Forthcoming in *Philosophical Studies*

How effects depend on their causes, why causal transitivity fails, and why we care about causation

Gunnar Björnsson

Göteborg University

ABSTRACT: Despite recent efforts to improve on counterfactual theories of causation, failures to explain how effects depend on their causes are still manifest in a variety of cases. In particular, theories that do a decent job explaining cases of causal preemption have problems accounting for cases of causal intransitivity. Moreover, the increasing complexity of the counterfactual accounts makes it difficult to see why the concept of causation would be such a central part of our cognition. In this paper, I propose an account of our causal thinking that not only explains the hitherto puzzling variety of causal judgments, but also makes it intelligible why we would employ such an elusive concept.

Sometimes, an event seems to fail as a cause of another *exactly because* the latter event was independent of – would have occurred without – the former.

But in other cases – cases of redundant causation – one event is a cause of another *even though* the effect would have occurred without that cause. Despite a recent flood of papers on causation and dependence, no known analysis of the concept of causation gives an adequate account of typical causal intuitions in all these cases: the relation between dependence and causation has remained something of a mystery.

In this paper, I try to dissolve the mystery. After some methodological preliminaries, I remind the reader of the important aspects of the problem, briefly discussing a number of cases where our causal intuitions are at odds with various recent attempts to analyze the dependence of effects on causes in terms of counterfactual conditionals. However, the main point of this section is not to criticize, but rather to canvas a variety of common causal intuitions and remind the reader how difficult it is to subsume these intuitions under a unified explanation. In the latter parts of the paper, I propose and elaborate on such an explanation of our causal judgments, one that handles intuitions of both redundant causation and causal intransitivity while making it intelligible why we commonly employ such an elusive concept. Even though the upshot is an account of causal *judgment*, it strongly suggests a kind of account of the features of reality to which our causal judgments are supposed to correspond: a kind of account of causal *facts*.

Preliminary methodological note

Throughout this paper, I will talk about 'typical' or 'standard' intuitions or judgments about what is and isn't a cause in various imagined scenarios: these intuitive verdicts are what analyses of causation are measured against. This raises three methodological issues that I want to address briefly at the outset. The first concerns the *evidential relation* between causal intuitions and analyses or theories of causation, the second concerns the *imaginary* character of the cases about which judgments are made, and the third concerns the *typicality* of the intuitions or causal judgments in question.

The evidential relation, as I will understand it, is one of inference to the best explanation. Our analysis of causation is primarily an attempt to describe the criteria employed when we make our causal judgments – it is in this sense an analysis of the *concept* of causation – and the fact that we employ these criteria is supposed to explain that we make the causal judgments we make. Obviously, this means that if our analysis of causation predicts that we make certain judgments about certain cases, that analysis is corroborated if this is indeed the judgments we make, and undermined if not. But it also means that we should strive to find a theory that fits into a more general understanding of the human mind and our conceptual capacities, thus yielding a more unified understanding of our causal judgments.

Some philosophers have been interested in the *physical nature* of causation much in the way that scientists have been interested in the physical nature of light, or heat, or biological inheritance, and it has been suggested that causal processes should be understood in terms of transference of preserved quantities.[1] If the physical nature of causation had been our subject matter, imaginary examples would have very little to offer our investigation. However, since our goal is to find the criteria employed in our causal thinking, the fact that the judgments against which the analyses are tested concern merely imagined or even unrealistic cases should not be a problem. Many of our everyday causal judgments are no doubt made in response to direct observation of the events concerned, but many others are made at least partly in response to descriptions of and various kinds of indirect acquaintance with the events. The assumption here, which I see no reason to question, is that our judgments are guided by the same fundamental criteria in both these kinds of cases. (If that should turn out to be erroneous, the theory sought for here would be a theory of our less observational causal judgments.)

As people working on the analysis of causation are well aware, not everyone makes the same causal verdict as everyone else in all cases. However, although causal intuitions differ, my experience based on the philosophical literature and discussions with non-philosophers is that some verdicts are very much more common than others, especially if we discount various theoretically motivated judgments that philosophers make while acknowledging a ‘pre-theoretic’ pull towards a different verdict. It is these ‘typical’ or ‘standard’ judgments that an analysis of the concept of causation can hope to explain. That is a challenging enough task.

A final preliminary remark: Sometimes, I will talk about events as causes; at other times, I talk about conditions or states-of-affairs as causal relata. In the final analysis, I would take events to be causal relata in virtue of being instantiations or non-instantiations of properties by objects at a time, but I don’t think that I presuppose any particular non-standard view in the arguments of this paper.

Counterfactual dependence and causation: some puzzles

The idea that effects *depend* on their causes seems to be an important part of our concept of causation, and it is natural to express this dependence in terms of counterfactual conditionals. In most cases where *c* is a cause of *e*, it seems correct to say that if *c* had not occurred, neither would *e* have occurred. Moreover, in cases where we deny that *c* is a cause of *e*, we often say things of the form ‘*e* would have been the case with or without *c’*. But as cases of redundant causation show – cases where *c* is a cause of *e*, but where, had not *c* occurred and caused *e*, some other event, *d*, would have – the concept of causation and the concept of counterfactual dependence cannot be identified. Take an everyday example:

> *The Elevator*: You are waiting for the elevator to come, and you push the button to make it stop at our floor; I also want it to stop at

> our floor and would have pushed the button if I had not seen you do it. Your pushing the button causes the elevator to stop, but the elevator would have been stopped without it.

The well-known counterfactual theory of causation proposed by the late David Lewis solves this problem:[2]

> *Lewis-causation I*: *c* is a cause of *e* if and only if *e* or some cause of *e* is counterfactually dependent on *c*.

This would make that causal relation *transitive*, and transitivity gives the right result in the elevator case. Think about the state of the elevator-guiding mechanism that is responsible for the elevator's stopping at our floor. Call the mechanism's being in this state at the time where I have just seen you push the button '*d*'. The stopping of the elevator is counterfactually dependent on *d*: at this time I already believed that my pushing the button would make no difference and so wouldn't have pushed it myself had *d* not been the case. (For the sake of argument, let us assume some principled constraint on admissible counterfactuals that rules out 'backtracking' counterfactuals such as 'Had *d* not been the case, you would have to not have pushed the button, in which case I would have pushed the button.') Moreover, *d* is counterfactually dependent on your pushing the button: if you had not pushed the button, I would have pushed it, but *at a later time*, too late to produce *d*. Hence, your pushing the button qualifies as a cause of the elevator's stopping.

