philosophies

MDPI

Article

Does Lewis' Theory of Causation Permit Time Travel?

Phil Dowe

School of Philosophy, Australian National University, Canberra, ACT 0200, Australia; phil.dowe@anu.edu.au

Abstract: David Lewis aimed to give an account of causation, and in particular, a semantics for the counterfactuals to which his account appeals, that is compatible with backwards causation and time travel. I will argue that he failed, but not for the reasons that have been offered to date, specifically by Collins, Hall and Paul and by Wasserman. This is significant not the least because Lewis' theory of causation was the most influential theory over the last quarter of the 20th century; and moreover, Lewis' spirited defence of time travel in the 1970s has shaped philosophers' approach to time travel to this day.

Keywords: time travel; counterfactuals; causation; miracles

Citation: Dowe, P. Does Lewis' Theory of Causation Permit Time Travel? *Philosophies* **2021**, *6*, 94. https://doi.org/10.3390/philosophies6040094

Academic Editor: Alasdair Richmond

Received: 14 September 2021
Accepted: 17 November 2021
Published: 23 November 2021

1. Introduction: Lewis' Theory, as Advertised

In 'Causation' [1], David Lewis presents the theory that *c* causes *e* if *c* and *e* occur and 'had *c* not occurred, e would not have occurred' is true (='counterfactual dependence'), or if there is a chain of such counterfactual dependence connecting *c* and *e*. The counterfactual 'had *c* not occurred, *e* would not have occurred' is true iff *e* does not occur in any of the closest ~*c* worlds. This analysis is intended to work under determinism. In that 1973 paper, Lewis says " [We should not reject] a priori certain legitimate physical hypotheses that posit backward or simultaneous causation" [1] (p. 566). I will argue that despite Lewis' best intentions, he has unwittingly done so. I will also show that the argument applies equally to Lewis' final version [2] of the counterfactual theory.

In 'The Paradoxes of Time Travel' [3], Lewis defends the possibility of time travel on the basis of an eternalist metaphysics of time and a purdurantist–causal theory of persistence. Since two stages of a time traveler need to be connected by appropriate causal connections in order for the two stages to be part of the one person, and hence to be a time *traveler*, "... travel into the past necessarily involves reversed causation" [3] (p. 147). In this paper, Lewis says, 'Elsewhere I have given an analysis of causation in terms of chains of counterfactual dependence, and I took care that my analysis would not rule out casual reversal a priori" [3] (p. 148). I will argue that despite significant effort on Lewis' part, it turns out that he did not take enough care.

In 'Counterfactual Dependence and Time's Arrow' [4], Lewis spells out a more detailed account of comparative overall similarity. His celebrated 'similarity measure' orders worlds as follows:

(A) It is of the first importance to avoid big, widespread, diverse violations of law.
(B) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(C) It is of the third importance to avoid even small, localized, simple violations of law.
(D) It is of little or no importance to secure approximate similarity of particular fact, even in matters which concern us greatly [4] (p. 472).

Lewis claims that a major advantage of this analysis over alternatives (e.g., [5]) that hold the past fixed by 'brute force' is that his account allows for backwards causation. Lewis adds "Careful readers have thought they could make sense of stories of time travel... speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves. Most or all of these phenomena

would involve special exceptions to the normal asymmetry of counterfactual dependence. It will not do to declare them impossible a priori" [4] (p. 464). I will argue that Lewis, in effect, has done exactly that.

For ordinary forwards causation, according to Lewis' similarity relation, the usual structure of the relevant ~*c* worlds where event *c* is an actual cause of event *e* will be: perfect match with actuality up until a time t_{c-} just before *c*; a small miracle leading to ~*c*; no further miracles; and future divergence including ~*e*. Specifically, those worlds will be closer than worlds like this: perfect match after t_{c+} including *e*, numerous diverse miracles leading back to ~*c*, no other miracles, and past divergence. Thus, to have backwards causation of an earlier event *e* by *c*, the idea is that we should look at worlds like this: perfect match after t_{c+}, a small miracle leading back to ~*c*, no other miracles, and past divergence including ~*e*. Lewis says:

> "I think I can argue (but not here) that under my analysis the direction of counterfactual dependence and causation is governed by the direction of other de facto asymmetries of time. If so, then reversed causation and time travel are not excluded altogether, but can occur only where there are local exceptions to these asymmetries" [3] (p. 148).

Lewis claims that on his analysis, the direction of counterfactual dependence and causation is governed by what he calls the 'asymmetry of overdetermination'. A 'determinant' is a minimal set of conditions jointly sufficient, given the laws of nature, for the event in question. The designation 'asymmetry of overdetermination' is ambiguous. On the one hand, it refers to a global asymmetry, and on the other hand it refers to an extrinsic property of a particular event, the putative cause. I will use the term exclusively in the latter sense. The property concerns how the event connects lawfully to its surroundings. An event is overdetermined if it has more than one determinant. A particular event, say *c*, displays an asymmetry of overdetermination if the overdetermination in one direction in time is significantly greater than its overdetermination in the other direction in time. For forwards causation, the cause *c* must exhibit a significantly greater overdetermination in the future than in the past, and vice versa for backwards causation [4] (p. 474).

