The Philosophical Quarterly Vol. 0, No. 0 *2025*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqae149

Backwards counterfactuals

Stephanie Rennick [1] and Neil McDonnell [2]

[1] *University of Stirling, UK and* [2] *University of Glasgow, UK*

This paper offers two novel conceptual tools: one concerning the semantics of counterfactuals and what should be held fixed when assessing them (the modal moat), and the other concerning the pragmatics of counterfactual assertions and how to avoid the potential pitfalls of meaning more than we say (antecedent gluttony). These allow us to address existing issues with the assessment of backwards counterfactuals within a framework that applies equally to forwards cases. In addition to solving a thorny problem from the time travel literature, what we learn teaches us something quite general about our evaluation of counterfactuals.

Keywords: counterfactuals; backwards; time travel; causation; Lewis.

I. Breadcrumbs

> *Niamh stumbles upon entering her time machine*[1] *in 2020, banging her knee in the process.*
> *Upon exiting in 1970, she looks down to discover a bruise blossoming on the aforementioned joint.*
> *'Jings, crivens!', exclaims Niamh. 'If I hadn't banged my knee, I wouldn't have this awful bruise!'*

How do we make sense of Niamh's claim? It seems plausible that the counterfactual conditional expressed is true, and that she might reasonably infer on this basis that the knee banging caused her bruise. But unlike ordinary cases where antecedents precede consequents, and causes take place earlier than effects, this is a case of backwards causation, where the antecedent of the

Correspondence to: Neil McDonnell, neil.mcdonnell@glasgow.ac.uk

[1] Given our focus on time travel we will assume an eternalist/four-dimensionalist conception of time throughout (as per Lewis 1976). However, our argument can be applied to any immutable theory of time (i.e. any theory that posits moments only occur once) that permits backwards causation.

counterfactual lies five decades in the future of the consequent. Nonetheless, Lewis (1979) writes, '[c]areful readers have thought they could make sense of stories of time travel…it will not do to declare them impossible *a priori.*'

Articulating how to evaluate counterfactuals of backwards causation is the purpose of this paper. We call these 'backwards counterfactuals'.[2]

The standard Lewisian treatment of counterfactuals has us consider worlds in which the antecedent is true (unlike the actual world) but which are otherwise as similar as possible to the actual world, and then consider whether the consequent would also be true in those worlds. Importantly, we are to consider worlds which share the laws and history of our world as much as possible—they are considered *closest* on Lewis's account (see Section II). So, when considering these alternate possible worlds, we *hold fixed* certain features of the actual world, and *leave open* others (such as the antecedent).

With this in mind, here is one way we might proceed when considering whether Niamh's counterfactual is true:

1. Hold fixed everything in the past of the antecedent

A rule of thumb many philosophers of causation use when evaluating ordinary causal counterfactuals[3] is to hold the past of the antecedent fixed.[4] When we contemplate whether the light would have come on if we had pressed the switch, we don't vary the lead-up to the switch pressing: we assume that there was a switch, and a light, that we're not underwater, that the earth is round, and so forth. It is a useful heuristic and generally yields truth values for the counterfactuals that match our causal intuitions.

But holding the past fixed up to the antecedent doesn't work in cases of backwards causation, where the antecedent comes after the consequent. In Niamh's case, the standard heuristic would have us hold fixed the consequent—the fact that Niamh has a bruise in 1970[5]—whether or not the antecedent (the knee banging) occurs, because the consequent lies in the past of the antecedent (and thus in the region we hold fixed to its actual world state). In other words, by fixing the past in which Niamh has the bruise, we *screen-off* the influence of the bash on whether the bruise appears. This would render all backwards counterfactuals with a nonactual consequent false by fiat. That is no use.

[2] As distinct from 'backtracking' counterfactuals (see Section II).

[3] We take it that the total class of counterfactual conditionals is broader than the sub-class that concern the causal relata, and we take it that the criteria that Lewis (1979) derived through his examples are tailored specifically to those concerning the latter. This is somewhat controversial but won't be defended here. What matters is that we restrict our claims to only those counterfactuals that concern distinct events in Lewis' sense, and thus we refer to these as 'causal counterfactuals'. Unless we specify otherwise, 'counterfactual' is used as shorthand for this particular subset.

[4] For example discussed in Lewis (1979), Mackie (2014), and Fernandes (2021).

[5] For clarity: the bruise becoming visible is the start of the event we are concerned with, and this event occurs after Niamh has left the Time Machine in 1970.

However, 'hold the past fixed' isn't quite the rule as written. The system for evaluating causal counterfactuals is designed to hold fixed potentially confounding factors and vary only that which is being considered as a candidate cause. In practice—given the ubiquity of forwards causation—this leads to an asymmetric rubric which involves holding fixed everything in the past of the antecedent. But what Lewis (1979) actually writes is that we should 'maximize the spatio-temporal region throughout which perfect match of particular fact prevails' (without creating wide deviations in law). Unlike the heuristic, this is symmetric in respect of past and future: it doesn't specify *which* spatio-temporal region we want to match to the actual history (i.e. which bits to hold fixed).

Given that holding the past fixed tends to give us the right answer in ordinary cases, we might invert it for cases of backwards causation:

2. Hold fixed everything in the future of the antecedent

But this doesn't work either. As discussed below, our motivation for holding anything fixed is to ensure that we evaluate against worlds similar to ours in the relevant respects, and to avoid confounding factors. We want to know if the antecedent made a difference to the consequent—if the former caused the latter. But if we pick out the closest worlds by holding fixed the future of the antecedent, among those closest worlds may be those with significantly different circumstances (up until 2020). Perhaps Niamh doesn't bash her knee because there is no time machine, she has no knees, or Caesar never crossed the Rubicon. In attending too much to the future and not at all to the past, Niamh's counterfactual ceases to be a reliable test for whether bashing her knee made a difference to her bruise.