Problems remain, however, when we turn to another family of cases, including:

> *John and Jill*: John and Jill each throw a pebble at a window and Jill's pebble gets there first, breaking the glass. Had not Jill's pebble been thrown, John's pebble would have broken the glass, but Jill's throw was nevertheless a cause of the breaking of the window.

This is a case of what Lewis calls 'late preemption' or 'late cutting', where the alternative cause is doomed – cut off from the effect – only when the effect is achieved by the actual cause: only when the glass has been broken by Jill's pebble is John's pebble rendered inefficacious, flying through the hole in the window. In such cases, there is no intermediate part of the process ranging from Jill's throw to the breaking of the window on which the breaking depends: in contrast to what common sense tells us, Lewis's early counterfactual theory does not count Jill's throw as a cause of the breaking. (As is well known, this will depend on how the effect is individuated: Jill's throw would certainly qualify as a cause of the breaking *exactly as it occurred*, for if Jill had not thrown her pebble, the window would have broken differently. The assumption here is that the breaking is individuated such that the counterfactual 'If Jill had not thrown her pebble, the window would still have broken' comes out correct.)

Lewis's response was to define causation in terms of a notion of *quasi-dependence*, where an event, *e*, quasi-depends on another event, *c*, if and only if they are distinct parts of a process – events taking place in a spatiotemporal region – sharing intrinsic character with a nomologically possible process in which the counterpart of *e* depends on a counterpart of *c*:[3]

> *Lewis-causation II*: *c* is a cause of *e* if and only if *e* or some cause of *e* counterfactually depends or quasi-depends on *c*.

When deciding whether Jill's throw counts as a cause of the breaking, we are thus allowed to consider the nomologically possible scenario where John does not throw his pebble. In that scenario, dependence is reinstated. This means that the actual breaking quasi-depends on Jill's throw, which thus qualifies as a cause of the breaking. Not only does this move give the right result for both early and late cutting, it also seems to latch on to some of our intuitive reason for our judgments in the above cases. Take the elevator case. Your pushing the button caused the elevator to stop because disregarding events external to what happened in the elevator mechanism following your push – this constituting a process – the stopping depended on the pushing. Since my desire had nothing to do with that process, the fact that it would have been at the start of such a process is quite irrelevant. Moreover, my desire is *not* a cause since the stopping didn't depend on it, and since there was no process along which my desire brought about the stopping.

Although some improvements have been made that I will try to preserve when giving my own account of the relevant kind of dependence, trouble abounds. For example, Jonathan Schaffer has recently proposed the following puzzle case for theories taking counterfactual dependence to be central to our concept of causation:[4]

> *Merlin and Morgana*: In a world somewhat like ours but with a magic realm, thorough magical experiment has shown that enchantments only take place at midnight, and that the enchantment invariantly matches the content of the spell cast at the highest altitude the preceding day. Spells cast at lower altitudes only match the enchantment if they also match the spell cast at the highest altitude.[5] One day, Merlin casts a spell at the top of the mountain to turn the prince into a frog; the same day, Morgana casts a spell with the same content from down the valley. No other spells are cast. The prince turns into a frog at midnight.

Most people seem to agree that Merlin's, but not Morgana's spell caused the enchantment, but it is unclear whether there are causal intermediaries such that Merlin's spell, but not Morgana's, could be seen as a cause of these. Hence, it is unclear whether Lewis's original theory avoids saying that neither spell is a cause: transitivity doesn't solve the problem in any obvious way. Furthermore, it seems that the appeal to quasi-dependence of Lewis's refined theory gives the opposite result: the enchantment would seem to quasi-depend on *both* spells. Neither is a welcome result if we want to account for the standard intuition. Now, Michael McDermott argues that Morgana's spell *did* cause the frogification, and part of his argument is that both Merlin's and Morgana's spells are lawfully sufficient conditions for the enchantment.[6] But his is hardly a standard reaction, and seems to be based on the unavailability of a theoretically pleasing way of explaining the distinction between the two spells: I hope to be providing such a way.

One might of course doubt the value of intuitions concerning examples that are pure fantasy when deciding what causation really *is*. But the primary objective of this paper is to give a unified explanation of our causal *judgments,* and most people are happy to make causal judgments about *Merlin and Morgana*. Moreover, Lewis himself was concerned enough to propose a new theory employing the notion of *influence*:

> Where C and E are distinct actual events, let us say that C *influences* E if and only if there is a substantial range, C_1, C_2… of different not-too-distant alterations of C (including the actual alteration of C) and there is a range E_1, E_2… of alterations of E, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on. Thus we have a pattern of dependence of how, when, and whether upon how, when, and whether.[7]

This is used to modify Lewis' original theory:

> *Lewis-causation III*: *c* is a cause of *e* if and only if *e* or some cause of *e* is influenced by *c*.

One striking problem with this definition is that it sees causation where we do not: a rain delays and changes the direction and rapidity of a forest fire, but we find it absurd to say that, for this reason, it caused the fire.[8] Perhaps, though, the notion of influence is enough to let us say why, in cases of redundant causation, the cause is a cause and the potential cause is not: the former has *more* influence on the effect than the other. Then we get something like:

> *Lewis-causation IV*: *c* is a cause of *e* if and only if *e* or some cause of *e* is either counterfactually dependent on *c* or more influenced by *c* than by any event on which it would have been counterfactually dependent had not *c* occurred.

As Lewis points out, this seems to give the right result in the case of John and Jill. Small alterations in the speed, spin or timing of Jill's throw or the size of her pebble makes a lot of difference to the details of how the window breaks – to the shape of the shards and the direction they fall, say – whereas small alterations in the speed, spin or timing of Jack's throw or the size of his pebble makes much less of a difference to the breaking, going little beyond minute variations in the gravitational field, say.[9] It also seems to give the right answer to the *Merlin and Morgana* case: small alterations in the content of Merlin's spell makes corresponding differences to the enchantment, whereas small alterations of Morgana's spell makes no such difference.

Unfortunately, this does not help with some other cases. For example, consider:

> *Merlin, Morgana and Cerridwen*: From the highest mountain in the magical world, Cerridwen is watching the tired Merlin make his way up his mountain to cast the spell to turn the Prince into a frog. Wanting to make sure that Merlin does not fail for some reason or other, Cerridwen is ready to cast a frogification spell at her mountaintop, should Merlin fail. Merlin casts his spell; Cerridwen does not. Unknown to both, Morgana has cast her frogification spell down in the valley. No other spells are cast that day, but had Merlin cast a different spell, or none at all, Cerridwen would have.