On Lewis' account, causation supervenes on the actual distribution of particulars, despite the inter alia reference to possible worlds. Thus, a non-standard way to think about Lewis' counterfactual theory of causation is to note some of its distinctive entailments about actuality: (1) A cause is a nomically necessary condition, given the actual circumstances, for its effect, where 'nomically' means according to the laws, which for Lewis are given by the best system analysis of the actual distribution of particulars; and (2) a cause exhibits asymmetry of overdetermination in the direction of the effect, that is, toward the past or toward the future depending on whether the effect is respectively in the past or in the future of the cause. Both conditions are necessary for causation. The reason (2) is necessary for causation is that, if a cause had no such asymmetry, that is it had the same number of determinants in the future as in the past, then a world with a perfect match in the past up to a small miracle leading to ~*c*, would be no closer than a world with a perfect match in the future back to a small miracle leading to back to ~*c*. In the former world, there would be no event *e*, in the latter there would be, and since *e* must be missing in all the closest ~*c* worlds for *c* to cause *e*, it follows that an event which lacks asymmetry of overdetermination cannot be the cause of anything. Actually (2) cannot be quite right, as I shall later illustrate, because it is defined in terms of the distinction between the past and the future of the event in question, whereas the similarity relation does no such thing, appealing only to 'spatiotemporal regions'.

As I have indicated, contrary to advertisement, Lewis' account fails to allow for backwards causation. I will sandwich my main argument for this (Section 3) between discussions of some specious arguments for the same conclusion.

2. Specious Argument #1

Suppose c is the cause of an earlier event e, that c has earlier causes including b, and also later effects including d. For e to be an effect of c, c must exhibit an asymmetry of overdetermination towards the past. Suppose the following world is closer than any $\sim c$ world containing e: perfect match after t_{c+}, small miracle leading back to $\sim c$, no other miracles, and past divergence including $\sim e$. However, that world contains d, so it follows that c cannot also be a cause of d; but that is to be expected, because for d to be an effect of c, c must exhibit an asymmetry of overdetermination towards the future. Since it is logically impossible for an event to exhibit an asymmetry of overdetermination in both the past and future directions, no event can have both a past and a future effect. Yet typical time travel scenarios, and the 'speculative physicist's' hypotheses mentioned by Lewis, do involve events with past and future effects. Therefore, Lewis does not allow for typical time travel or backwards causation.

This is a specious argument. The similarity relation does not trade on the distinction between past and future—it is formulated in terms of spatiotemporal regions. Space-time divides, for example, into three regions and the similarity relation rules on the relative closeness of a world where part of the past matches actuality, and part of the past diverges (Compare [6] p. 10). Such a world—with reference to the above case—might be: Perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future including $\sim d$, and divergence in the past in a region stretching from $\sim c$ back through a region including $\sim e$. Then indeed, c causes both e and d. The condition for causation implicit in the similarity relation is not 'a cause exhibits asymmetry of overdetermination in the direction (past or future) of the effect', but rather, (2)* a cause exhibits asymmetry of overdetermination in the spatiotemporal region containing the effect compared to some other region very close to c. To reject this point is to accept the specious argument, I would claim, and among other things that would leave one wondering how Lewis ever thought his account could get off the ground.

This brings us to an exception that should be allowed to the argument I will put below. I will discuss the exception first, then give the argument in the next section. Suppose again that c is the cause of an earlier event e, that c has earlier causes, including b, as well as later effects. Call the region containing all of c's past effects the 'time travel region of c', and call the region containing all of c's past causes the 'causal past of c'. Now we make two assumptions. (1) Suppose the time travel region of c and the causal past of c do not overlap, or more precisely, there exists regions R, S, s.t. R contains all of c's past effects, S contains all of c's past causes, and R and S do not overlap. (2) Suppose, more generally, nothing in the time travel region of c causes anything in the causal past of c. For any case where these assumptions hold, the closest world will indeed be: perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future, and divergence in the time travel region of c including $\sim e$. Thus, c causes e, and Lewis' analysis does indeed permit time travel. The assumptions are very restrictive, perhaps too restrictive to be of interest, but it does give us an exception to the argument I am now going to give.

3. The Main Argument

Suppose I want to see a comet pass near the earth, but I sleep through the alarm[1]. Suppose someone (my older self) subsequently gives me instructions for a time machine (event b). I find the time machine, set the controls as instructed, and press the button (c). The machine immediately travels back in time; nevertheless, c has future effects, including the sound of the button being pushed reaching (d) the ears of a nearby possum. I then get to see the comet (e), and then I find my younger self and give him the instructions (b). c causes e, a case of backwards causation (Figure 1). Suppose the scenario just described plays out in the actual world, $W_@$. Let us make two assumptions, both purely heuristic: (1) nothing earlier than e has a future cause, and (2) no future event (including c) has any effects in the region of b except via the region of e. It does not matter for the following argument

whether the time travel occurs via closed timelike curves, which does not involve any local backwards causation, or via a 'Wellesian' time machine, which does.