To evaluate counterfactuals like Niamh's, we need a different rule for what we hold fixed. In this paper we identify three desiderata of a theory of evaluating counterfactuals in order to reach such a heuristic, and thereby provide a way to understand Lewis's similarity rules symmetrically with respect to time. First, we sketch Lewis's account (Section II), and then consider and reject other candidate heuristics (Section III). In Section IV we turn to the desiderata and present our semantic proposal—the *modal moat.* Then we identify a much-overlooked pragmatic element in the assessment of counterfactuals, which we name *antecedent gluttony* (Section V). Finally, we consider a problem case (Section VI) and bring all the pieces together (Section VII). By journey's end, we endorse a familiar (Lewisian) framework with two crucial tweaks.

II. Background

As is familiar, Lewis (1973) used his possible-worlds semantics for counterfactual conditionals to develop a counterfactual theory of causation. Due to his reductive ambitions, his account of the closeness/similarity of worlds is not based on *causal* similarity.[6] Lewis offered the following rubric for his similarity ranking:

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact obtains.
3. It is of the third importance to avoid even small localized simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular facts, even in matters that concern us greatly (Lewis 1979).[7]

Much of our discussion will focus on the second condition: which spatio-temporal region we should 'hold fixed' (seek perfect match within).[8]

We follow Lewis's lead on three points of methodology. First, the theory behind the semantics of counterfactuals is considered successful only insofar as it concurs with the intuitive reading of the truth value of (ordinary) counterfactual statements. This classically Lewisian approach privileges common sense.

Second, Lewis's rules are aimed at codifying the assessment of counterfactuals in the 'standard' context. This rules out certain known-but-outré ways that counterfactuals might be used (such as backtracking counterfactuals; Lewis 1979). We follow Lewis in setting these cases aside, but note that the backwards counterfactuals that interest us here are ones that Lewis thought of as *non*standard. We think his rules *as written* probably do cover such cases, but require certain clarifications.[9]

The final methodological point is more controversial: we allow our causal intuitions to help guide what the counterfactual semantics should report. The nature of the relationship between causation and counterfactuals is itself controversial, and (in the Lewisian project) the direction of fit is supposed to be that we analyse causation in terms of counterfactuals, not the other way around. However, it is uncontroversial that there is some close connection between the counterfactuals that we are willing to assert and the causal claims we think are true. What *is* contentious, in light of a range of problematic cases (e.g.

[6] For benefits of relinquishing this ambition, see Schaffer (2004) and Wasserman (2015).

[7] It is interesting to note that the broadly recognized (Ichikawa 2011; Lewis 2016) context sensitivity of counterfactuals is not captured directly in this rubric. Our Section V offers a route to capturing at least some such context sensitivity, but is not offered as a general solution.

[8] There are some well-known problems with this condition which we do not have the scope to discuss here (cf. Pollock 1976); ultimately we provide a modification.

[9] Much of the literature has been concerned with whether Lewis's account adequately rules *out* illegitimate backwards causation (e.g. backtracking) rather than in explaining how we might evaluate counterfactuals where backwards causation *does* occur. Our focus is the latter.

overdetermination/pre-emption), is the idea that counterfactual dependence between two events is necessary for there to be a causal connection between them. For the task at hand, we stick to cases without the known problematic structures so as to allow ourselves the additional diagnostic data of causal intuitions.[10] Further, we use intuitions only as an aid in refining the counterfactual semantics, not as an analysandum. We are not assuming that this is the direction of explanatory or ontological fit. Rather, we use the intuitive data about causal claims to help us derive what a consistent and holistic account of counterfactuals and causation *ought* to say. The success of the resultant account will be judged on its ability to shed light on our problem case of backwards counterfactuals.

III. Possible solutions

At the outset, we rejected (1) the standard heuristic—hold fixed everything in the past of the antecedent—and (2) the backwards rule (hold everything fixed in the future of the antecedent). The former holds fixed too much in cases of backwards causation, not leaving open the matter for testing, and the latter holds the wrong parts fixed for relevant similarity of worlds. Here we consider and reject other candidates before presenting our own solution in Section IV.

3. Hold fixed the external past of the antecedent and the personal past of the time traveller

Standardly, discussions of time travel differentiate between external time (time itself) and personal time (which fulfils the same role as external time does for the nontime traveller). In time travel cases these come apart: Niamh's time machine traverses fifty years of external time, but within the time machine it may only have been five minutes (as measured by Niamh's smartphone clock or the aging of her cells) (Lewis 1976). When she arrives in 1970, 2020 lies in the external future, but Niamh's personal past. So, we might think, we could hold fixed the external past of the antecedent, as usual, and additionally hold fixed events in Niamh's personal past (some of which occur in the external future relative to the time in question). This allows us to overcome the problem with (2) by ensuring that relevantly similar worlds are considered.

[10] Some might wonder how much weight to put on our intuitions about time travel scenarios. This isn't something we have scope to litigate here, but it seems to us that the plethora of time travel narratives, and the relatively constrained range of tropes therein, suggests that there are limitations as to what people are willing to entertain in time travel cases, and some widely held views about what is possible and/or likely (Rennick 2021). Many philosophers of time travel, and metaphysicians more broadly, seem to be operating under the assumption that our intuitions are at least a valuable starting point when thinking about issues like backwards causation. Suffice to say that our use of intuitions here is in keeping with many of the interlocutors we are engaged with in this paper.

But this solution doesn't work because it holds fixed that Niamh is present in 1970 whether or not she gets into the time machine in 2020. Option (3) holds strictly more fixed than the already-problematic option (1) and so both screen-off the influence of the antecedent on the consequent. We need to hold less fixed.

4. Hold fixed the external past of the consequent and the personal past of the time traveller prior to the antecedent

This improves on (3) by leaving open the important region around the consequent in 1970, and thus avoids problematically guaranteeing that Niamh will be nursing a bruise in 1970 whether or not she gets into the time machine. It remains problematic, however. Suppose that Niamh remains in 1970 and, despite a brush with death in 1980, lives to attend her own birth as an adult in 2000. It seems true for Niamh to say, 'If I had died in 1980, I wouldn't have attended my own birth as an adult'. However, under (4) this would be false since the consequent (Niamh's birth) lies in Niamh's personal past and so should be held fixed.[11]

One might respond by refining what 'the personal past of the time traveller' includes, for example, only things which Niamh remembers. But since time travelling rocks presumably remember nothing, this won't generalize to all time travellers. We could instead restrict the personal past to only cover the regions of history that contain the object itself, but this won't help once adult Niamh gives her baby self a cuddle. There are objects whose 'personal' past includes regions that we must hold open for the sake of assessing the counterfactual. We need to reject (4).