In this case, Morgana's spell does not in the least influence the fate of the Prince, but neither does Merlin's. On all reasonably small variations of Merlin's spell, the Prince will turn into a frog, and it is not obvious that transitivity will help one bit. By contrast, Cerridwen's intention to cast a spell should Merlin fail has obvious and great influence on the outcome: there are many small variations of her intention that will make a great difference to what happens at midnight. And yet it seems clear that Merlin's spell caused the outcome whereas Cerridwen's intention did not.

Other problems arise from the fact that all Lewis's theories take causation to be transitive. This assumption is at odds with our intuitions concerning the following example, discussed by Ned Hall:[10]

> *Plain Switching:* I switch a train from track A to track B. Both tracks run parallel and merge just before the station a few miles later. The train arrives at the station, but my switching made no difference in that regard: it would have reached the station equally well along track A.

The standard (pre-theoretic) intuition about this case seems to be that my switching *didn't* cause the train's arrival at the station, although the arrival depended on the train's running on track B later on, which depended on the switching. Hence, the causal relation, as ordinarily grasped, can't be identified with a chain of dependence holding between the relata. Moreover, the reason why the arrival wasn't caused by the switching seems to be exactly that the arrival didn't *depend* on the switching. So although both Lewis's revised theories keep counterfactual dependence at the core of our concept of causation, they seem to demand *too little* dependence.

Now, Hall argues at length that our intuitions in this kind of case should be discounted. However, his main reason for doing so seems to be that there is no relevant difference between this case and other cases where we have quite different intuitions. But the only way to decide whether some difference is relevant is to have an adequate analysis of the concept of causation, and one of the criteria of adequacy for such analyses would seem to be whether they accommodate common intuitions. So, if we can come up with an analysis that handles switching cases and is doing as well as other analyses, *that* analysis should be the judge. And suggesting such an analysis is exactly what I propose to do later on in this paper.

However, let me say a few words about a response of Hall's that does not question our intuitions about the switching. Consider this modified case:

> *Contrived Switching:* "… hold the details of the arrival fixed, but alter the extraneous events so drastically that the way the train gets to its destination, in the counterfactual situation in which it travels [along A], is completely different from the way it, in fact, gets to its destination: it stops after a short while, gets taken apart, shipped piecemeal to a point near its destination, reassembled, and all this in such a way as to guarantee that nothing distinguished its counterfactual from its actual arrival."[11]

The added complication of the counterfactual scenario radically decreases the urge to disqualify the switching as a cause. Why? One explanation would be that, in spite of instructions to keep actual and counterfactual arrival indistinguishable, we take the contrived journey to delay the arrival. To avoid this source of error, let a group of magicians handle the dissembling, shipping and reassembling at the speed necessary. I do not think that this changes the reaction much, however. Hall himself has a different explanation of the difference in reactions to *Plain Switching* and *Contrived Switching*. Using an idea from L. A. Paul,[12] he suggests that it might be because in the former, the actual and the counterfactual routes were *similar* enough for us to see them as the same event – traveling towards the station – although with different aspects (along A, and along B, respectively). Suppose that he is right in that regard, and suppose that we accept Lewis's original version of the counterfactual theory. Then we could say that although traveling towards the station caused the arrival, the switching didn't cause the traveling towards the station but only an aspect thereof on which the arrival did not depend, and so didn't cause the arrival, even assuming transitivity. Apparently, Hall's suggestion could save some version of the counterfactual analysis from this kind of counterexample. However, a closer look reveals that the *similarity* of actual and counterfactual journeys along tracks is largely irrelevant to our negative causal judgment in the initial case. Consider:

> *Reversed Contrived Switching:* I switch a train from track A to track B. Both tracks run parallel and merge just before the station a few miles later. The train arrives at the station, but my switching made no difference in that regard: it would have reached the station equally well along track A. What it did, however, was to completely change the way the train got to its destination: in stead of going straight along the tracks in its normal fashion, it was now stopped after a short while, got taken apart, shipped piecemeal to a point

> near its destination, reassembled, and all this in such a way as to guarantee that nothing distinguished its counterfactual from its actual arrival (perhaps some magic was needed to get there in time).

Was changing the switch among what caused the train's arrival at the station? I believe most people regains the intuition from *Plain Switching*, denying that the switching caused the arrival. However, the two journeys are as dissimilar as in *Contrived Switching*: this is not at all what one should expect given Hall's explanation. A more straightforward explanation – which I will substantiate and qualify later on – seems to be that the contrived counterfactual journey was *less obvious*, and that this reinstates an appearance of counterfactual dependence of the arrival on my switching.[13]

Here is a different case of failing transitivity:

> *Cory's Scurvy:* Cory brought a bottle of vitamin pills when boarding for her sail across the Atlantic. Unfortunately, the pills were lost during the first storm, and after some weeks Cory had a bad case of scurvy going.

Assume that bringing her pills did not cause Cory to board, or to miss out on some other source of vitamin C. In that case, Cory's contraction of scurvy depended on her not having her vitamin pills, which depended on her losing them during the storm, which depended on her bringing them on board in the first place. But it makes no sense to say that bringing a bottle of vitamin pills caused her scurvy. Why? Apparently because scurvy would have been a problem anyhow: because dependence fails. And yet all Lewis's theories seem to say that bringing the pills on board was among what caused Cory's scurvy.

We might want to say, as Hall does in discussing and dismissing a somewhat similar case, that causation demands a causal process: perhaps bringing the bottle didn't cause the scurvy because there was no *causal process* connecting the two.[14] Of course, to understand this suggestion we need some idea of what a causal process is. Perhaps we could say that a causal process is a series of counterfactually dependent *positive* events or states-of-affairs. It is because losing the bottle caused Cory *not* to get enough vitamin C that she got scurvy, but not getting enough vitamin C isn't a positive event or a positive state-of-affairs: hence bringing the bottle did not cause the scurvy. But whether or not absences or negative states-of-affairs are ontologically dubious, their being a part of the chain of dependence is not by itself what disqualifies Cory's bringing the bottle as a cause of her disease. For it seems straightforwardly true that lack of vitamin C causes scurvy, that losing the bottle caused a lack of vitamin C, and that Cory's bringing the bottle on board caused her to lose it. (Admittedly, the last claim might seem awkward. The reason for this, I suggest, is that for most practical purposes, we expect there to be more salient and proximate and practically relevant causes of the loss of the bottle than bringing it on board. You would not find the claim odd, however, if you were Cory's sister, constantly worried about her bringing too many things along and losing half of them. By contrast, whatever your interests and prior expectations, you will find it implausible that Cory's bringing the bottle caused her scurvy.)

