# Rescued from the Rubbish Bin: Lewis on Causation

Ned Hall†

Lewis's work on causation was governed by a familiar methodological approach: the aim was to come up with an account of causation that would recover, in as elegant a fashion as possible, all of our firm "pre-theoretic" intuitions about hypothetical cases. That methodology faces an obvious challenge, in that it is not clear why anyone not interested in the semantics of the English word "cause" should care about its results. Better to take a different approach, one which treats our intuitions about cases merely as guides in the construction of a causal concept or concepts that will serve some useful theoretical purpose. I sketch one central such purpose, suggesting, first, that an account of causation that, like Lewis's, gives a central role to counterfactuals is well-suited to fulfill it, and, second, that the most famous pre-emption-based counterexamples to a counterfactual account yield an important constraint on a successful account.

Let's start with Lewis's original analysis of causation, sticking to the deterministic case, here and for the rest of the paper. Lewis takes it to be a sufficient condition for event $C$ to cause distinct event $E$ that $E$ counterfactually depends on $C$, that is, that it be the case that, had $C$ not occurred, $E$ would not have occurred. Causation generally is analyzed as the ancestral of this relation.

This account is now almost 30 years old, and over the last three decades has come in for a quantity of criticism pretty much in proportion to its influence and importance—quite a lot, in other words. Some of this criticism focuses on foundations: Can the account be supplemented by an adequate theory of events? Of the counterfactual dependence relation? And so on. I will pass these criticisms by. Others take the form of specific counterexamples, either to the original analysis or to the dozens of refinements thereof that have appeared in the literature. I will pass most of these by, save for the most important, which is a kind of example that

†To contact the author, please write to: Ned Hall, Massachusetts Institute of Technology, Department of Linguistics and Philosophy, 77 Massachusetts Avenue, 32-d808, Cambridge, MA 02139-4307; e-mail: ejhall@mit.edu.

Philosophy of Science, 71 (December 2004) pp. 1107–1114. 0031-8248/2004/7105-0040$10.00
Copyright 2004 by the Philosophy of Science Association. All rights reserved.



https://doi.org/10.1086/428014 Published online by Cambridge University Press

has come to be called "late pre-emption." Here is an instance: Billy and Suzy both throw rocks at a window, each with deadly accuracy. But Suzy is a bit quicker, so her rock gets there first. Consequently, it is her throw, and not Billy's, that causes the breaking of the window. But the breaking depends on neither throw: subtract Suzy's throw, and Billy's throw would have been enough to break the window, and conversely.

Cases of late pre-emption earn their importance by their simplicity, intuitive force, and stubbornness: the literature is littered with failed attempts to solve the problem posed by them. What's more, we learn a lot about causation by getting clear on what makes them tick, and towards the end of the paper I will make a suggestion as to their proper diagnosis. I will get there by way of a detour through a discussion of a very different kind of challenge, the point of which is not merely to make yet more trouble for counterfactual analyses, but rather to draw out issues that I think such analyses must face squarely if they are to be developed in the most plausible and interesting fashion. (At any rate, the challenge is not very discriminating: if it succeeds in making trouble for a Lewis-style analysis, then it makes just as much trouble for a much wider range of accounts of causation.)

To see what it is, consider some attractive metaphysical theses, which taken together I will call the "basic ontological package." The first is that there is a non-trivial distinction, applicable to any possible world, between the categorical (or, if you like, non-modal) truths about that world, and all other truths. Lewis certainly took such a distinction on board; it is built into his famous doctrine of Humean supervenience. The second claim is that there is a distinction to be drawn between more and less fundamental categorical facts. For example, categorical facts about middle-sized objects are less fundamental than categorical facts about atoms and molecules, which in turn are less fundamental than categorical facts about fundamental particles. (Witness the appropriateness of calling those particles "fundamental." Note that it might be that there is no most fundamental categorical level, but for our purposes there is no harm in supposing otherwise.) The third claim is that the whole truth about the world is fixed by, first, its most fundamental categorical truths, and, second, what the fundamental laws are that govern it. Again, the word "fundamental" is meant to be doing real work here: e.g., the laws of the special sciences, if such there be, don't count. It is, rather, only in our fundamental physics that we find paradigm examples of genuine laws. In the deterministic case, it is useful to think of these laws as rules that determine how complete physical states of the world generate successive physical states. We will make use of that conception a little later.