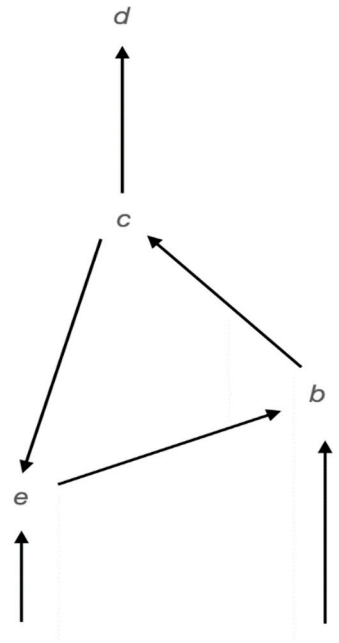

Figure 1. If I had not pressed that button?

I will start by asserting which worlds are among the closest worlds[2], then I will show that if those are the closest worlds, then *c* does not cause *e*, and then I will show that those worlds are indeed (in general) the closest worlds. The closest ~*c* worlds are worlds like W_1 and W_2, as follows:

W_1: Perfect match up to $t_{e-,}$ thereafter divergence, no miracles, ~*e*.

W_2: Perfect match up to $t_{e-,}$ thereafter divergence, no miracles, *e*.

(I say, 'Perfect match up to $t_{e-,}$', but this perfect match may extend a little further into the future of *e* in the region of the immediate past of *b*, depending on what one means by 'future'.) An example of W_1 would be: I sleep through the alarm, no-one appears out of a time machine, no-one gives me the instructions, I do not even know about the time machine, I do not time travel, and I do not get to see the comet. I simply regret sleeping through. This is a ~*c*, ~*e* world. An example of W_2 would be: I sleep through the comet, then someone (her older self) subsequently gives my wife instructions for a time machine. She takes me to the time machine, sets the controls as instructed, and she presses the button. We both time travel, I see the comet, and she passes the instructions to her younger self. This is a ~*c*, *e* world; that is, it is not true that I press the button, but it is true that I see the comet. In neither of these worlds are there any miracles. The past of *e* and *b* matches actuality, but what ensues depends on not only on that past, but also on what happens in the future, including whether or not the button is pressed. Vary the latter, and you vary what happens in the region of *e* and *b*, without miracles. These are the closest worlds, and if I am right about that then *c* does not count as a cause of *e* on Lewis' theory, since his theory requires that *e* does not occur in any of the closest ~*c* worlds. W_2: is one of the closest ~*c* worlds, *e* occurs in W_2, hence *c* does not cause *e*.

In fact, in general, there would be indefinitely many such closest worlds, each containing no miracles. This feature is well-known in the literature on wormhole time machines in the context of general relativity [7–9], but is not so well-appreciated in the philosophical literature concerning Wellesian time travel. In the physics literature, it has been shown for certain classes of cases that data fixed before a time travel region can be extended to indefinitely many consistent trajectories through the time travel region [7]. Thus, the general theory of relativity is, in one sense, an indeterministic theory (although not in another sense,

since all the local dynamics will be deterministic). What happens in the region of *e* depends deterministically on what happens before *e together with* what happens in the region of *c*. On the other hand, there is indeterminism in the sense that what goes on before the time travel region does not fix what occurs in the time travel region [8]. Laplacean determinism (i.e., that the particulars at a given time together with the laws fixes the particulars at later times) fails. This, however, does not enable us to define chances in the usual way, so it is appropriate that we seek to use Lewis' theory designed for determinism. In any case, Lewis' theory designed for indeterminism, where causation obtains on account of counterfactual chances, was not developed far enough to see how Lewis might account for backwards causation (see Lewis' comments at [10] (p. 274), for a discussion of the constraints on chances in loops see [11]).

Assuming still that I am right about the closest worlds, why does Lewis' theory go wrong? It is not because there is anything wrong with the idea that causes are necessary conditions for their effects in the actual circumstances. The problem in time travel cases is that Lewis' similarity relation fails to do something specific that it is intended to do: it fails to pick out the worlds in which the actual relevant background conditions are held fixed. If you allow the background conditions to vary, then in general, a cause will not be a necessary condition for its effect, even if it is a necessary condition for its effects in the actual circumstances. The reason it fails to pick out the worlds in which the actual relevant background conditions are held fixed is that large miracles are required to hold fixed the background of the cause on account of the fact that the cause has effects in that region. The asymmetry of overdetermination makes it impossible to avoid that, short of conditions that separate the regions containing the causal past and regions containing the backwards effects. That is why some of our closest $\sim c$ worlds contain *e*.