Lewis took an opposite line of defense, arguing that our reluctance to ascribe causation in cases like this stems from mistaken assumptions: in the end, we should accept that bringing the bottle caused the scurvy.[15] However, Lewis's defense rests on the assumed impossibility of accounting for early cutting in a theoretically pleasing way without invoking transitivity. The account presented in the following sections solves that problem, and so does an interesting recent version of the counterfactual theory by Christopher Hitchcock – or so it seems.[16] Consider again *The Elevator*. I walk to the elevator, just in time to see you push the button to make it stop at our floor: shortly thereafter, it stops. Had you not pushed the button, I would have, and the elevator would have stopped. One intuitive way of explaining why your push did cause the elevator to stop is to say that I did not in fact push the button, and *given that*, the stopping depended on your pushing. Hitchcock's presentation of the theory is technically complex, but the following sketch will do:

> *Hitchcock-causation:* An event *c* is a cause of an event *e* if and only if

> (a) there is a true counterfactual of the form ‘if some event alternative to *c* had been the case, *e* would not have occurred’ or
>
> (b) there is a counterfactual of the form ‘if some event alternative to *c* had been the case, and *d* had still occurred, *e* would not have occurred’ such that (i) *d* is some actual event (or set of events) that is distinct from the putative cause and effect and (ii) the antecedent of the counterfactual isn’t too remote, far-fetched or absurd.[17]

This handles our judgments regarding *The Elevator* nicely: ‘If you had refrained from pushing the elevator button and I still had not pushed it, the elevator would not have stopped’ seems just fine. Moreover, it seems to handle John and Jill fairly well: ‘If Jill had not thrown her pebble and John’s pebble had still not hit the window, it would not have broken’ sounds all right.

More importantly, however, it seems to handle *Merlin and Morgana* very well. Merlin’s spell qualifies as a cause, for ‘If Merlin had cast a spell turning the Prince into a lizard, the Prince would not have turned into a frog’ is obviously true. By contrast, Morgana’s spell does not pass the test. The counterfactual ‘If Morgana had cast a spell turning the Prince into a lizard and Merlin had still cast a spell with the same content as Morgana’s, the Prince would not have turned into a frog’ is of course true. Moreover, its antecedent might seem about as far-fetched as those of the counterfactuals used to establish your pushing the elevator and Jill’s throwing the pebble as causes. But the event that Merlin casts a spell with *the same content as*

Morgana's is not conceptually distinct from the putative cause, so this counterfactual is disqualified. Furthermore, *Hitchcock-causation* yields the right result in *Merlin, Morgana and Cerridwen*: 'If Merlin had cast a spell turning the Prince into a lizard and Cerridwen had still not cast any spell, the Prince would not have turned into a frog' seems just fine. So it seems to be a clear improvement on Lewis's suggestions.

One of the selling points of the account is that it handles cases of intransitivity quite well. That seems true if we consider *Cory's Scurvy*. Looking for true counterfactuals of the form 'If Cory had not brought her vitamin pills and … would still have been the case, she would not have contracted scurvy', nothing springs to mind, or at least nothing with a non-remote antecedent. The available treatment of *Plain Switching* is not quite so convincing, however. The claim would be that any correctly formed true counterfactual of the form 'If I had not moved the switch from A to B but … would still have been the case, the train would not have arrived at the station' would have an antecedent that is quite remote. But consider: 'If I had not moved the switch from A to B and the train still wouldn't have passed the midpoint of track A, it would not have arrived at the station.' That counterfactual seems true, and the possibility given in the antecedent seems no more far-fetched than my refraining from pushing the elevator button even when you do not push it, or John's pebble not hitting the window even if Jill doesn't throw hers, which Hitchcock takes to be perfectly all right.[18] At the very least, this calls for clarification of the notion of remoteness. And we need *substantial* clarification: to say merely that these are possibilities that we do not in fact consider to be relevant for the purpose of making causal judgments would amount to little more than *ad hoc* tinkering.

Given a suitable clarification of remoteness, Hitchcock's account might have the means to explain the difference between *Plain Switching* and *Contrived Switching*, and between *Contrived Switching* and *Reversed Contrived Switching*. The possibility relevant to *Contrived Switching* is that I would not make the switch and the train would still not go through the whole contrived route. The possibility relevant to *Plain* and *Reversed Contrived Switching* is that I would not make the switch and the train would not travel down track A. The former seems less remote than the latter, which suggests that my switching might qualify as a cause in *Contrived Switching* but not in the other two cases, just as it should. But *Hitchcock-causation* seems to have no way of handling the following:

> *Explosive Switching:* As before, I switch the train from A to B, but then I blow up track A. However, since I am a nice person I would have left the track intact if I had not made the switch first. So if I had not made the switch, the train would still have arrived at the station along track A.

As in *Plain Switching*, it still seems unintuitive to say that my switching caused the arrival at the station, and exactly for the reason that dependence fails. And yet, a counterfactual that reveals Hitchcock-causation is not hard to come by: 'had I not switched the train over but still blown up track A, the train would not have arrived at the station'. Here, the antecedent is no more far-fetched than what was employed in the elevator case: 'If you had not pushed the elevator button and I still had not pushed it, the elevator would not have stopped'. Or take the following case from Michael McDermott:

> *Two Servants:* "I order my two servants to push in opposite directions against a moveable object: it stays still. ... intuition denies that the order caused the object to stand still."

Clearly, the object's remaining at rest depends on the order in the way needed for Hitchcock-causation: 'If I had not ordered my two servants to push but one of them still would have pushed, the object would not have remained at rest' is just fine.[19] Again, Hitchcock's analysis gives the wrong results. At least, it gives the wrong results unless the notion of a far-fetched antecedent is specified in such a way as to rule out these results.

The Source of the Difficulty

This concludes my discussion of the kind of cases that make it so difficult to provide a theory that captures the extent to which we think that effects must depend on their causes. On the one hand, it seems clear that sensitivity to dependence of effects on causes is an important part of our typical concept of causation, as witnessed by *Plain Switching*, *Reversed*

Contrived Switching, *Explosive Switching* and *Two Servants*. On the other hand, there is a variety of cases where dependence is not required for causation, as in *The Elevator*, *John and Jill*, *Merlin and Morgana* and *Contrived Switching*. So, while we take dependence to be an important feature of causation, we also refrain from taking certain factors into consideration when assessing independence. We ignore my readiness to push the elevator button, John's throw, Morgana's spell, and, in some cases, the features in virtue of which an unswitched train would have reached its destination along a contrived route. The various attempts to refine the counterfactual analysis of causation can all be seen as attempts to capture the extent to which various factors can be ignored in assessing dependence.