5. Hold fixed 'what the relevant deliberating agent has reliable evidence of, independently of her decision (using a sufficiently externalist notion of evidence)' (Fernandes 2021).

Fernandes's approach is distinct from the options discussed above; she holds some future events fixed and the past open. Evidence is interpreted broadly and includes not only photographs and written records but also memories. Fernandes (2021) writes that 'whatever causal or nomic mechanisms allow us to have records of the past in the actual world allow backwards time travellers…to have records of the future that are as reliable as our usual records of the past'. The mere existence of such evidence justifies our holding fixed some future events when evaluating counterfactuals. We should not hold the past fixed when evaluating backwards counterfactuals, allowing 'past events to change in counterfactual worlds'.

[11] This last case is a forward counterfactual, but it remains a case where the putative solution in 4 fails to match intuition. Our own solution will not have this drawback.

Although Fernandes merely 'sketches' this approach, noting more would be needed to articulate and defend it, there are some apparent limitations which our own solution (Section IV) overcomes.

First, (5) seems to leave room for cases in which the time traveller lacks evidence of future events, because no such evidence exists (e.g. an amnesiac time travelling from now to the distant post-apocalyptic future, and then back to 1950). This means that there could be some 'close' worlds in which the immediate past of the antecedent is very different—if the time traveller forgets, or simply does not know, crucial aspects about how the time machine works, then it can work very differently in Fernandes's 'close' worlds. This inherits the problem discussed concerning (2) where the antecedent worlds are too different from our own.

Secondly, the account is built around deliberating agents. Fernandes (2021) notes that even if the time traveller can't deliberate, 'we can still consider what counterfactuals would obtain for a properly deliberating agent in a relevantly similar situation' and thereby 'use counterfactuals to recover the causal structure of the case'. It's not clear how this might apply to time-travelling particles or pieces of information (which presumably can be the subject of a backwards counterfactual, and yet cannot deliberate).

Thirdly, while Fernandes argues that we shouldn't hold the entirety fixed, she doesn't give an account of which past events we *should* hold fixed.[12] It's possible Fernandes could remedy this by means of her earlier criterion, holding fixed what the agent has evidence of, but it's unclear how much of the past this would apply to since it will presumably vary with the competence, knowledge, and species of the deliberator. This risks building in a level of subjectivity and variability in the truth conditions of backwards counterfactuals that we do not accept in their forward-looking equivalents. Our semantic account, by contrast, specifies which past events should be held fixed irrespective of who is deliberating, and ensures large regions of perfect match between worlds to guarantee their relevance to the counterfactuals under consideration.[13]

Finally, while Fernandes describes her account as temporally neutral, she concedes that we might reasonably evaluate forwards counterfactuals differently depending on context. Our proposal works without modification in either direction.

Fernandes (2021) notes that Lewis's method of evaluating counterfactuals is 'global in character, seeking perfect match between largest possible spatiotemporal regions, and so has trouble capturing the local variations of causal order

[12] Fernandes clarifies that she doesn't think we should hold the whole past fixed, but neither should we leave the whole past open.

[13] The pragmatic dimension of our account is sensitive to these deliberations but it only fixes *which* counterfactual is being asserted, not its truth.

in cases like backwards time travel'. We now turn to our method, which seeks to preserve the former while overcoming the latter.

IV. The modal moat

The rejected solutions are all consistent with Lewis's (1979) four-fold rubric. They were rejected for holding too much, too little, or the wrong things fixed. We identify three desiderata for a functioning theory of counterfactuals: trajectory, openness, and relevance. Lewis's rubric implicitly satisfies these in forwards cases; by making them explicit, we can also evaluate backwards counterfactuals. (NB these desiderata—and the resulting amendments—are specific to evaluating counterfactuals, they are not general principles regarding the ordering of worlds within this kind of modal metaphysics).

Some period of the past must be fixed just prior to the antecedent to ensure the set-up of the case is relevantly similar to the actual world (the atmosphere contains oxygen, Niamh still has knees, the time machine exists). This is the *trajectory* requirement. It ensures that when we consider the possible worlds where Niamh does not bash her knee, any further deviations in that world relative to ours are traceable to that counter-to-fact alteration, not spurious confounding factors that would be present if we didn't hold the recent past fixed. For illustration, consider a still-frame of a game of snooker after a shot has been taken. The position of each ball is evident, but the movement or *trajectory* is not. Such a snapshot is consistent with a wide range of different game states—every ball could be moving, in any direction, or none could be moving at all. With so many live possibilities, we cannot confidently assert even simple counterfactuals such as 'if I were to lift the green ball, the white would hit the cushion'. This is analogous to specifying the antecedent region and nothing else—that state of affairs is consistent with a wide range of different histories, futures, and laws. However, if we see even a very brief video of the snooker shot prior to that still-frame, we would be able to rule out many of the possibilities consistent with that frame but not consistent with the video. That would restrict our considerations to only those possibilities relevant to *this* game; the additional information provided by the video would make it clear what difference lifting the green ball would make. This is analogous to specifying what happens in the region of the antecedent *and* fixing at least a brief period of the history building up to it. It serves to align the immediate historical trajectory with the actual world and ensure local relevance.

As we saw above, in cases of backwards counterfactuals, we must hold at least some of the past open (minimally the region of the consequent) on pain of screening-off the impact of 'wiggling' the antecedent. This is the *openness* requirement. (Extending the snooker analogy: holding the consequent region fixed is like gluing the target ball in place.) In ordinary forwards counterfactuals we automatically leave the consequent open—it contributes not-at-all to closeness considerations—and fulfil this requirement by default whether we hold some or all of the past fixed: in forwards cases it's safe (non-confounding) to hold the entirety fixed because the counterfactual relata both occur after the fixed period.[14]

So, when evaluating counterfactuals, we need to hold a region fixed (for trajectory) and another open (for openness); in backwards cases, these are both in the past of the antecedent. We also need to ensure as much is held fixed as possible to maximize match, without accidentally screening-off dependencies of interest. As the former is what fixes the similarity of the world more broadly, we call it the *relevance* requirement. In forwards counterfactuals, the entire past of the antecedent is held fixed. In backwards cases, we need to fix a period in the past of the consequent.[15] Instead of having two heuristics here, however, we note that the past of the *earliest* relatum is held fixed in both cases. We will typically gain much in terms of overall match, and accrue no additional risk of confounds, by fixing the entirety of the past prior to the earliest relatum.