So let us turn to the question of whether worlds like W_1 and W_2 are indeed (in general) the closest worlds according to Lewis' similarity relation. Worlds like W_1 and W_2 contain no miracles. A world with more perfect match gained at the cost of a large miracle will not be as close as W_1 and W_2. The only way to get a closer $\sim c$ world is to buy increased perfect match at the cost of a small miracle or two. Let us try that. What we would like to do in our example is to have perfect match in the region of *b*, and a small miracle leading to $\sim c$. Indeed, such a world can have divergence in the future and in the region of *e*, such that we get the required result that *e* and *d* do not occur at that world. The problem is, in such a world we need a large miracle to go from divergence in the region of *e* to perfect match in the region of *b*, if Lewis' similarity relation is right about the asymmetry of overdetermination. Each and every change in the *e* region which would have effects in the *b* region each of which would need to be somehow deleted by a miracle in order to maintain perfect match in the *b* region, and that set of diverse miracles would add up to a large miracle. Alternatively, a closer world (W_3 below) than the one just described is one where the reconvergence miracle occurs between *c* and *e*, not in the region between *e* and *b*, and this would be a world containing *e*. Both of these worlds contain a large miracle, and hence, neither is as close as W_1 or W_2.

> W_3: Perfect match up to t_{e^-}, perfect match in the region of *b*, small miracle leading to ~c, future divergence, large reconvergence miracle, *e*.

It may be objected that I have only shown that there is no counterfactual dependence of *e* on *c*, not that there is no causation given that on Lewis' theory causation is the ancestral of counterfactual dependence. However, it is easy to show that there can be no chain of counterfactual dependence either. The first link of any chain of counterfactual dependence between *c* and *e* must be a counterfactual dependence of something, call it *f*, on *c*. For *f* to counterfactually depend on *c* it must be that none of the closest $\sim c$ worlds contain *f*. These $\sim c$ worlds are exactly the ones we have already been considering. The closest are worlds like W_1 and W_2. By the reasoning above, some of these worlds contain *f* and some do not. Hence, there is no counterfactual dependence of *f* on *c*, and hence, there can be no chain of counterfactual dependence between *c* and *e*, hence *c* does not cause *e*.

Alternatively, it may be objected that I have only shown that there is no backwards causation on Lewis' original theory, but not for his final theory [2]. On this theory *influence* is "a pattern of dependence of how, when and whether upon how, when and whether" [2] (p. 190), and causation is the ancestral of influence. This allows for more cases to count as causation compared to the original theory, namely, all those cases which do not exhibit whether on whether dependence yet do exhibit how, when or whether on how, when or whether dependence. In fact, influence can obtain when any one of nine kinds of dependence obtains, whether on whether being one of those. So to show that there is no whether on whether dependence (counterfactual dependence) in a case of putative backwards causation, as I have done, is not to show that there is no influence. However, it is straightforward to show that the argument I have given applies equally to any case of influence. On this theory, causation obtains if at least one of the nine types of dependence obtains. But for any type of dependence to obtain, the defined change ('alteration') must feature in *all* the relevant closest worlds, just as for counterfactual dependence. Lewis' semantics for such counterfactuals remains the same. Thus, my argument will apply to each kind of dependence. Take when on when dependence. Counterfactually, suppose by a small miracle the button is pushed slightly later (c^*), and that the time travel set-up is such that I see the comet slightly later (e^*). However, by the argument given above, closest c^* worlds are worlds with no miracles and no perfect match in the time travel region, and therefore some of those worlds do not contain e^*. The same argument applies to any alteration to the time of the cause. Hence, there is no when on when dependence. A similar case can be made for each of the nine patterns of dependence, hence c does not influence e. Hence, Lewis' influence theory does not allow for backwards causation.

Return now to the two assumptions of the argument, namely, (1) nothing earlier than e has a future cause, and (2) no future event (including c) has any effects in the region of b except via the region of e. Would it make any difference if we relaxed these assumptions? No. Take assumption (1). This would hold for certain events around a time machine wormhole: events near enough the so-called Cauchy horizon [7]. In fact the Cauchy horizon could be characterized roughly as the set of points which divides the region which can contain events with future causes (time travel region) from the region which cannot (Cauchy region). However, in the Wheeler-Feynman advanced/retarded potential theory assumption (1) would not hold as backwards effects would extend indefinitely into the past, with intensity dropping off with $1/r^2$. Suppose in our example e does have past effects (i.e., assumption (1) is false). If these effects extend indefinitely into the past, then the closest $\sim c$ worlds will be worlds with no miracles but no perfect match—these worlds are still closer than any world with a large miracle. Suppose on the other hand that although e does have past effects so assumption (1) is false, nevertheless there is some earlier event a which is an effect of e but which itself has no earlier effects and indeed that assumption (1) holds of a. Then the closest $\sim c$ worlds would be worlds with perfect match up to a time just before a, and no miracles. Thus the argument still holds if we drop assumption (1).

What about assumption (2)? Suppose c for example has direct effects throughout the region of b, that is, effects which are not 'mediated' by events in the region of e. For simplicity retain assumption (1). Then again, the closest $\sim c$ worlds will be worlds like W_1 and W_2. Now, to get perfect match in the region of b we not only need a large miracle to remove the backward effects in the region of e due to changing c, we also need a large miracle to remove the direct backward effects in the region of b due to changing c. No reason here to doubt my argument.