Although all versions of the counterfactual theory of causation that have been considered here have problems with some of the cases discussed, new modifications of the theory might of course provide solutions.[20] But there is a further problem of method: the modifications and complications considered so far leave us in the dark as to *why* we should be constantly occupied with causal relations; why we should attend to certain factors while ignoring others in assessing dependence. Unfortunately, this also leaves us without guiding principles when we try to accommodate anomalies, thus giving our complications a strong *ad hoc* appearance, not being guided by what originally motivated the theory.

In the following sections, I will sketch an account of our causal thinking that not only predicts the puzzling variety of causal judgments we typically make about the cases discussed *and* captures our reasons for these judgments, but also makes it quite intelligible that we have a concept of causation and care about causal relations. As I will try to show, the criteria to which our mechanisms for causal judgment are sensitive are exactly the *primary* criteria employed when considering something as a means to an end. Since these are criteria employed in all our practical dealings with the world, their involvement in such a conceptually basic notion as that of causation comes as no surprise.

Before I can state my account of causal thinking, however, I need to put forth some assumptions about the workings of our minds on which the account relies. In the next three sections, I will propose hypotheses about how we represent aspects of the world; how we employ these representations in instrumental reasoning; and finally about how our causal thinking is a form of virtual instrumental reasoning.

Representing Aspects of the World: The Constancy Hypothesis

You have been assigned to record in a notebook the seconds at which the sun is shining at a particular spot on your favorite beach next month. A very cumbersome way of doing this is to have an entry for each and every second, and put an ‘s’ after entries representing seconds during which the sun was shining, leaving the others empty. However, when temporally ordered, seconds at which the sun is shining on that spot will tend to come in large groups, by the hundreds or thousands. In order to save work and notebooks, it is therefore much wiser to record *intervals* during which the sun was shining. This could be accomplished by noting the second at which the sun starts shining and put an 's' there; then wait for the first second at which the sun isn't shining and write it down without the 's'; then repeat the procedure if needed.

In a procedure like this, there is a default or *ceteris paribus* assumption that the shining as well as the non-shining continue. What I suggest is that our mind works according to similar principles, primarily recording changes and differences while assuming constancies: call this the '*Constancy Hypothesis*'. For example, in considering that an object has a certain property at a certain time, we will naturally consider that state as part of an interval that continues, ceteris paribus, in both temporal directions.

Importantly, some of the things we can keep track of are states that come with default assumptions of certain *changes*. For example, clouds above your beach will normally be moving at a fairly constant speed relative to the sun and the beach, and the assumption will therefore be that the position of the cloud relative to sun and beach will change. Similarly, we can learn that the movements of clouds has implications for what times the sun will be shining at the beach: if the movement of the cloud is constant, the shining will vary. Movement is just one kind of state that implies changes, ceteris paribus; intentions form another important category, where, ceteris paribus, intentions to realize *g* is followed by the realization of *g*; and there are countless others such as states of decay, precipitation, and rejuvenation.

The psychological role of ceteris paribus assumptions is motivated by cognitive economy. If we make a ceteris paribus assumption that something will continue to be the case, we continue to assume that it will be the case unless *positive* evidence to the contrary is brought to attention: that is the very point of making a default assumption. Moreover, positive evidence against an assumption will equally have a ceteris paribus character, and can be defeated by further evidence. Notice what this means: in assuming that *p* will be the case, ceteris paribus, and assuming that the ceteris paribus clause is satisfied, we can conclude that *p* will be the case without *attending* to – investing cognitive resources in – various possible defeaters of the ceteris paribus clause. The possibility of this kind of cognitive strategy obviously relies on a cooperative environment, one in which the absence of evidence that a certain ceteris paribus constancy is ended is itself reliable enough evidence that constancy holds. Fortunately, we live in an environment where this holds for a wealth of constancies.

Determining Aspects of the World: Instrumental Reasoning

In performing a piece of instrumental reasoning, classically conceived, you start with a desire for some goal, *g*, to be realized; you believe that some 'supporting condition', *s*, obtains such that if you perform a certain action, *a*, under *s*, *g* will indeed be realized: these beliefs combined with the desire make you decide to perform *a*, ceteris paribus. (This is very simplified, of course, but in ways that are peripheral to my argument.) In many cases, instrumental reasoning has several steps. In such cases, the decision will rest on beliefs and intermediate desires in what I will call an ‘instrumental hierarchy’:

(1) P decides to make it the case that *a* holds.

(1) rests on (2), (3) and (4)

(2) P believes that if *s*' holds and P decides to make it the case that *a* holds then *a* is realized.

(3) P believes that *s*' holds.

(4) P has an intermediary desire that *a* is realized.

(4) rests on (5), (6) and (7)

(5) P believes that if s holds and *a* holds then *g* will hold.

(6) P believes that *s* holds.

(7) P desires that *g* is realized.

Instrumental reasoning leads to decisions and often enough to the realization of goals. But decisions are costly. They typically lead to actions that demand energy, time and attention, and might call for rethinking of previous plans. And they put constraints on further planning, unless revoked, which again takes rethinking. For that reason, the primary focus in decision-making is naturally on decisions given which the realization of our goals is necessary or highly probable. We also avoid making decisions for the purpose of goals that we know will be achieved without changes in our plans. We therefore tend to make decisions that seem necessary for the realization of our goals or without which their realization seems improbable.

The latter feature brings the following constraint: in making decisions, our confidence that intermediary and final goals will be realized is conditional on our confidence that the decision is made. Intuitively, this means that conditional beliefs, such as (2) and (5) above, will be included in instrumental hierarchies only insofar as it is supported by some basic assumption of constancy, where the supporting conditions include that the ceteris paribus clause for that assumption is satisfied. For simplicity, I will express this by saying that we include conditional beliefs in our instrumental hierarchies only insofar as we take the consequent to follow lawfully from the antecedent. (Since none of the cases of causation discussed here invokes probabilistic assumptions, I will say nothing about the interesting relation between probabilistic reasoning and causal intuitions.)