Bringing trajectory, openness, and relevance together, it is clear that when evaluating backwards counterfactuals, we need the past fixed before the consequent, open at the region of the consequent (minimally), and fixed again for some period before the antecedent. In other words, we need an open or fluid period between the antecedent and the consequent, which has fixed regions at either end. The fixed-open-fixed structure resembles a moat with solid ground abutting something fluid, and so we call it a *modal moat*. The modal moat is what is needed to meaningfully assess the truth of backwards counterfactuals.

Fig. 1 depicts the trajectory, openness, and relevance requirements applied to the original Niamh counterfactual: 'If Niamh hadn't bashed her knee, she wouldn't have had a bruise.' Note the distinctive 'modal moat': the modally fluid open period sandwiched between two fixed periods.

Fig. 2 shows an ordinary forwards case featuring Niamh's cousin Eve. The counterfactual under consideration is: 'If Eve had not bashed her knee, she would not have had a bruise the next day'. Note how the trajectory and relevance requirements overlap (darker grey) to give two independent reasons to fix the period before the antecedent; we may have never realized that we needed them both had we only ever attended to forwards cases.

[14] Openness is not a feature of a given world, but rather of a set of worlds: in any world, what happens in the region in question is 'fixed' one way or another, but if what happens in that region varies *across* the worlds under consideration, then that region is not 'fixed' to actual values. Regions that are 'open' in our sense do not contribute to the assessment of closeness of a given world.

[15] In forwards cases, this rule would cement the antecedent at its real-world truth-value, rendering the conditional trivially true.

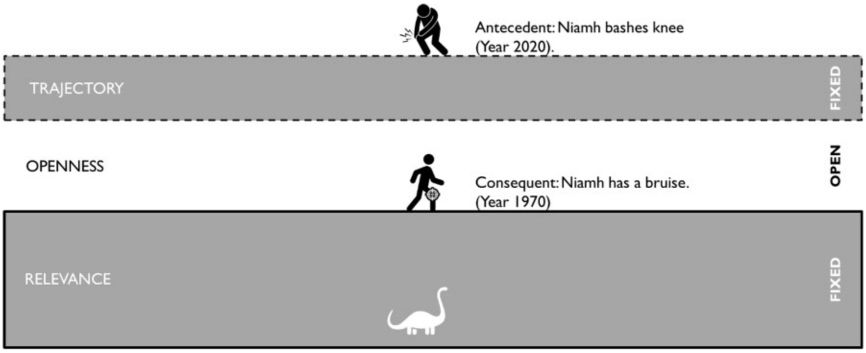

Figure 1. Niamh's modal moat.[16]

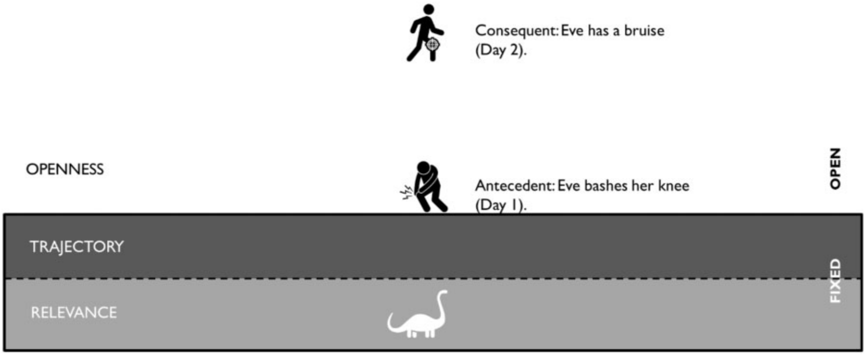

Figure 2. Eve's modal shore.

IV.1 Size

Having identified the modal moat structure, we can start to consider the principles governing where, and how large, these fixed and open regions should be.

The relevance requirement guides the position and extent of the first region: everything in the past of the earliest relatum is held fixed. In backwards cases, this is the entire past of the consequent.[17]

We need to fix the period of trajectory to ensure relevant similarity on the build-up to the antecedent of the conditional. When evaluating forwards

[16] We believe our account is compatible with relativity; the diagrams are meant to serve as a visual aid to the examples, but are not meant to suggest e.g. a literal (nonrelativistic) time slice of the universe.

[17] When exactly the bruise event begins is another dimension of vagueness that influences our view, and any others concerning events. We won't explore it here, but suffice it to say that wherever the event begins, the past prior to that point is held fixed. In our story, the bruise emerges in 1970 after Niamh leaves the Time Machine.

counterfactuals, the period fixed for trajectory overlaps with the larger period fixed for relevance (Fig. 2) and so its size is irrelevant; by contrast in backwards cases the two are separate—in our case, the larger before the consequent and the smaller before the antecedent (Fig. 1). The greater the extent of match with the real world, the more confident we can be of a world's relevance, so there is upwards pressure on how much of the past of the antecedent matches actuality in backwards cases.

However, the open region around the consequent acts as a limiter on how much of the antecedent's past matches actuality. It is difficult to be precise about the size of the open region, as it requires balancing the match that comes with fixing a bigger region against the type of law violation (typically large and widespread) required to ensure the history converges with the actual history before the antecedent occurs. Lewis (1979) extolls the vice of such convergence miracles in the pursuit of closeness. As currently presented, a consequence of the modal moat structure is that an open region must converge with a fixed region—future similarity (Fine 1975; Wasserman 2006) from the perspective of the consequent is *required*. This would suggest that it is simply a feature of backwards counterfactuals that they require convergence miracles. We address this (Section VI), but there is one last puzzle piece to be introduced first (Section V).