Quite obviously, this basic ontological package is heavily informed by modern physics. One might think of it as metaphysics under the tutelage

https://doi.org/10.1086/428014 Published online by Cambridge University Press

of physics. It is equally obvious that Lewis subscribes to it, although it is perhaps worth noting that he would consider it redundant to mention the fundamental laws as part of what everything reduces to, given that we have already included in the reduction base the fundamental categorical facts. Of course this package will strike many as controversial, but that does not matter: it is sufficient to recognize that it is plausible and interesting enough to be worth adopting as a working hypothesis, in order to see what metaphysical inquiry will look like when constrained by it.

Metaphysical inquiry into causation seems to face a problem, for what room does the basic ontological package leave for a relation of causation as part of the furniture of the world? There is a quick and easy answer, which is that it leaves plenty of room, just so long as causation can be shown to reduce to the facts drawn from the basic ontological package. Whatever its other shortcomings, Lewis's account of causation appears to succeed brilliantly on this score. The facts about what events occur where and when presumably are, or are fixed by, the fundamental categorical facts, and relations of counterfactual dependence between events are presumably fixed by the categorical facts together with the facts about the fundamental laws. It is worth just pausing to note that for those enamored of the basic ontological package, this is a non-trivial accomplishment on Lewis's part, for observe that other popular approaches to causation—for example, Davidson's, which seems to need special purpose laws in order to "cover" any causal relation, at least under suitable descriptions of the relata—do not square so easily with a position in ontology that takes the only genuine laws to be ones of the sort exemplified in fundamental physics. But that's another story.

At any rate, the quick answer obscures a deeper challenge. To see what it is, we need to remember that for Lewis, a proper philosophical account of causation is supposed to be an analysis of our ordinary concept, and as such needs to pay careful heed to our intuitions about cases. The challenge is to say why anyone should care about such a thing. Compare our ordinary concept of "object," the concept that is highly discriminating, in that not just any old motley assortment of things collectively counts as composing a genuine object. One could imagine a philosophical account of this concept that made heavy metaphysical weather of it, and that indeed refined and extended it. My spies tell me that something like that appears in Aristotle, where, at the level of deepest ontology, one finds distinctions drawn between living objects, artifacts, non-artifactual inanimate objects, and, as against all these, mere aggregates. That is a metaphysics that makes our ordinary concept of object quite interesting and important. It is also, unfortunately, a metaphysics thoroughly undermined by those developments in modern physics that underwrite the

https://doi.org/10.1086/428014 Published online by Cambridge University Press

basic ontological package. It's not that we could not endorse the basic ontological package and, at the same time, go about pursuing a reductive analysis of our ordinary, discriminatory concept of object—presumably, by looking for feeble surrogates within the materials provided by the basic ontological package for the sorts of distinctions that Aristotle takes so much more seriously. But it is just not obvious why success in such an endeavor would be anything worth caring about.

This is a familiar story, and could have been told about any of a number of ordinary concepts, e.g. our ordinary concept of *life*, or of natural biological *function*, or our ordinary concept of *intelligence*, etc. Why isn't it a story we should tell about our ordinary concept of causation? And if so, then it seems that while the sort of philosophical account of causation that Lewis is pursuing—one whose standard of success is that it reproduce as faithfully as possible our intuitive judgments about application of the English word "cause"—*might* be of interest to semantics, it will have very little value to metaphysics, and none at all to philosophy of science.

The issues here are murky enough that I imagine there is a host of ways that one could respond to this sort of challenge. I will focus on just one. To see what it is, consider that philosophers have, for a while now, recognized that it is a legitimate exercise to begin with some ordinary concept, and then provide an account that refines or modifies it in some way so as to meet some important theoretical need. While such an account must not completely ignore intuitions about the application of the ordinary concept, it also need not defer to them wholesale. One way to put the point is that in pursuing such an account, we shift from treating intuitions as bedrock "data" to treating them more as guides or signposts to where interesting and useful concepts in the vicinity might be found.

Consider what has happened in philosophy of biology with discussions of the concept of *biological function*, as applied to traits of organisms. Typical analyses of this concept advert to the evolutionary history of the organism in question, and in so doing aim to frame a concept that will be useful in understanding the structure of certain kinds of biological explanation. These accounts do not ignore our ordinary judgments about the application of the concept of biological function; it is taken for granted, for example, that such accounts had better deliver the verdict that one of the functions of the heart in humans is to pump blood. But intuitions about certain sorts of cases are simply rejected as irrelevant. If Swamp Creature, formed by fluke collisions of molecules just moments ago, happens to be built exactly like a human, with a fully functional heart pumping its blood, intuition may say that its heart is performing its proper biological function—is "working properly"—even though Swamp Creature has no evolutionary history. Bad news for the foregoing philosophical analyses? No: evidence, rather, that intuitions about cases

https://doi.org/10.1086/428014 Published online by Cambridge University Press

have only limited value for the kind of inquiry those philosophers are engaged in.