Now, given what I have just said, it might seem that we should take the following to be a condition for forming a decision: the non-realization of a goal should follow lawfully from the non-realization of intermediary goals and decisions, just as the realization of the goal should follow lawfully from the realization of intermediary goals and decision. In other words: we should take the realization of the goal to *depend* on the decision. However, when we do not know in advance whether a certain goal will be achieved *without* our decision, checking whether this is so might be cognitively cumbersome (if not practically inconvenient or impossible) and a certain degree of myopia rather expedient. Or more accurately, *some* checking for dependence will come for free: in focusing only on the building blocks in an instrumental hierarchy, one might already be considering conditions in virtue of which the goal follows without the decision. It is checking for further dependence defeaters that will take a further cognitive effort, beyond what is needed for the most basic instrumental reasoning.

Another aspect that will be part of basic instrumental reasoning is awareness of action-defeating side effects. Attention to side effects can of course be generally beneficial, for achieving one goal is no good if it means foiling another. Unfortunately, *general* scanning for potential bad consequences is an open-ended business with potentially huge demands on one's cognitive economy, going well beyond the most basic instrumental reasoning. However, one kind of awareness of side effects is *intrinsic* to forming an adequate instrumental hierarchy: an agent needs to be sure that her action to achieve a goal does not annihilate the supporting conditions for achieving the goal by that action. For that reason, it very likely that basic instrumental reasoning with respect to some goal should include awareness of side effects *qua* side effects, *should they be attended to when focusing on the building blocks of the instrumental hierarchy*. Looking for *further* side effects takes cognitive effort beyond what is necessary for simple instrumental reasoning, just like looking for further dependence defeaters, but ensuring the integrity of the instrumental hierarchy is no cognitive extra.

The suggestion, then, is that there is a basic *myopic* procedure of instrumental thinking the point of which is to put together a correct instrumental chain such that the decision will be lawfully sufficient for the realization of a given goal. This procedure *is* sensitive to whether the realization of the goal depends on the decision, but only if this can be determined on the grounds of factors already attended to in determining whether the decision would be sufficient for the realization of the goal: factors to which attention had already been *forced* by the effort of putting together an instrumental hierarchy, we might say. In normal decision-making, this myopic procedure is of course typically surrounded by various degrees of awareness of possible dependence defeaters and side effects and sensitivity to various degrees of uncertainty, but it is the basic unit to which such further reasoning is added and also the procedure which is used in thinking about various side effects. And, I suggest, it is the procedure by means of which we decide whether one event is among what caused another.

Thinking about Causes: Virtual Instrumental Reasoning

There is of course *prima facie* reason to assume an intimate connection between our causal thinking and our instrumental reasoning, as it seems that in trying to achieve a certain goal by some means, one is trying to *cause* the realization of that goal by the realization of the means. This has spurred some philosophers to try to define causation in terms of action, hoping that this will explain such things the asymmetry of the causal relation. None of these attempts have become very popular, primarily because it has seemed that causation must be a more primitive notion than action – it would seem that all action involves causation but not all causation involves action.[21] Whatever the merits of this argument are, however, my primary ambition here is not to define causation in terms of action but to explain our puzzling variety of causal judgments in terms of what criteria we would need to employ in instrumental reasoning: the question of what causation *is* must wait.

In the remainder of this paper, I will try to show how the variety of typical causal judgments discussed in the first section is what we should expect if causal judgments resulted from an application of the basic myopic procedure in a piece of virtual instrumental reasoning. If this is indeed the case, we not only have a unified explanation of a puzzling variety of cases, but also an explanation of why the concept of causation is central to our understanding of the world. If causal judgments are made according to principles that seem to be fundamental in instrumental reasoning, such judgments probably function to prime and adjust cognitive structures crucial to realizing goals in an effective way, letting us understand the world as opportunities for action.

There is of course one notable difference between instrumental and causal reasoning. Whereas the former is necessarily limited by a real lack of knowledge – prior to knowing our decision, we do not know whether the goal will be realized – we can equally well talk about what *did* cause a certain known event as we do about what *would* cause a certain kind of possible event. The resulting claim, then, is this:

> *The Sufficiency and Restricted Dependence Hypothesis*: In determining whether *c* is a cause of *e*, we try to determine whether both of the following hold:
>
> *Sufficiency:* some supporting conditions obtain in virtue of which *e* follows lawfully from *c* and
>
> *Restricted Dependence:* the realization of *e* does not follow lawfully from the non-realization of *c* together with conditions that we are forced to consider to determine that *e* followed lawfully from *c*, given full knowledge about facts about the situation (other than facts about what caused or would have caused what).

In a different context, I argue that an elaboration of this hypothesis concerning the connection between instrumental and causal thinking explains our intuitions regarding causal asymmetry (the movement of the tree caused the movement of the shadow rather than the other way around) and spurious correlations (the fall of the barometer did not cause the rain-storm), as well as the apparent intelligibility but apparent absence of backward causation. However, what has been said should be enough to let us explain the variety of typical causal judgments introduced earlier on.

Causal preemption and Restricted Dependence

I will now use the *Sufficiency and Restricted Dependence Hypothesis* (*SRDH*) to explain our judgments about the cases of early and late cutting and trumping. Start with *The Elevator*, our case of early cutting, where you pushed the elevator button and thereby caused the elevator to stop, even though I would have stopped if I had not seen you do it. According to *SRDH*, your action qualifies as a cause since, given an intact elevator mechanism, the stopping of the elevator follows lawfully from the pushing of the button. My desire for the elevator to stop and my readiness to push the button should you not do it are irrelevant, since they are beyond what we need to consider in order to determine that the stopping followed lawfully from your pushing. However, it might be thought that my desire also qualifies as a cause of the stopping, for the case involves circumstances such that the consequents of the following conditionals follow lawfully from their antecedents:

(1) If you push the button, the elevator will stop.

(2) If you don't push the button, I will believe this.

(3) If I believe that you don't push the button, I will believe that the elevator won't stop unless I push the button.

(4) If I desire that the elevator stops at our floor and believe that it won't stop unless I push the button, I will push the button.

(5) If I push the button, the elevator will stop.

From these conditionals it follows that:

(6) If I desire that the elevator should stop, the elevator will stop.

Now, (6) seems to express a law-like connection between my desire and the stopping of the elevator, for one could *ensure* that the elevator stops by making me desire for the elevator to stop. (This, of course, is a standard problem for theories taking causes to be *sufficient* conditions for the occurrence of effects.) The question, then, is whether *Restricted Dependence* is satisfied: whether considering conditions in virtue of which the stopping follows lawfully from my desire forces attention to conditions from which the stopping follows lawfully under the assumption that I do *not* desire for the elevator to stop. Here it might seem that we can know that (1) through (5) and hence (6) is true without being forced to consider the fact that you pushed the button. If that were the case, *SRDH* would erroneously say that my desire was a cause of the stopping of the elevator. But a closer look at the cognitive mechanisms involved in establishing *Sufficiency* reveals that *Restricted Dependence is* violated. Notice that (1) through (5) only yields (6) given:

(7) You either push the button, or you don't.