For the moment, we can attend to Lewis's observations about convergence miracles: that they are generally less widespread, and less deviant from our own laws, if they happen earlier (in worlds where causation is predominantly past to future). Thus, we have reasons to *maximize* the region fixed prior to the antecedent, but also to find the 'cheapest' convergence miracle available. This will typically mandate having an early convergence miracle after the open period, but cases may vary.[18] The governing principle remains the same as Lewis laid out: first minimize big miracles, then maximize match.

IV.2 Symmetry

When characterizing the modal moat, we used the temporally loaded terms 'past' and 'earlier' to specify which regions of the world are held fixed. This may sound like built-in temporal asymmetry, which is something that we objected to in other proposals; but it isn't.

We take 'past' ('earlier') to refer to the temporal direction relative to Time's Arrow at a world which, following Lewis (1979), we take to be derivable from the relative size of miracle required in each direction to achieve

[18] This is for the general reason that as effects propagate through the world they create more and more changes, each of which requires a minor miracle to undo. An earlier intervention on the process will therefore typically require a smaller miracle to achieve reconvergence. See Lewis (1979, p. 471) for the origin of this reasoning.

match with counterfactual worlds. Thus, the asymmetry of miracles tells us which direction should be considered 'past' ('earlier'), and which should be considered 'future' ('later') at a given world.[19] In a world where the arrow is inverted relative to the actual world, the referents of 'past' and 'future' are too.

What if there is a world with no such direction? Lewis (1979) considers a single atom world, but perhaps more salient is a world where the prevalence of causation is roughly equal in each direction. There, Time's Arrow points both ways, and we have no good reason to select which is the past and which is the future. Does that undermine our view? Again, we think not. The reference to the past and the future when thinking about what to hold fixed reflects the common heuristics by which we can ensure the real target: overall similarity. In a world with no clear temporal direction, those heuristics are not (or less) useful. Nevertheless, the principles behind similarity remain unscathed: minimize big miracles, ensure maximum match, then minimize small miracles. A strange world without clear temporal direction may complicate our assessments, but it needn't undermine the overall Lewisian approach that we are aiming to refine.

V. Gluttonous antecedents

The modal moat approach to what we hold fixed seems to get the right result that Niamh spoke truthfully when she said 'If I had not bashed my knee, I wouldn't have this bruise'. This treatment refines Lewis's rubric, and retains the attractive temporal symmetry that Lewis considered a desideratum of a theory of counterfactuals. However, once we embellish the Niamh case complications arise which lead us to propose a pragmatic device in our interpretation of counterfactuals more generally.

Suppose that Niamh, having travelled to 1970 and noticed her bruise, seeks out a pharmacy for treatment. Having no appropriate currency from the era, she steals arnica ointment (a bruise will wreck her mission, we suppose). She curses her clumsiness, exclaiming: 'Bashing my knee caused me to steal!'. This sounds like a true causal claim, and thus we should expect a corresponding true counterfactual such as: had she not bashed her knee, she would not have been stealing. However, our initial modal moat approach doesn't get that result because it advocates holding fixed the past of the consequent (the earliest relatum). Suppose that Niamh arrived in 1970 with a bruise at noon and stole the arnica at 1 pm. In this scenario, the consequent is the act of stealing, which happens after the bruise has developed.

[19] As per footnote 13, we take our account to be compatible with relativity; 'past' and 'future' should not be read as referring to an absolute past or absolute future.

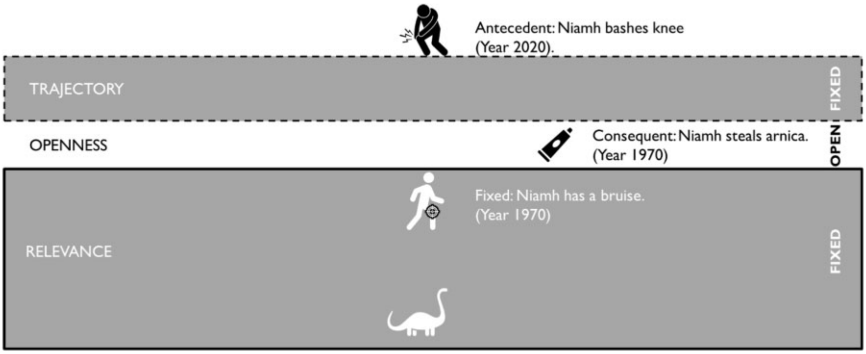
Antecedent: Niamh bashes knee
(Year 2020).
TRAJECTORY
FIXED
OPENNESS
Consequent: Niamh steals arnica.
(Year 1970)
OPEN
Fixed: Niamh has a bruise.
(Year 1970)
RELEVANCE
FIXED

Figure 3. Niamh steals arnica.

As Fig. 3 illustrates, the modal moat holds fixed an intermediate causal point (the bruise) on the path from bashing the knee to stealing the arnica. By holding that point fixed, we screen-off the influence that the bash had on the shoplifting, since the bruise is guaranteed in all the scenarios under consideration. This is a variant of the problem that led us to reject the standard treatment in Section I; something has gone wrong here.

It is tempting here to appeal to the solution Lewis (1973) uses in response to early pre-emption cases and break the causal claim down into component steps (A causes B, B causes C) and argue via transitivity that A causes C without there being C-A counterfactual dependence. That would only work for the causal claim, however. It would not work for the closely related counterfactual assertion: ‘If I hadn’t bashed my knee, I wouldn’t have been stealing!’. The modal moat approach considers this statement (taken at face value) to be false. Niamh is just wrong. Is that the right result?

If it seems like Niamh has said something *true* here, then either the modal moat approach is incorrect, or there is an additional interpretive step required between the words Niamh utters and the proposition she expresses. The need for such an interpretive step was previously identified by Bennett (2003: §65–66, see especially p. 162) in relation to antecedent strengthening and transitivity in the logic of counterfactuals. We think that the phenomenon is perhaps more general than previously identified, and that it deserves its own name, so we dub this interpretative step *antecedent gluttony*. Next we introduce the concept and argue that it explains the apparent, but not literal, truth of the uttered counterfactual without affecting how we read the associated causal claim. This lesson generalizes to the interpretation of counterfactuals more broadly.