Assumption (2) does not hold in standard examples of time travel. Lewis' example of Tim who tried to kill his grandfather involves a certain spatiotemporal region replete with both the effects and the causes of future events. Tim's aiming his gun at Grandfather is the effect of Tim's later decision to travel back to kill him. Tim's Grandfather surviving is the cause of Tim's later decision to travel back to kill him. But again, in order to hold fixed the background condition of Tim's pressing the time travel button, we need large miracles to erase the effects of his counterfactually not doing so.

Finally, consider again the exception I granted at the start of this section. In this case, we made two assumptions: (1) the time travel region of c (the region containing all of c's past effects) and the causal past of c (the region containing all of c's past causes) do not overlap, and (2) nothing in the time travel region of c causes anything in the causal past of c. I said the closest worlds will be: perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future, and divergence in the time travel region of c including $\sim e$. This entails that there are no causal loops. So, should the constraint on my argument simply be that there are no causal loops? No, suppose there are 'epiphenomena' of c in the regions of e and b, where epiphenomena are events with no effects and the epiphenomena of c are the effects of c in with no effects. Suppose—as in the exception I granted—that there are no causal loops. Then again, to get perfect match in the region of b in a $\sim c$ world, we need a large miracle to remove the epiphenomena of changing c. Thus the exception cannot simply be granted to worlds with no causal loops.

We now recap the argument. Time travel hypotheses and other backwards causation hypotheses generally require regions containing both effects and causes of some pertinent event, say c. Under such circumstances Lewis' semantics rule out c being the cause of anything in that region (except for causes, if there are any, that are necessary for their effect in all physically possible circumstances). Hence, there can be no time travel on Lewis' theory except where there is no such spatiotemporal and causal overlap.

4. Specious Argument #2: Collins, Hall and Paul

Collins, Hall and Paul [12] (pp. 9–11) give a different argument to the conclusion that Lewis' semantics does not allow backwards causation. In their example a single billiard ball approaches a Time Machine Portal, entrance on the right, exit on the left. It is on course to pass harmlessly between the entrance and exit but a ball emerges from the exit at $t = 0$, collides with the first ball and sends it into the entrance at $t = 1$. Surprise, it is one and the same ball. Suppose:

c—the ball rolling into the entrance
e—the ball emerging from the exit

"Clearly, c causes e", they say [12] (p. 11). So which are the closest $\sim c$ worlds? Consider these $\sim c$ worlds:

W_1: Small miracle leading to $\sim c$ (ball vanishes before rolling into the entrance); a small miracle to re-instate the ball leaving the exit; e; perfect match except for the small region where the ball has disappeared.
W_2: Perfect match in the past, the ball goes straight through with no collision; no ball through the time machine; $\sim e$.

W_1 closer than W_2 because it has the greater region of perfect match which trumps the fact it has more (two) small miracles. Therefore c does not cause e.

We must reject this analysis. The problem is that it trades on an example which is 'simple and ideal'. The example contains just one particle which could be, just as well for the example, a Newtonian point particle. But it can be shown that Lewis' semantics does not apply to simple ideal worlds (compare [13]); causation, on Lewis account, only applies in systems with sufficient complexity. This point will be established in Section 5.

To be convinced at least that the example trades on being 'simple and ideal', one just has to add 'sufficient complexity'. Suppose we have a billiard ball, and suppose there is a strong light source so that there is plenty of light being continually reflected off the ball. Suppose in addition there is a rough surface which results in a noise when the ball rolls over it. Suppose for simplicity we have a short-lived wormhole, so that there is a Cauchy horizon at t_{cauch}. Light will of course enter the time machine, and thereby be propagated back near but not beyond t_{cauch}. Suppose again a ball is on course to pass harmlessly between the entrance and exit but a ball emerges from the exit at $t = 0$, collides and sends the ball into the entrance at $t = 1$. Again, it is one and the same ball. This time, however, if we remove c, it would take a large miracle to regain perfect match anywhere in the

time travel region. Say we have again a small miracle to delete the ball as it rolls into the entrance, and another small miracle to reinstate it leaving the exit. The ball is missing through the wormhole, and any light that would have been reflected off the stage now missing will also be absent, as will any sound waves that the missing stage of the ball would have initiated. We could for example begin to recover perfect match everywhere by a large number of miracles which 'reflect' the light as if it were reflecting off the ball (where it is missing). This might occur by numerous small miracles producing photons with the right properties one by one; and in the same way removing the light that actually travels straight past (where the ball would have been); together with miracles to generate the noise the ball would have made. This array of small miracles adds up to a large miracle. So compare:

W_1: A small miracle leading to $\sim c$ (ball disappears), a small miracle to re-instate the ball at exit, e; a large miracle to ensure perfect match.
W_2: Perfect match in the past until t_{cauch}; ball goes straight through, no collision; no ball through the time machine, $\sim e$.

W_2 is closer than W_1 since the latter contains a large miracle.

