On what causes what

“Barring a fundamental change in approach, the prospects of a relatively simple, elegant and intuitively attractive, unified theory of causation, whether ontological reduction or conceptual analysis, are dim...” - Laurie Paul and Ned Hall

“I think that the prospects for a conceptual analysis of causation are dim... Conceptual analyses seem to fail generally... Causation itself is conceptually central [and] well-explored... [And] intuitions about causation prove pliable... A conceptual analysis of causation is not on the cards...” - Jonathan Schaffer

“Causation is not to be analyzed in terms of counterfactual dependency at all, no matter how many equants and epicycles are appended to the original rough draft...” - Tim Maudlin

1. Counterfactual analyses of causation are impossible; or so says the consensus in contemporary metaphysics. (As the quotations above make clear.) This paper disagrees: it offers a counterpossible counterfactual analysis of causation, and shows how counterpossibles let us slip cleanly around the counterexamples which’ve posed problems for counterfactual analyses past.

The plan: §2 sketches the relevant technical apparatus. (The non-trivial-counterpossible-tolerating counterfactual semantics. A new property. And something else.) §3 sketches an analysis in that technical apparatus’ terms, and shows how it lets us say the right things about a number of famous counterexamples from the causal literature. And §4 discusses the question of whether the analysis I offer is an “analysis” of causation at all.

2. In this section I discuss technical stuff. I’ll begin by sketching a semantics. (The semantics is not new to this paper; but it may be unfamiliar to philosophers raised on a more traditional view.)

As I said above, I am interested in counterpossible counterfactuals: counterfactuals whose antecedents are impossible. Different counterfactual semantics have different things to say about such counterpossibles. On the standard Lewis-Stalnaker semantics, every counterpossible is trivially true. A number of recent philosophers, however, have suggested that our counterfactual semantics should allow us to endorse the following claim:

Counterpossibles

---

1 Acknowledgments removed
2 Hall and Paul (2013: 249).
5 The standard caveats apply (see Lewis 1973a for more): I will be analyzing only causal relations which obtain within deterministic settings; assuming that causation relates only particular events (whatever those are!); etc.
6 There are different conceptions of what the “impossibility” mentioned here amounts to, which I won’t discuss: see Jago (2014).
7 See Stalnaker (1968) and Lewis (1973b). See also Williamson (forthcoming) for several related proofs, and Starr (2019) for a fuller story than the standard potted history.
There are non-trivially true and false counterpossible counterfactuals.  

And I will accept Counterpossibles, in the rest of the paper. 

There are ever so many semantics – much discussed, by those recent philosophers and logicians – that might make Counterpossibles true. I will not take a stand here on what the true and Counterpossibles-tolerating semantics really is. But as an helpful heuristic, I will adopt: 

Nolan\(^{10}\)

‘If it were the case that \(p\), then it would be the case that \(q\)’ expresses a truth in world \(w\) iff, in the most similar\(^{11}\) worlds to \(w\) – whether those worlds are possible or impossible – in which \(p, q\) 

To get a grip on Nolan, consider the counterpossible counterfactual expressed by ‘if Hobbes were to square the circle, then mathematics would not be any different’.\(^{12}\) According to Nolan, that counterpossible is true iff, in the most similar worlds to our own in which Hobbes does square the circle, mathematics isn’t any different. (Those most-similar worlds, moreover, are impossible: for squaring the circle would violate the laws of mathematics, which are naturally thought of as necessary.) And such a counterpossible is plausibly false: for in the most-similar worlds to our own in which Hobbes’ mathematical demonstration genuinely succeeds, the laws of mathematics are different.