In the case described, both we as observers, and Niamh as a participant, are all-too-well aware of the intermediate event (B)—the bruise blossoming in 1970. When we think about the closest possible worlds in which she does *not* bash her knee entering the time machine (worlds in which event (A) does

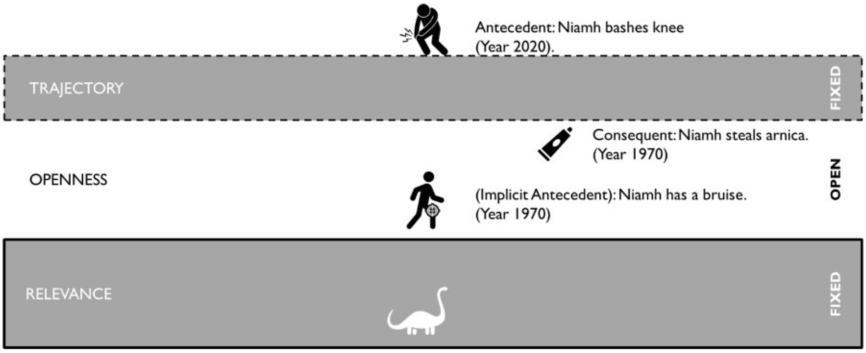
Antecedent: Niamh bashes knee
(Year 2020).
TRAJECTORY
FIXED
Consequent: Niamh steals arnica.
(Year 1970)
OPENNESS
(Implicit Antecedent): Niamh has a bruise.
(Year 1970)
OPEN
RELEVANCE
FIXED

Figure 4. Niamh steals arnica (open).

not occur), we also implicitly consider her not to have a bruise in 1970 (we assume that event (B) does not occur either). So, we propose that the connection between (A) and (B) is not actually separate from what Niamh asserts in the counterfactual statement: it is *built-in* to it. More precisely, we take the fact that (B) occurred to be an implicit element within the scope of the negation of the antecedent in Niamh's asserted counterfactual, and thus it must remain *open* to variation and not *fixed* to (B)'s actual world state (Fig. 4).

Instead of representing Niamh's counterfactual statement literally as ¬A □→ ¬C, it should instead be represented by something like [¬A (& ¬B)] □→ ¬C, where the round parentheses indicate the implicit element. If Niamh had not bashed her knee in 2020 (and gotten a bruise in 1970), then she would not have been in the pharmacy stealing arnica in 1970.

By way of justification for this reading, consider how someone would read the counterfactual had they only been told part of the story: merely that Niamh bashed her knee getting into the time machine, and that she was shoplifting in 1970. Absent the information about (B)—the bruise—they would plausibly fail to see the connection between the two events, and be dubious about Niamh's asserted counterfactual. But Niamh is fully informed about the bruise, and so are we as we search our intuitions about the case. Thus, when we interpret the counterfactual, we implicitly build in that additional content into the antecedent.

Once we see that antecedents can be implicitly enriched by the speaker's knowledge in this way, we might worry that the floodgates open. Niamh hasn't forgotten that Caesar crossed the Rubicon, grass is green, or that she ate an egg for breakfast that morning. All manner of additional information could be smuggled-in to the antecedent in this way, making for a very rich antecedent indeed. However, it is implausible that Niamh had all of these in mind at the time of her assertion, so we think it implausible to consider them all part of her assertion.[20] On the other hand, it is both plausible and likely that Niamh was thinking about her bruise when she asserted the counterfactual about the bash and the visit to the pharmacy it precipitated.

None of this mind-dependent interpretation of Niamh's assertion alters the causal structure of the world. The bash *caused* the stealing and no mistake. We know already that causal connection does not imply *overall* counterfactual dependence, so the causal structure does not settle the truth of the counterfactual about the bash and the stealing. That is, the counterfactual 'If I hadn't bashed my knee, I wouldn't have been stealing!' can still be false if there is a causal connection here, and we claim that it *is* false unless you implicitly build into the antecedent more than is said. We think most people do just that, and so hear the counterfactual as true.

Let us call any antecedent that implicitly builds in more than its literal interpretation *gluttonous*. The most important thing to note about gluttonous antecedents is that they can, and often do, change the truth value of the counterfactual. The closest $\neg$A (& $\neg$B) worlds will often be further away than the closest $\neg$A worlds, and thus counterfactuals that take the first as the antecedent may differ in truth value from those that take the second (even though the consequent remains the same). Gluttonous antecedents are easy to mistake for literal antecedents, in part because the additional content is implicit, and also because antecedent strengthening in other contexts is innocent: it does not alter the truth of the more-familiar material conditional, for example.[21]

Three further notes about gluttonous antecedents:

First, the pejorative term here signals that there is something deleterious about the impact smuggling implicit information has on our philosophical attempts to analyse counterfactuals. That isn't to say that the enriching of the antecedent by implicit information must lead to falsehood or is unjustified. On the contrary, enriched antecedents are a natural and efficient way to communicate. They just create a gap between our intuitive reading and attempted analyses that take the words uttered at face value.

Second, there is nothing to stop *future* information being built into the antecedent too (as becomes important in Section VI). Consider Fred who was due to board the Titanic as a watchman but was waylaid and missed the boat. When reading of the sinking in the newspaper, he may say, 'If I'd boarded, I

[20] It won't matter for our view if you think Niamh holds more fixed (consciously or otherwise) than we do—that just means that you take Niamh to be building in more detail about the counterfactual scenarios she is evoking. So long as the details Niamh builds in are true of the actual world too, and remain outwith the scope of the negation, that will just *ensure* the closeness of those worlds, not confound the assessment. See the Holly case below for false presuppositions, and Section VI for more on gluttony's role in ensuring closeness.

[21] Lewis highlighted antecedent strengthening as a potential fallacy of counterfactual reasoning (1973).

would've had to abandon ship'. Here it is clear that Fred is holding fixed the Titanic's sinking: 'If I'd boarded (and the ship had sunk), I would've had to abandon ship'. Read literally, it could look as though Fred speaks falsely since, as watchman, he may have averted the sinking altogether. But that is not the possibility Fred is considering, and so we are uncharitable if we interpret him that way. Identifying the gluttony in antecedents enhances interpretive charity.