What I will argue is that if we know that you pushed the button and use the disjunction expressed by (7) in establishing *Sufficiency*, we are *forced* to consider the fact that you pushed the button. And given attention to that fact, *Restricted Dependence* fails. The stopping of the elevator follows lawfully from the fact that you pushed the button, taken together with the supporting conditions for (1), which were brought to mind when seeing that the stopping followed lawfully from my desire. And this, I suggest, is why we do not think that my desire qualifies as a cause of the stopping. My desire ensured that the elevator would stop, but your pushing the button caused the stopping and wasn't caused by my desire.

So, why does using the disjunction expressed by (7) force attention to the fact that you pushed the button? Because of the following:

> *Activating Belief*: If one believes that *p* and engages in causal reasoning employing a representation of *p*, or a representation that contains a representation of *p*, this will force activation of one's belief that *p*.[22]

Using the disjunction expressed by (7) in establishing *Sufficiency* for my desire involves employing a representation containing a representation of your pushing the button. By *Activating Belief*, this forces activation of the belief that you pushed the button.

As I see it, *Activating Belief* has strong theoretical support, given the hypothesis about causal thinking proposed in this paper, and given a representational model of human psychology. On this model, which I take to be quite well known (although not universally accepted), psychological states like believing, disbelieving, hypothesizing, desiring, etc involve inner representations as content-carrying parts. My belief that Todd is at home consists, in part, of a representation constituted by my concepts of Todd and of being-at-home, and so does my desire that Todd should be at home.

Where the belief and the desire differ is with regard to how they let the representation – their content-carrying part – participate in practical and theoretical reasoning. The function of the belief is to guide action that relies on Todd's being at home; the function of the desire is to produce actions that make it the case that Todd is at home.

Now, inner representations can, in turn, contain or be constituted by further inner representations in the way that complex sentences can contain or be constituted by less complex sentences. For example, if I consider the possibility that *Todd is either at home or at work*, my grasp of this possibility involves a complex (disjunctive) inner representation which in turn involves as constituents inner representations of Todd's being at home and of Todd's being at work, respectively. Moreover, just as I can assert the disjunctive statement without asserting that Todd is at home, I can have a belief that has the disjunctive inner representation as its content-carrying part without having a belief that has one of the disjuncts as its content-carrying part.

Suppose now that we believe that Todd is at home. The question we need to answer to assess *Activating Belief* concerns what happens if we employ the complex representation that Todd is either at home or at work when we believe that Todd is at home. Will employing the complex representation force activation of our belief that Todd is home in the way relevant for *Restricted Dependence*? Differently put: Suppose that we direct the kind of myopic attention hypothesized for virtual instrumental reasoning at the assumption that Todd is either at home or at work. Is that enough to activate our belief that Todd is at home? Yes. We have good reason to think that the cognitive activation of a representation that is the content-carrying part of a belief goes hand in hand with activation of that belief, for a substantial gap between the two would be detrimental to the point of having beliefs and indeed complex beliefs. Generally speaking, a belief is something one takes to be reliable enough to guide action. Consequently, for an inner representation to be the content-defining part of a belief is for it to have a certain practical and epistemological authority, an authority that shows in theoretical and practical inferences. Moreover, the general point of having beliefs constituted by *complex* representations is to transmit this kind of authority from representations constituted by one part of the complex to representations constituted by other parts. It is essential for such inferences to be sensitive to conflicts of authority, keeping us from jumping immediately from *p* and *if p then not-q* to *not-q* when we already believe in *q*, instead provoking further deliberation, say. And to have this sensitivity, the prior *belief* that *q* must be immediately available when the representation of *q* is activated. For that reason, I suggest that *Activating Belief* holds.

Activating Belief explains why we deny that my desire for the elevator to stop is a cause of the stopping. And it is equally at work in *John and Jill*, the case of late cutting where Jill's throw caused the window to break, but where the pebble thrown by John would have broken the window if Jill's had not. In considering conditions in virtue of which the breaking of the window follows lawfully from the occurrence of Jill's throw, it seems quite clear that we can find such conditions that do not mention the whereabouts of John's throw or his pebble. It is enough that the direction and kinetic energy of Jill's pebble at the point of release and the location and brittleness of the window were related in suitable ways, and that there were no obstacles along the trajectory of the pebble from the point at which Jill let it go to the point at which it reached the window. John's throw, by contrast, will fail as cause of the breaking for the same reason that my desire for the elevator to stop failed as a cause of the stopping. To understand why, notice that we cannot see John's throw as a lawfully sufficient condition for the breaking in the same way that Jill's throw was. There, the surface of the window was located in the trajectory of the stone; here, it is not. This, of course, corresponds to the intuitively crucial difference between the two throws: when John's pebble arrives, the surface of the window is no longer there to be hit and broken. Absent the condition that the surface of the window is located in the trajectory of the stone, we need to invoke some weaker condition to achieve sufficiency, perhaps that the surface of the window *was* in the right location when the stone was thrown and that is has either not been dislocated or been dislocated by being broken. And by *Activating Belief*, this forces attention to the fact that it was broken already and leads to a violation of restricted dependence.

Finally, consider Merlin and Morgana's spells as putative causes of the prince's transformation. Here, there is a sense in which Morgana's spell ensures the transformation, for the content of a spell matches the enchantment at midnight unless the spell cast at the highest altitude the same day has different content. But in considering Morgana's spell as lawfully sufficient for the enchantment, we are forced to consider the spell cast at the highest altitude that day and its content. The fact that Merlin cast his spell to turn the Prince into a frog at the highest altitude at which a spell was cast that day is thus part of the supporting conditions in virtue of which the enchantment follows lawfully from Morgana's spell. And given that condition, the Prince's enchantment follows lawfully even when it is assumed that Morgana does not cast her spell. Hence, Morgana's spell does not qualify as a cause of the enchantment. By contrast, there is no need to consider the content of Morgana's spell when seeing how the enchantment follows lawfully from Merlin's spell, because Merlin's spell *is* the spell cast at the highest altitude that day. As desired, Merlin's spell, but not Morgana's, comes out as a cause of the Prince's enchantment.[23] And the addition of Cerridwen to the story makes no difference to the outcome: Cerridwen's intention in *Merlin, Morgana and Cerridwen* is just as irrelevant here as my intention was in *The Elevator*.