Apart from a new counterfactual semantics, I will also need a new property. Intuitively, some things are alone in a world iff they all exist in the world, and every thing in the world is identical to one of them. But more precisely: 

Aloneness

\(a, b, ..., F, G, ...\) are alone in a world \(w\) iff (i) they all exist in \(w\) and (ii) for every \(x\) in \(w\), it is identical to \(a\) or \(b\) or… or \(F\) or \(G\) or… 

As I hope is clear, this definition is plural: ever so many things can be alone within a world. And as a term of art, let’s say that any plurality of things which are alone in a world form a lonely

---

\(^{8}\) Aficionados will know that there is a dispute in the literature about whether to understand counterpossibles in epistemic (Vetter 2016) or metaphysical (Brogaard and Salerno 2013, REMOVED) terms. For simplicity – as my adoption of the semantics I’m about to mention in the main text will make clear – I’ll assume that the latter way to read counterpossibles is correct. But I also don’t believe that debate will matter much for what I say, and I also believe that much of what I say might be paraphrased into terms which the epistemicist might accept. 

\(^{9}\) So for example we could make Counterpossibles true with semantics built in terms of Fine-style states (Fine manuscript), or Yablo-style states (Yablo manuscript), or any of the different kinds of impossible worlds mentioned in fn. 12. 

\(^{10}\) Nolan (1997). As will be familiar to those who work on counterfactuals, this heuristic takes a stand on a number of questions (e.g., on the Uniqueness principle [Stalnaker 1968]). But this is merely an heuristic, and I do not mean to pretend that I have somehow resolved those questions here. 

\(^{11}\) This sort of similarity is not the standard kind of similarity: see the discussion between Fine (1975) and Lewis (1979). As another heuristic, I will assume that the kind of similarity developed by Lewis (1979) is the right kind of entity in terms of which to understand the talk of similarity in Nolan. Nolan himself (1997: 543) appears at least open to adopting this Lewisian metric. 

\(^{12}\) The inspiration for this example, of course, is Nolan (1997: 544). 

\(^{13}\) I won’t have anything to say about what impossible worlds are, here; but for different theories about their nature, see Nolan (1997), Vander Laan (1997), Brogaard and Salerno (2013), Jago (2014), and Berto and Jago (2019).
(For ease of expression, I will sometimes say that this or that element of a lonely plurality is alone; but strictly speaking that is obviously false.) As I also hope is also clear, the quantifiers on its left-hand side are second order: the elements of a lonely plurality may include properties, events, states of affairs…

As may not be clear, the quantifiers in part (ii) of the definition’s right-hand side are restricted. They do not range over propositions (so I may be alone in a world in which there are also ever so many propositions about me). They do not range over the laws (so your property of being tall may be alone in a world where ever so many laws govern it). They do not range over the properties of those propositions or laws (so I may be alone in a world in which there are ever so many true propositions about me). And perhaps more.

As may also not be clear, aloneness is usually impossible. Suppose, for example, that existence were a property which every entity necessarily has. Then most ascriptions of aloneness could not possibly be true. (Thus, for instance, I cannot be alone: for if I were alone, then my existence would not itself exist.) Or suppose that the numbers exist necessarily. Then, again, most ascriptions of aloneness will be impossible. (Thus, for existence, the planets cannot be alone: for if they were alone, the numbers would not exist.) Or suppose that the vast majority of physical objects have parts. Then, once more, most ascriptions of aloneness will be impossible. (For if some physical objects are alone, then they will be partless.)

On to the last element of our technical apparatus. By way of introducing it, let me discuss a paradigm case of causation. A plane’s going into a stall, suppose, is caused by its angle of attack growing too steep (at a particular altitude, in the particular wind conditions, at the plane’s particular speed, etc.). And the existence of that causal relation, within the world where it obtains, requires a number of other things to be true, or to exist, or to obtain. Thus, for instance, it requires that the relevant physical laws operate on the wind and the velocity of the airplane in a certain way. It requires that the plane be built out of material relevantly similar to the material it’s built out of. It requires the existence of the plane, and the wind, and the astronomical body in whose gravity well the plane is operating. And so on.

Here’s a definition. Let me say that the things which are necessary for the causal relation between the plane’s going into a stall and its angle of attack growing too steep are the preconditions for that causal relation. Or, more generally, let’s say that the preconditions of a causal relation are the things whose obtaining (or being true, or existing, or…) is necessary for that causal relation to obtain.