Third, false presuppositions held by speakers should not make false counterfactuals seem true. Suppose that superstitious Holly was also supposed to be on the Titanic, also missed the boat, and believes (falsely) that her lucky socks kept boats afloat. She thinks 'had I been on the Titanic it wouldn't have sunk' is true, and implicitly holds fixed in the antecedent both that she was wearing the lucky socks on the day in question and that they had the powers she superstitiously believed. Thus, in the closest worlds where her socks are as magic as she imagines, the boat does not sink and so the counterfactual would appear to be true. It would stretch charity beyond breaking to read Holly as having spoken truly here, but that is what the gluttony of the antecedent would deliver if we built her superstitious worldview into the assertion. So, we constrain the acceptable deviations from actuality in the antecedent to only those which are both *psychologically available* to the speaker and which are also *taken to be deviations from actuality* by the speaker. Holly does not realize that she is building in *two* counter-to-fact elements into her antecedent: boarding the Titanic and having magic socks. Our constraint rules out the latter as contributing to the gluttony of the antecedent precisely because she thinks it isn't counter-to-fact. Holly would (we trust) revise her counterfactual assessment that the Titanic would not have sunk with her aboard upon learning that her socks weren't magic. So, gluttony can build more detail into the antecedent—be that information that fits with the real world, or deviates from it in specific ways—and change *which* counterfactual is really being asserted. This does not change the truth conditions of any settled counterfactual assessment, or render them problematically *subjective* (which we criticized Fernandes' account for in Section III). Note that in our proposal, subjectivity plays a part in identifying which counterfactual is being asserted, not how we evaluate it. Gluttony belongs to the pragmatics of counterfactual assertion, and in its current presentation, the modal moat belongs to the semantics.

VI. Future similarity redux

We now return to the issue of *big* miracles, backwards counterfactuals, and a problem case that we think is everybody's problem. This is a redux of the old Future Similarity problem.

If Nixon had pressed the button, there would have been a nuclear holocaust (Fine 1975).

This counterfactual seems true, but a world with no holocaust is more similar (closer) than a world with one. So, in the closest button-press worlds, there should still be no holocaust. Lewis (1979) responds that once Nixon presses the button, countless tiny traces ripple out throughout the universe, creating widespread deviation from actuality. Only an equally widespread deviation from the actual laws of nature could 'put the genie back in the bottle' so as to ensure a perfectly-matched future thereafter (convergence). If we prioritize having no such 'big miracles' ahead of having perfect match over a region, then the button-pressing-but-no-holocaust world—turns out to be further from actuality, despite perfect future similarity, than some button-pressing-holocaust-world. The first requires a big miracle (or several) to achieve the future match.

Whatever one thinks of this solution, it is clear that Lewis offers an indirect way to avoid Future Similarity problems—indirect because rather than ruling that future similarity does not count towards closeness, the work is done by the presence of big convergence miracles. With this in mind, we return to Niamh.

VI.1 Bruises, bashes, and future similarity

So far, we have proposed that when evaluating a backwards counterfactual, we leave open what happens in one period—openness—and fix (to actuality) what happens in a subsequent period: trajectory. This is a recipe for a big convergence miracle since, whatever else happens in the world in the open period, we must be back in *perfect* agreement with actuality by the time we reach the trajectory period. This guarantee of a big miracle allows for a redux of the Future Similarity problem.

The guiding intuition in the Niamh case is that the counterfactual 'if Niamh had not bashed her knee, she would not have had a bruise' comes out true. Now, consider two counterfactual worlds: w1 with no bash and no bruise, and w2 with no bash but nevertheless a bruise. If w2 is closer to the actual world (w@) than w1, the counterfactual is false on the Lewisian semantics.

In this example, w2 mirrors Fine's case where Nixon presses the button but the holocaust miraculously fails to occur, and thus a long stretch of that world's future (relative to the antecedent in the Fine case, and the consequent in Niamh's case) matches actuality. The miraculous presence of the bruise in 1970 ensures match between w@ and w2 up through the openness and trajectory periods—the period of match extends to 2020. By contrast, w1—by lacking the bruise in 1970—diverges from w@ thirty years earlier. Lewis would note that in both the Nixon and Niamh cases, the miracle required to ensure match (given the button press in the former, and the lack of bang in the latter) is a large convergence miracle, which renders the world in question distant.

In Fine's case, as Lewis diagnoses, the rival button-press-holocaust world does not contain such a convergence miracle, and so is closer despite having less overall match. In Niamh's case, the rival world w1 *does* require such a miracle—it is guaranteed by the need to converge with the trajectory period (which, as per the modal moat, we are holding fixed). So now we are trading big miracles between w1 and w2. The general rule of thumb is that the earlier a convergence miracle is deployed, the smaller it can be (the ripples have not gone so far). The convergence miracle in w2 occurs in 1970 with the emergence of the spontaneous bruise. In w1, it occurs later—somewhere between the failure of the bruise to form and the start of the trajectory period. Thus, w2 is closer than w1 and our guiding counterfactual comes out as false.

Lewis' indirect solution to the original Future Similarity problem does not solve our redux version as it stands.

VI.2 Solving the problem

What should we do about it?

We could abandon the prioritization of avoiding Big Miracles. This would tempt those who thought it a poor approach already (e.g. Wasserman 2006). More conservatively, and drawing on the resources established above, we propose that the trajectory requirement is weakened to allow for *approximate* match within that period, and not *perfect* match.[22] This means that we no longer need the miraculous returning of every photon to its actual world position, every electron to its valence, and every ripple of gravity to be undone. We no longer require a convergence miracle.

Lewis argued against the relevance of imperfect match in formulating his rubric, but he was considering forwards cases where relevance and trajectory overlapped. When they come apart—as they do in backwards cases—there is scope to pay closer attention to trajectory and what it requires, without changing the verdicts in the forward cases Lewis was considering.

To meet the *relevance* desideratum we still require perfect match over a large period prior to the earliest relatum (in addition to avoiding big miracles). To ensure that the counterfactual we are enquiring about is what we take it to be, however, we only need the *trajectory* period before the antecedent to match enough to fit the *gluttonous* antecedent specification. In other words, we only need the trajectory period to match enough that we are referring to the antecedent we think we are.[23] To see this, reflect on Niamh's story: in 2020 she stumbles and bashes her knee while entering her time machine. We can assess our guiding counterfactual about the bruise without knowing whether she had a cold, or what was playing on the radio at the time. We can well imagine a

[22] For an independent motivation for adopting approximate match, see Dorr 2016.