What seems to me an interesting and underexplored line of inquiry is whether a similar approach might succeed in the philosophy of causation. Could we find some useful theoretical role that a concept of causation could play, one that would vindicate the tacit assumption of philosophers working in this area that they are doing more than just the semantics of the English word "cause"? I think so, and will close with a very brief sketch of one possible role.

Here is the main idea. We have an obvious interest in understanding the nomological structure of our world—how it is that any possible complete physical state of the world would evolve, under the laws, into successive physical states. We also face obvious limitations in coming by such an understanding, limitations both of information and of computational power. (Only the simplest physics homework exercises show the student *exactly* how some complete physical state evolves into successive physical states.) So if, in pursuing an understanding of the nomological structure of our world, the only sort of knowledge we could aim for was *that* sort of knowledge, then we would be sunk. We would seem to have need, then, of more manageable targets: propositions that give us useful information about the nomological structure of the world, but that do so in a more easily digestible form. My suggestion is that the concept of event-causation serves the purpose of providing us with just such propositions.

To see how this might play out, consider first an idealized case. We have a possible complete physical state of the world at some time $t_1$, and the physical state it evolves into at time $t_2$. Imagine carving up space into very tiny "cells," and identifying the complete physical states at the two times simply by identifying the states in each cell. (In the limit, the world we're imagining might be a cellular automaton world, of the kind described by Conway's game of "Life.") Now we focus on some cell at $t_2$, and ask: What, at $t_1$, is responsible for that cell's time-2 state? (We'll leave it just vague and intuitive for the moment what we mean by "responsible.") If the laws are particularly nice, it might be that we can single out for attention some small number of cells at $t_1$, and say that it is the fact that they have the states they do, then, that accounts for the $t_2$-state of the target cell. Or we could say the same thing by means of a bit of obvious terminology: instead of talking about the fact that the target cell at $t_2$ has a certain state, we could talk about an *event* that occurs at $t_2$, in that cell's region; likewise for the cell-states at $t_1$ that we pick out as the responsible ones. And we could summarize our findings by saying that these latter events are the time-1 *causes* of the former event.

Now let's relax the idealization slightly, recognizing that given physics as we know it, one would have to look at the entire backward light cone

https://doi.org/10.1086/428014 Published online by Cambridge University Press

of the given cell to say exactly why it has the physical state it does, at some time. But instead of asking a fine-grained question of the sort "Why does this cell have exactly this state at this time?", we might ask a more coarse-grained question, by focusing on some relatively gross feature of the cell, and asking what about the earlier state of the world is responsible for its having that feature. And then we could reasonably expect our answer to be fairly discriminating. Consider the example of Billy, Suzy, and the broken window. We notice a feature of the spatiotemporal region of the world that contains a breaking window—namely, that it contains a breaking window. We ask what, at the earlier time, was responsible for this feature. The answer does not need to advert to the physical state of every region in the back lightcone of the shattering. And so we say that it was only Suzy's throw that caused the breaking, and in the same breath that a multitude of other events contemporaneous with this throw were causes of its breaking *in exactly the way it did*.

Now, in laying out this sketch of a role for causation, I have been leaning heavily on the word "responsible," and of course that will look suspicious. But remember the guiding idea: we are looking for a way to capture, in a kind of piece-wise fashion, the way in which complete physical states evolve under the laws into other complete physical states. In slogan form, the concept of *event* allows for the piece-wise representation of physical states, and the concept of *causation* allows for the piece-wise representation of nomological relations between them. With that as our picture, we can take talk of bits (or features of bits) of the earlier physical state being "responsible" for bits (or features of bits) of the later one as a placeholder for some legitimate but yet to be specified relation that will serve our purposes.

And now we are ready to hear from the counterfactual analysis again. For counterfactual dependence between events seems a prime candidate for a relation that will allow us to capture, in an easily digestible fashion, niblets of information about the nomological structure of the world. If that's right, then a counterfactual analysis, if attractive at all, should be seen to be attractive not merely because of its success (if such there be) at recovering our ordinary intuitions about cases, but also because it succeeds in presenting us with a causal concept that serves a manifestly useful theoretical purpose.