However there are other $\sim c$ worlds which do contain e, and which are equally close to W_2. For example, there are closest $\sim c$ worlds containing 'lions'—loops objects which have no beginning or end[3], and in particular have no presence before t_{cauch}. In our case the 'lion' could be a second billiard ball, but let us stick with the literal lion. Consider this $\sim c$ world:

W_3: Perfect match in the past until t_{cauch}; no miracles; lion picks up the ball and takes it through the time machine, releases it so it comes out as before, e.

W_2, W_3 are equally close worlds; in fact they are among the closest worlds and both are closer than W_1. Thus c does not cause e. However, now we see the argument conforms to the argument of Section 3 and does not rely on the fact that components of the set up are simple and ideal. Thus, the analysis of Collins et al. is the wrong analysis: Lewis fails to account for backwards causation in their example because their example does not have sufficient complexity for there to be any causation, backwards or otherwise. We will now defend that premise.

5. Lewis' Emergent Causation

"Overdetermination in Lewis' sense . . . "fades out" as we approach the micro level", says Huw Price [13] (p. 150). Suppose the actual world contains just two ideally elastic inert point particles, which collide just once. Suppose c is a certain instantaneous state of the first ball at a point before the collision, and e is a certain instantaneous state of the second ball at a point after the collision. We want to say c causes e. What are the closest $\sim c$ worlds? Compare:

W_1: Perfect match up until t_{c-}; small miracle leading to $\sim c$; no further miracles; future divergence; $\sim e$.
W_2: Perfect match after t_{c+}; small miracle leading back to $\sim c$; no other miracles, past divergence; e.

W_1 and W_2 are to be taken as equally close, hence c does not cause e. This arises because there is no asymmetry of overdetermination at the level of the simple and ideal, at least when you have time-symmetric dynamics. The asymmetry, as pointed out above, is necessary for causation. However, there are worlds closer than either W_1 or W_2. Compare:

W_3: Perfect match up until t_{c-}; small miracle leading to $\sim c$; small miracle leading to the production of an identical particle *ex nihilo* at t_{c+}; no further miracles; perfect match after t_{c+}; e.

W_3 is closer than W_1. It is worth a second small miracle to purchase a large swath of perfect match, according to Lewis' similarity measure. Hence, c does not cause e.

So there are two reasons there can be no causation on Lewis' theory at the level of the simple and ideal. The first is that following the counterfactual absence of an actual

event a world can reconverge to actuality without a large miracle. The second is that an event can appear without causal precedence and without a large miracle, thereby buying additional perfect match. Once we have sufficient complexity and (presumably, although the connection is quite controversial) an entropy gradient like ours, then it will take a large miracle to reconverge in the first case, and in the second case although it still just takes a small miracle for miraculous replication, now that will not buy any additional perfect match, and so will be ruled out on account of having an extra miracle, on Lewis' similarity measure. In addition, when we consider large scale events or objects, there must be some point at which a miraculous replication requires a large miracle. The miraculous appearance of anything with a considerable number of diverse parts will surely count as a large miracle. For Lewis' theory of causation, then, causation is emergent: it arises only where there is sufficient complexity for events to exhibit the asymmetry of overdetermination.

6. Specious Argument #3: Wasserman

Wasserman [16] gives an argument somewhat similar to that of Collins, Hall and Paul in that it also trades on the simple and ideal, but different in that it is based on the phenomenon of action at a temporal distance. In his example a single electron persists from t_1 to t_3, at which point it enters a time machine. At that moment, a button is pressed and the electron is sent back, discontinuously, to t_2, where it continues on to t_4 and on into the future. Assume that no other electron appears at t_2, ie there are no preempted backups. [16] (pp. 143–144). Wasserman claims this counterfactual is true: 'If the button hadn't been pressed at t_3, an electron wouldn't have appeared at t_2'.

A world W_1 where a small miracle leads to the appearance of an electron at t_2—call this miraculous replication—and a second small miracle leads to the button not being pressed is closer than a world W_2 with no appearance of an electron at t_2 and a single small miracle leading to the button not being pressed. The latter has perfect match only up until t_2 while the former has perfect match up until t_3, *viz.*:

> W_1: perfect match until t_3; a small miracle; appearance of an electron at t_2; second small miracle at t_3; the button is not pressed.
> W_2: perfect match until t_2; no electron at t_2; a small miracle at t_3; the button not pressed.

Therefore it is false that if someone hadn't pressed the button at t_3, an electron wouldn't have appeared at t_2. And so, on Lewis' account, pressing the button is not the cause of the electron landing back at t_2. Further, the culprit can be identified, Wasserman claims:

> Note that this line of reasoning does not apply in the case of ordinary future-directed counterfactuals . . . In the case of time travel, the traces of the button-pressing are irrelevant, since those traces do not disrupt perfect match in the past. This is the fundamental problem for Lewis—in the case [of] backward counterfactuals, perfect match can be purchased without a miraculous cover up. That is why we get the wrong results . . . [16] (p. 147).