[23] These requirements hold for both forwards and backwards counterfactuals, but recall that in the former, trajectory and relevance overlap, so we default to the more stringent requirement (perfect match).

counterfactual scenario in which she enters the time machine without stumbling, and goes on to live a rather different life between 1970 and 2020 as a result. What we require for making sense of the counterfactual is that this different life does not confound the antecedent postulated in which she does not bash her knee. Putting these two requirements together, we can tolerate some part of the world—Niamh's alternate life—being very different from actuality, and a great deal of the world—all regions affected by the ripples of that life, including the period of our counterfactual antecedent itself—being a little different.

Thus, the trajectory period should only require approximate match, not perfect match (which generated the Redux problem). Specifying what counts as approximate enough remains tricky, but one natural guide to what is important in the match is what is contained in a gluttonous reading of the antecedent: whatever that reading says was the case in the trajectory period is implicitly part of the antecedent. The upshot of this is that our *trajectory* requirement is not an additional element in the rubric for assessing the closeness of worlds, but is instead a function of understanding the true scope of the antecedent.

VI.3 Bonus

An important silver lining from this Redux and our alignment with Lewis regarding large miracles emerges. The prohibition on Big Miracles is what protects the openness of the period between Nixon's press and the nuclear disaster—matching actuality across this period despite Nixon's press would require a convergence miracle, rendering the world in question distant and thus irrelevant.

The very same prohibition ensures the openness we need for our Modal Moat too. A world where the bash does not happen in 2020, but which matches actuality (including bruise) in 1970 is a world with a Big Miracle appearing somewhere between the antecedent and consequent. By holding onto Lewis's first rule (Section II), we ensure the openness of the consequent that we have argued is essential to the correct assessment of our target backwards counterfactual. We get the open region of the moat for free.

VII. Journey's end

We have come a long way and so it will help to summarize where we have landed.

When assessing any counterfactual statement of the form 'if it were the case that *c*, it would be the case that *e*', we should first seek to clarify what is implicitly expressed by the antecedent. This is to guard against *antecedent* *gluttony* and to ensure approximate match to the actual world of the antecedent period as required for *trajectory*. When we make explicit what was implicit, we 'de-gluttonize' the counterfactual.

Next, when assessing the truth of any (de-gluttonized) counterfactual, the Lewis semantics hold: the counterfactual is true iff there is no c and ¬e world that is closer than some c and e world.

Importantly, we differ from Lewis with respect to two closeness criteria for worlds when evaluating the counterfactual conditional (deviations in italics):

> (2) It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact obtains, *excluding the regions of the counterfactual relata*.
> (4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly, *except in the region preceding the antecedent* (Lewis 1979).

We maintain Lewis's (1) and (3): the avoidance of big and small miracles. Rule 2 provides match of the history up until, but not including, the earliest relatum. That delivers *relevance*. Rule 1 provides an open period between the relata since a large widespread miracle would be required to ensure match in that period *following* the counter-to-fact region of the earliest relatum. That delivers *openness*. Attending to the implicit content of the antecedent provides approximate match of the period just prior to the antecedent. That delivers *trajectory.*

Together, this delivers the modal moat for backwards counterfactuals: fix the entire period before the earliest relatum (consequent), (approximately) fix the period leading up to the antecedent, and keep as much as you can—minimizing big miracles—open in between.

And now, as with all great time travel stories, we can end where we began: '*If I hadn't banged my knee, I wouldn't have this awful bruise!*' When evaluating Niamh's counterfactual, we first tease out what she meant by the antecedent: *that* situation modulo the bang. We then hold everything fixed in the past of the earliest relatum—in this case, the consequent bruising—and leave open the regions of (no) bang and (no) bruise. It would take a miracle to have *that* scenario, and no bang, but still have a bruise. Therefore, there exists some closer world in which Niamh does not bang her knee, and does not have that awful bruise. We always knew Niamh spoke truly, and now we have a way of evaluating counterfactuals that agree. Distilled into our sought-after heuristic, we say: get clear about what is entailed in the antecedent, and hold fixed the past of the earliest relatum.

References

Bennett, J. (2003) *A Philosophical Guide to Conditionals*. New York: Oxford University Press.
Dorr, C. (2016) 'Against Counterfactual Miracles', *The Philosophical Review*, 125: 241–86.
Fernandes, A. (2021) 'Time Travel and Counterfactual Asymmetry', *Synthese*, 198: 1983–2001.
Fine, K. (1975) 'Review of Counterfactuals', *Mind*, 84: 451–8

Ichikawa, J. (2011) 'Quantifiers, Knowledge, and Counterfactuals', *Philosophy and Phenomenological Research*, 82: 287–313. https://doi.org/10.1111/j.1933-1592.2010.00427.x

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

Lewis, D. (1976) 'The Paradoxes of Time Travel', *American Philosophical Quarterly*, 13: 145–52.

Lewis, D. (1979) 'Counterfactual Dependence and Time's Arrow', *Noûs*, 13: 455–76.

Lewis, K. S. (2016) 'Elusive Counterfactuals', *Noûs*, 50: 286–313. https://doi.org/10.1111/nous.12085

Mackie, P. (2014) 'Counterfactuals and the Fixity of the Past', *Philosophical Studies*, 168: 397–415

Pollock, J. L. (1976) 'The 'Possible Worlds' analysis of Counterfactuals', *Philosophical Studies*, 29: 469–76.

Rennick, S. (2021) 'Trope Analysis and Folk Intuitions', *Synthese*, 199: 5025–43.

Schaffer, J. (2004) 'Counterfactuals, Causal Independence and Conceptual Circularity', *Analysis*, 64: 299–309.

Wasserman, R. (2006) 'The Future Similarity Objection Revisited', *Synthese*, 150: 57–67.

Wasserman, R. (2015) 'Lewis on Backward Causation', *Thought*, 4: 141–50.