There is much more to say about preconditions for causal relations than I’ve said here. But I should make one clarifying note explicitly, before moving on. For it can be unclear in what sense of ‘necessary’ preconditions are necessary for the causal relations for which they’re preconditions to obtain. So here is what I mean.

---

14 Perhaps the threat of certain paradoxes, which I won’t discuss here, will require the fans of this kind of talk to reduce (in some strong sense of ‘reduce’) pluralities to their elements. I’m convinced this can be done, but can’t say more here.

15 Some philosophers deny this. I’m not sure why: the property of existence is logically definable. (See Salmon (1998).) Some philosophers still deny that existence is a property, when confronted with such definitions. I would respond by changing the above argument to be about schmexistence. (See Kripke (1980: Lecture III.).)

16 Here I assume the falsity of certain views about the fusion relation, for more on which see Cotnoir and Baxter (2014).

17 Though the idea of a precondition is widespread within the philosophy of science, the idea of a precondition developed here is derived from certain things said by Railton (1980). Obviously he does not use the idea in the same way I use it here, and obviously he is not guilty of whatever sins I’ve committed in its use.
Suppose you were to construct a world which was a qualitative duplicate of the world in which there is a causal relation between the plane’s going into a stall and its angle of attack growing too steep, except that this new world lacked one of the preconditions. (So the only difference between this new world and the first one is a difference in one of the preconditions’ obtaining, or being true, or existing.) Then, in this new world, the causal relation would not obtain. (For one of the preconditions for its obtaining would not obtain.) And it is in that sense that every precondition \( P \) is “necessary” for the causal relations they support: every world which contains all the preconditions except \( P \) does not contain.\(^\text{18}\)

3.
In the last section I sketched a technical apparatus: a new counterfactual semantics; a new property; and the notion of a precondition. In this section I sketch a counterfactual analysis of causation in that technical apparatus’ terms. The analysis will look as follows:

**Counterpossible causation**

(i) For every world \( w \) with events \( a \) and \( b \), there is a set \( I \) of the impossible worlds which are the most similar worlds to \( w \) in which there is a lonely plurality with the following elements: \( a, b \), all the causal relations from \( a \) to \( b \) in \( w \), and all the \( w \)-ly preconditions of those causal relations

(ii) \( a \) causes \( b \) in \( w \) iff \( B \) depends counterfactually on \( A \) in every element of \( I \)

But before turning to objections, it’s worth walking through a simple case to show how the analysis is intended to work.

Suppose Ruzy throws a sock at a window, which causes the window to break. (It is a very heavy sock, and Ruzy is an expert thrower.) Our analysis entails that the window’s breaking depends counterfactually upon the sock’s being thrown in the most similar impossible worlds containing a lonely plurality with the following elements: Ruzy’s throwing the sock; the window’s breaking; all the actual causal relations between the corresponding events; and those causal relations’ preconditions. And it seems that, within those most similar impossible worlds, there is such a relation of counterfactual dependence. For go to those worlds most similar to our own which contain only those two events, and the actual causal relation between them, and all the preconditions for their standing in a causal relation. It falls out of standard Lewis semantics that, in the worlds most similar to those in which we remove the sock-throwing, the window breaking doesn’t obtain. And that’s sufficient for a counterfactual dependence relation to obtain.

But let’s turn now to counterexamples.

**First objection:** I throw a rock at a window, and you throw just after (at the same speed, from an equidistant position, with no wind…). My rock breaks the window, but your rock would’ve broken it, had I not thrown. What does Counterpossible causation say?