But now there is a problem. For, from everything I've said, it would seem that the theoretical purpose I've sketched could be served simply by bare counterfactual dependence between events, full stop—with no need for the bells and whistles that counterfactual analysts inevitably attach in order to deal with recalcitrant cases such as late pre-emption. That is, if one of our principal aims in giving an account of causation is not, as it is with Lewis, to do old-fashioned conceptual analysis, but rather

https://doi.org/10.1086/428014 Published online by Cambridge University Press

to use our intuitions about cases as guides in the construction of a concept whose primary importance resides in its ability to serve the sorts of theoretical aims I have outlined, then why not simply take causation to be counterfactual dependence? Well, comes the reply, because we have powerful counterexamples to such an analysis, e.g. in the form of Suzy, Billy, and the broken window. Well, comes the rejoinder, those counterexamples are only of interest to someone who is doing good old-fashioned conceptual analysis; for someone who is not, they can be rejected just as thoroughly as can examples involving Swamp Creatures, in the philosophy of biology.

But I don't think that rejoinder is right. I think that late pre-emption has something important to teach us, even once we have turned our backs on the sort of philosophical project that makes accommodating firm intuitions about cases a non-negotiable desideratum. I'll just say what I think this lesson is, and plead limitations of space as my excuse for lack of detailed argument. I think that in order for a concept of causation to serve the purpose I have outlined, it needs to exhibit a kind of stability. Very roughly, it needs to be the case that when we recognize a certain sequence of events as possessing a certain causal structure, and when we see that type of sequence reproduced, in other circumstances, we should be entitled to assume that the given causal structure is likewise reproduced. Start with a case where Suzy is all alone, and throws a rock through the window. Duplicate the sequence of events connecting her throw to the breaking in an environment where Billy is also present, though too slow. I suggest that if causal concepts are to help us in mapping out the nomological structure of the world bit by bit, we should be entitled to "transfer" the causal structure the process exhibits in the simpler case over to the more complex case.

It helps to compare our concepts of dispositional properties of objects. Suppose I have an object that is soluble in water. That fact bears some intimate but tricky-to-spell-out connection to the following counterfactual: if the object were placed in water, it would dissolve. The two claims—that the object is soluble in water, and that this counterfactual is true—are obviously independent. For the object might be soluble, yet protected by devices that will render it insoluble, should it ever be placed in water; conversely, it might be insoluble, yet watched over by devices that will render it soluble, should it be placed in water. On the other hand, it is ridiculous to suppose that the truth of the counterfactual is never relevant to whether the dispositional property is present. So an adequate account of dispositional properties ought, on the one hand, reserve some important role for the counterfactual, and, on the other hand, respect a kind of stability requirement—perhaps the requirement that the object not gain

https://doi.org/10.1086/428014 Published online by Cambridge University Press

or lose the dispositional property merely by changing its extrinsic circumstances.

I think that cases of late pre-emption teach us an exactly parallel lesson about causation. Consider a case where Suzy is all by herself, and throws a rock at the window, breaking it. In that case, her throw is a cause of the breaking (and, significantly, it is also true that the breaking depends on the throw). But that causal fact cannot change merely because we introduce, as a feature of the environment of the causal process connecting Suzy's throw to the shattering—a feature *extrinsic* to this process—Billy, his throw, and the flight of his rock toward the window. Compare: the fact that a certain object is soluble in water cannot change merely because we introduce protective devices. Just as dispositional properties should remain stable under modifications of the environments of the objects that exhibit them, so too causal relations should remain stable under counterfactual modifications of the environments of those processes that exhibit them. The lesson of late pre-emption is to make vivid the need for such a stability requirement.

The sort of counterfactual dependence that Lewis took to be at the heart of causation quite obviously does not, all by itself, exhibit this sort of stability, most notably because it is far too easy for it to go out of existence (as in cases of pre-emption). That emphatically does not mean that relations of counterfactual dependence are irrelevant to causation. But it is a nice question—nice enough, perhaps, to be closing with—exactly how to construct an elegant account of causation that will reserve some privileged role for relations of counterfactual dependence, while at the same time respecting the thesis that causal relations should be stable. At any rate, in addressing questions such as this, we would do well not merely to consult our by now sizable roster of intuitions about hypothetical cases in an effort to "curve-fit" our account of causation to them. We should, in addition, ponder carefully what exactly we want this account of causation to accomplish.

https://doi.org/10.1086/428014 Published online by Cambridge University Press