 258-275

Woodward, J. (2003). Making Things Happen: A Theory of Causal Explanation. Oxford: Oxford University Press.