**Reply:** according to Counterpossible causation, the window’s breaking still depends upon my throw in the relevant worlds (just as it should). For in the most similar impossible worlds where *deep breath* my throwing the rock and the window’s breaking and all the actual causal relations between those events and their preconditions for those causal relations are alone… the window’s breaking depends upon my throwing the rock. For in the most similar worlds to those

\(^{18}\) You will note that this definition of the kind of necessity relevant to preconditions plausibly requires me to appeal to impossible worlds. (For the near-qualitative duplicate worlds may very well be impossible.) Evidently, such an appeal doesn’t bother me.
worlds which don’t include my throwing the rock, the window doesn’t break. (For you aren’t around, in that first set of impossible worlds; and so your throw isn’t around to cause trouble for my analysis, in that second set of them.)

Better still: according to Counterpossible causation the window’s breaking does not depend on your throwing the rock, in the worlds relevant to evaluating whether there’s a causal relation between those events. (And again that’s as it should be.) For in the world where your throwing the rock and the window’s breaking and all the actual causal relations between them are alone, the window’s breaking does not depend upon your throwing the rock. After all, there is no actual causal relation between your throwing the rock and the window’s breaking. So there is no causal relation between those events in the closest impossible worlds specified by Counterpossible causation. Nor are there any preconditions in such worlds (since there aren’t any causal relations for them to be preconditions for!). So there is no basis for saying that the window’s breaking (as it does, in such a world) depends counterfactually upon your throwing the rock, in the first set of impossibilia.

Second objection: consider now the famous trumping cases: “The major and the sergeant stand before the corporal, both shout ‘Charge!’ at the same time, and the corporal decides to charge.” The trumping case is a counterexample to a wide range of counterfactual analyses: for the major’s order causes the corporal to advance; but there appears to be no plausible way to say that the corporal’s advancing depends counterfactually on the major’s order. What does Counterpossible causation say about it?

Reply: if the major’s shouting and the corporal’s advancing and the causal relations between them from the actual world (and preconditions thereof) were alone, then the corporal’s advancing would depend counterfactually upon the major’s shouting ‘Charge!’, by just the same argument already provided. And if the sergeant’s shouting and the corporal’s advancing and the (non-existent) causal relations between them from the actual world (and the [non-existent] preconditions thereof) were alone, then the corporal’s advancing would not depend upon the sergeant’s shouting ‘Charge!’ – again, by the same argument just provided.

Third objection: the simple version, developed by Goosen (1979), of the cases which Hall and Paul (2013) call “virulent” (131) and “a particularly recalcitrant form of preemption.” (132) In such a case – and this presentation is oversimplified – for every time before the effect occurs there is some event such it occurs between that time and when the effect occurs, and which is also such that it would have caused the effect, had the actual cause not done so. (So no matter how temporally close or far you are from the effect, there are always blockers, within the relevant time-frame, acting on the counterfactual dependence relation between cause and effect.) What does counterpossible causation say about that?

Reply: well, there are causal relations (and thus counterfactual dependence relations) in the most similar impossible worlds in which the actual cause and effect and other things – the causal relations and preconditions specified by Counterpossible causation – are alone. But there are no causal relations (and thus no counterfactual dependence relations) in the most similar impossible worlds where the preempted causes and the effect – and those other things – are alone. So it seems like we can employ the same argumentative strategy as before: there is reason to think that, within our “first” set of impossible worlds, the effect does depend upon the cause; and also reason to think that, within the corresponding “first” set of impossible worlds, the effect does not depend upon any one of the (infinitely many) spurious causes.

19 Schaffer (2000).
4.
I have been terribly selective in my treatment of counterexamples thus far; and I have also
eschewed saying a number of things – which I try to say in a longer version of this paper – about
the counterpossible analysis. But hopefully I have given you some reason for hope that, where
analyses of the simplest form of causation in ordinary counterfactual terms failed, analyses in
counterpossible terms might succeed. In this last section I turn to the very different problem I
promised to address.

The very different problem has to do with whether or not I’ve offered a genuine analysis
of causation at all. After all, an analysis would let us reduce facts about causation to facts about
counterfactuals. But the right hand side of my biconditional mentions causal relations! (I is the
set of impossible worlds most similar to w where – among other things – certain causal relations
obtain.) So my alleged analysis looks like it clearly fails.