A straightforward counterargument shows that it is not only in cases of backwards counterfactuals that we meet this problem. Assume a Newtonian spacetime which allows action at a spatial distance. Suppose when the button is pressed at p_3 (now using p for a spacetime point) an electron is teleported instantaneously to some far away location p_i. As with Wasserman's example there are closer no-button-pressing worlds than the world with perfect match in the past until by a small miracle the electron disappears just before p_3, thence no perfect match or miracles, and no appearance of an electron at p_i. One such closer world would be a world with perfect match in the past until by a small miracle the electron disappears just before p_3, but then by a small miracle an electron appears at p_i. As with Wasserman's example, this world exhibits a greater region of perfect match at the cost of a second small miracle, assuming only that any other effects of the button pressing do not reach that region until some time later. However, in response Wasserman can easily adjust the diagnosis to encompass simultaneous causation (or for that matter causation

outside the lightcone in a Minkowski spacetime). Again, perfect match can be purchased without a miraculous cover up.

The reasoning can be extended to a limited set of forwards in time cases. Suppose in our Newtonian spacetime the forward lightcone limits causation except for rare action at a distance. Then if the button pressing sends the electron to some future location outside the forward light cone of the button pressing, we again find miraculous replication in the closest worlds rules against causation. Lewis' account fails to account for (a limited kind of) forwards in time causation.

The correct diagnosis is that Wasserman's example trades in the simple and ideal, just like that of Collins et al.. As I have argued, on Lewis's account causation is emergent at higher levels of complexity. The reason "perfect match can be purchased without a miraculous cover up" in the appearance of an electron ex nihilo derives from the 'simple and ideal' nature of the micro events and so that we shouldn't expect there to be causation on Lewis' theory, quite apart from any backwards causation. According to Wasserman there is an implausible asymmetry entailed by Lewis' account [16] (p. 147). If we supposed a person rather than an electron was sent back in time then we wouldn't have the problem he gives, because it would take a large miracle to bring forth a person *ex nihilo* at t_2. Such a world would not be a closest world, and hence pressing the button comes out as the cause of the person landing back at t_2. I agree it would take a large miracle to bring forth a macro object *ex nihilo*. I also agree there is an asymmetry between the two cases, but not because one is a case of causation and the other is not; but because there are different reasons in each case for why there is no causation.

It is not the case on Lewis' theory that pressing the button comes out as the cause of a person landing back at t_2. The reason is not that a small miracle could be responsible for miraculous replication of a person, but that it could happen without any miracle. Suppose actually I pressed the button and showed up discontinuously at an earlier time t_2. Consider the worlds where I do not press the button. One such world is where I do not show up at t_2. Another is where I do not press the button, I go home, someone appears from the future, convinces me to go back to the machine, presses the button for me, and then I show up discontinuously at the earlier time t_2. These two worlds are equally close: perfect match until t_2; one small miracle. One has the putative effect, the other does not; so on Lewis' theory, my pressing the button is not the cause of my turning up at the earlier time t_2. This is essentially the same problem as the one I outlined in Section 3. Anything can happen in the time travel region.

But there is another feature of Wasserman's example in addition to the time travel and the simple and ideal aspect, that might be considered to raise a red flag: spooky action at a distance. Should we identify the problem in this example with the spooky action at a spatiotemporal distance, rather than with the time travel or the simple and ideal aspect? Action at a temporal distance is problematic in its own right [17], although Lewis wants to allow for the possibility [3] (p. 148). If we disallow action at a distance then there is no putative counterexample here to backwards causation. Wasserman thinks otherwise: "I focus on discontinuous time travel for the sake of simplicity. The same point can be made in the case of continuous time travel" [2] (n. 11, p. 149). This appears to be an error. Suppose on pressing the button at t_3 the electron turns around in time and travels back continuously to t_2 where it turns around and travels forwards in time[4]. We can consider the world where a small miracle leads to the appearance of an electron at t_2 and a second small miracle leads to the button not being pressed, but this does not generate any further perfect match, since there is no electron traveling backwards between t_3 and t_2 whereas in the actual world there is. On account of its second miracle, that world is not as close as any world with no appearance of an electron at t_2 and a single small miracle leading to the button not being pressed. However, again there are other equally close worlds where an electron does appear without a miracle—some other use of the time machine could be responsible provided only that it does not count as the event pressing the button at t_3. Hence the continuous example does not count as causation, but it is not "the same point".

So could we argue that the problem here really lies in the action at a distance? No, we are going to find that when action at a distance is ruled out on Lewis' theory, the reasons it is ruled out are exactly those reasons we've already seen: either the problems arise because we are dealing with simple and ideal cases, or because we find we cannot hold the right things fixed in the action at a distance region. And those problems can arise without action at a distance, cf Sections 3 and 4. We have already seen how this works for simple and ideal cases of simultaneous causation and for a limited set of cases of future action at a temporal distance: it works because one can achieve a replication miracle to reproduce the effect by a small miracle and thereby generate further perfect match. In the case of a large complex effect like the appearance of a person, we have seen that we do not get causation because it is not possible to hold fixed the events in the background of the cause, and in fact nothing in the time travel region can be held fixed by the similarity measure.