Let me start responding to that objection by offering a distinction which (I think) we
should all accept.20 Metaphysical fundamentality is a property which a thing has when it’s
among the set of things which give rise to everything else, while analytical fundamentality is a
property which a thing has when it’s unanalyzable. And these properties come apart. Electrons
are plausible candidates for metaphysical fundamentality, but any physicist could tell you that
they aren’t analytically fundamental. And – given the assumptions that (i) only words analyze
other words, (ii) the finitude of the English language, and (iii) the anti-symmetry of the analysis
relation21 – some words must be analytically fundamental; but words aren’t part of the
metaphysical fundament.

Now a clarification. I didn’t mean to be suggesting, when I offered you Counterpossible
causation, that causation isn’t analytically fundamental. (For all I know it is.) What I meant
Counterpossible causation to suggest was that causation isn’t metaphysically fundamental. For
Counterpossible causation shows that facts about causation can be given rise to (in some suitably
metaphysical sense) purely by facts about worlds.

Perhaps you think that capital-A Analysis, when philosophers engage in it, is a matter of
offering a linguistic or conceptual analysis of words or concepts in terms of other words of
concepts. If you think that, then by your lights I misspoke: for I have not offered such an
Analysis.22 But what I have offered is a picture of how facts about worlds (impossible worlds
determined by the specification in Counterpossible causation) give rise, in some suitably
metaphysical sense of ‘give rise’, to facts about causation. Whether or not that picture is an
Analysis, I hope you’ll agree that it is at least worth offering.

Perhaps you worry that the inclusion of causation among the specificatory criterion in
Counterpossible causation imperils my claim that facts about worlds alone give rise to facts

20 Thanks to (REMOVED) for teaching me about this distinction, long ago. This paragraph also owes its form,
though not its content, to Turner (2012).
21 The proof I have in mind here is structurally analogous to well-known proofs that, if justification is asymmetric
and each agent’s set of beliefs is finitely large and only beliefs stand in justificatory relations, then every agent must
have some fundamental beliefs. Note that I’m not assuming that the set of constructible sentences is finite, only that
the number of English words is. And note also that I offer this proof, not because I necessarily believe all of that
proof’s premises, but because it does a nice job of letting me spell out the distinction between metaphysical and
analytical fundamentality. (Thanks to comments by (REMOVED) on a related point for suggesting I include this
last sentence.)
22 I am not sure what such an Analysis is, though obviously my sympathies lie with those who put a metaphysical
spin on their account of it. For different and interesting takes, see (among others) Bealer (1982), Jackson (1998),
about causation. But compare: the specification of my dog might also require me to appeal to facts about the causal relations operating on him. (‘The dog who’s chasing the Pomeranian!’ ‘The dog who made me late for work!’ *Etc.*) Just as those facts about specification do not somehow make causation an element of the set of things which metaphysically give rise to my dog, facts about the specification of these impossible worlds do not make causation an element of the set of things which metaphysically give rise to those worlds. Or so it seems to me.\(^{23}\)

**Word count (w/o footnotes or works cited):** 2,935 (w/o works cited, footnotes, and this line)

**Works cited:**

Fine, Kit. Manuscript. What is meaning? Delivered at the 2019 Eastern APA.

\(^{23}\) Some might fuss about my use of ‘require’ in the last paragraph: can’t we use non-causal language to specify the dog? But imagine an entity – call her Emily – which could only utter the two sentences provided in the main text. It seems like Emily’s specification of my dog does require her to use causal language. (And again, that hardly makes causation among the grounds of my dog.) So those worried about my saying that our specification of the dog might require us to employ causal language would need an argument against Emily’s possibility. Thanks to [REMOVED] for a conversation, in another context, which inspired me to come up with the case just provided.


Yablo, Stephen. Manuscript. What is meaning? Delivered at the 2019 Eastern APA.