7. Conclusions

Contrary to his explicit advertisment, Lewis' semantics does not allow for backwards causation or time travel. Essentially, the reason is that when a cause c has effects in its own causal past, the closest $\sim c$ worlds will be worlds which do not hold fixed the background of c, hence in general, an earlier effect of c will not be absent from all the closest $\sim c$ worlds; $\sim c$ worlds which do hold fixed the background of c contain large miracles, and hence are not the closest worlds. The exception to this argument is the case where the time travel region and the causal past of a cause do not overlap, and where nothing in the time travel region causes anything in the causal past. The constraints necessary for this exception are difficult to instantiate. They do not in general allow for the kinds of physical hypotheses Lewis hoped to allow for, namely tachyons, advanced potentials, and cosmological models with closed timelike curves. These constraints may allow backwards causation in, for example, the case of a very specific trajectory of a tachyon, although they do not in general allow tachyon trajectories to count as backwards causation. Finally, these constraints certainly do not make possible the kinds of time travel Lewis has in mind—stories such as Tim attempting to kill his grandfather. This is significant not the least on account of the level of influence that Lewis' theory of causation and his defence of time travel have had on philosophy over the last 50 years.

Funding: This research was supported by the Australian Reasearch Councile grant #DP110101815.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 This argument was first presented at the American Philosophical Association Central Division meeting in Chicago, 18–21 February 2009.

2 Strictly speaking Lewis' similarity measure is a three place relation between two sets of worlds A, B say and some specific world, call it actual. It might give the result that each and every world in set A is closer to the actual world than is any world in set B. However, in rough and ready parlance, it is common practice to speak of the closest worlds.

3 The term 'lion' arose from consideration of Lewis-type time travel consistency claims: Tim attempts to kill his grandfather before his parents' conception, but fails for 'commonplace reasons', the gun jams, he loses his nerve etc. Do these reasons have their source in the region prior to the Cauchy horizon? Not necessarily: "an unexpected hungry lion behind the door of a time machine could effectively reconcile the traveler's freedom of will and his grandfather's safety" [14] (p. 064013–0640134); [15] (p. 183).

4 This is somewhat like the theory of Feynman [18] that positrons could be interpreted as electrons traveling backwards in time except that the turnaround at t_2 would require energy input which needs to be absent from our example.

References

1. Lewis, D. Causation. *J. Philos.* **1973**, *70*, 556–567. [CrossRef]
2. Lewis, D. Causation as Influence. *J. Philos.* **2000**, *97*, 182–198. [CrossRef]
3. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
4. Lewis, D. Counterfactual Dependence and Time's Arrow. *Noûs* **1979**, *13*, 455–476. [CrossRef]
5. Jackson, F. A Causal Theory of Counterfactuals. *Australas. J. Philos.* **1977**, *55*, 3–21. [CrossRef]
6. Fernandes, A. Time Travel and Counterfactual Asymmetry. *Synthese* **2019**, *198*, 1983–2001. [CrossRef]

7. Friedman, J.; Morris, M.; Novikov, I.; Echeverria, F.; Klinkhammer, G.; Thorne, K.; Yurtsever, U. Cauchy Problem in Spacetimes with Closed Timelike Curves. *Phys. Rev. D* **1990**, *42*, 1915–1930. [CrossRef] [PubMed]
8. Arntzenius, F.; Maudlin, T. Time Travel and Modern Physics. In *Time, Reality and Experience*; Callender, C., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 169–200.
9. Dowe, P. Constraints on Data in Worlds with Closed time-like Curves. *Philos. Sci.* **2007**, *74*, 724–735. [CrossRef]
10. Lewis, D. A Subjectivist's Guide to Objective Chance. In *Studies in Inductive Logic and Probability, Volume II*; Jeffrey, R., Ed.; University of California Press: Berkeley, CA, USA, 1980; pp. 263–293.
11. Dowe, P. Causal Loops and the Independence of Causal Facts. *Philos. Sci.* **2001**, *68*, S89–S97. [CrossRef]
12. Collins, J.; Hall, N.; Paul, L. Counterfactuals and Causation: History, Problems and Prospects. In *Causation and Counterfactuals*; Collins, J., Hall, N., Paul, L., Eds.; MIT Press: Cambridge, MA, USA, 2004; pp. 1–57.
13. Price, H. *Time's Arrow and Archimedes' Point*; Oxford University Press: New York, NY, USA, 1996.
14. Krasnikov, S. Time Travel Paradox. *Phys. Rev. D* **2002**, *65*, 064013. [CrossRef]
15. Krasnikov, S. *Back-in-Time and Faster-than-Light Travel in General Relativity*; Springer: Cham, Switzerland, 2018.
16. Wasserman, R. Lewis on Backward Causation. *Thought* **2015**, *4*, 141–150. [CrossRef]
17. Adlam, E. Spooky Action at a Temporal Distance. *Entropy* **2018**, *20*, 41. [CrossRef] [PubMed]
18. Feynman, R. The Theory of Positrons. *Phys. Rev.* **1949**, *76*, 749–759. [CrossRef]