Notice that, apart from giving the right verdicts in the above cases, *SRDH* seems to capture our intuitive reasons for the negative verdicts. Although my desire ensured the elevator's stopping given some further conditions, it fails as a cause because it only ensured the occurrence of either of two possible chains of events from which the stopping would follow lawfully, and the one that *did* occur did not depend on my desire. And although John's throw ensured the breaking of the window, it only did so by ensuring the occurrence of either of two possible chains of events from which the stopping would follow lawfully, and the one that did occur – starting with Jill's throw – ensured the breaking without itself being caused by John's throw. Finally, Morgana's spell did not cause the enchantment of the Prince because it only ensured the enchantment given the content of Merlin's spell, which ensured the enchantment without depending on Morgana's spell. In all cases where possible causes were preempted or trumped, they failed to satisfy the kind of restricted dependence to which instrumental reasoning is immediately sensitive.

Intransitivity and the Constancy Hypothesis

Although *SRDH* allows for chains of causation, it does not render the causal relation transitive. This is all well, for, as we have seen, transitivity frequently fails. First, consider *Plain Switching*. Here is, in rough outline, how we think that the arrival at the station follows lawfully from my switching, where (3) and (6) express lawful relations between antecedent and consequent:

(1) I move the switch from A to B.

(2) The train is heading towards the switch and the station.

(3) *If* I move the switch from A to B *and* the train is heading towards the switch and the station *then* the train will be heading towards the station along track B.

(4) follows from (1), (2) and (3):

(4) The train will be heading towards the station along track B.

(5) *If* the train will be heading towards the station along track B *then* the train arrives at the station.

(6) follows from (4) and (5):

(6) The train arrives at the station.

Notice that there is no mention of the status of track A anywhere in the deduction. Moreover, without recourse to the fact that track A was intact, we cannot see that the train would have arrived at the station had I not made the switch. So it could seem that *SRDH* is satisfied. If that were indeed the case, my suggestion as to how causal judgments are formed erroneously predicts a positive judgment in this case: that my switching caused the arrival.

However, attention to the *Constancy Hypothesis* will reveal that *Restricted Dependence* is violated. In explaining why my switching in *Plain Switching* didn't cause the train's arrival at the station, it is quite natural to say that since the train was *heading* in that direction already, it would have arrived there in any case. And the fact that the train was heading towards the switch and the station – item (2) in the deduction above – is indeed among the conditions needed to determine that train's arrival followed from the switching. Now, according to the *Constancy Hypothesis*, thinking about this fact brings in the constancy assumption that:

CA The train will arrive at the station, ceteris paribus.

Moreover, I suggest that in grasping the scenario of *Plain Switching*, we have assumed that there are no violations of the ceteris paribus clause for the train's continued counterfactual journey up to the station. Furthermore, the assumption that the ceteris paribus clause is in fact satisfied in the counterfactual scenario where I do not move the switch needs no further cognitive work: that is the nature of ceteris paribus assumptions. We can thus conclude, attending only to the factors involved in seeing how the switching was sufficient for the train's arrival at the station, that if I had not made the switch, the train would have arrived at the station. Hence, *Restricted Dependence* is violated, and this is why we think that the switching fails as a cause of the arrival.

Before turning to other kinds of switching, I will re-examine *The Elevator* in order to eliminate a common misinterpretation of my thesis, clarifying what *isn't* implied by the *Constancy* and *Sufficiency and Restricted Dependence Hypotheses*. Here is how these hypotheses might *seem* to predict the judgment that your pushing the button didn't cause the elevator to stop. First, the fact that you pushed the button seems to force attention to the fact that the button wasn't pushed before your action. This, in turn, brings in the constancy assumption:

CA* The button will continue not to be pushed, ceteris paribus.

That, in turn, might seem to force attention to conditions for this being the case, which in turn would force attention to the fact that if you had not pushed the button, I would have. Moreover, in establishing *Sufficiency* for your action, we assumed that the elevator mechanism was intact. This, together with the fact that if you had not pushed the button, I would have, seems to establish that if you had not pushed the button, the elevator would have stopped. So it might seem that *Restricted Dependence* would be violated. However, even if one would grant that we *are* forced to see that the button wasn't pushed before you pushed it (and this is quite doubtful), there is one *major* error in this reasoning. It is just false that an assumption of constancy such as CA* would force attention to the particular conditions in virtue of which the assumption does or does not hold true. In fact, if that were the case, focused cognition would be practically impossible, since pretty much *every* known fact would be brought to attention: through intermediaries, pretty much everything is lawfully related to pretty much everything. So it is just false that attention to CA* brings attention to my readiness to push the button. And since my readiness to push the button is not forced into attention, *Restricted Dependence* is preserved, along with the conclusion that your pushing caused the elevator to stop.

The variations of *Plain Switching* are handled equally well by the *Sufficiency and Restricted Dependence* and *Constancy Hypotheses*. First, consider *Contrived Switching*, Hall's variation where the counterfactual journey involves stopping the train, disassembling and reassembling it and having it start anew. The fact that the train was heading in a certain direction prior to the switch and would so continue does not help us all the way to the counterfactual arrival in this case. The ceteris paribus clause of CA is violated when the train ceases to head towards the station, and we are not forced to consider the factors responsible for the counterfactual journey following after the stopping and disassembly. Hence, we think that my moving the switch from A to B *does* qualify as a cause of the arrival in this case. Also, it is no wonder why, in *Reversed Contrived Switching* – the case where the *actual* journey was contrived and complex – we would lose the intuition from *Contrived Switching* that the switching caused the arrival: in deciding that my switching was lawfully sufficient for the arrival we are again forced to see that dependence was violated, just as in *Plain Switching*.

Turn to *Explosive Switching*, the variation of the switching case in which I blow up track A after having set the switch to track B. The reasoning would follow the same line as in *Plain Switching*, but in considering the train's counterfactual journey along track A, the fact that track A was destroyed would come to mind. And without attention to the fact that the destruction of track A was an effect of my switching, it would seem that *Restricted Dependence* is satisfied, for the ceteris paribus clause of CA would be violated. But as I argued when discussing instrumental reasoning, the basic myopic process of putting together an instrumental hierarchy will involve awareness of side-effects *qua* side-effects, *should they be brought to* *attention*. And the fact that track A had been destroyed was indeed brought to attention. Hence, the fact that the destruction of track A was an effect of my switching is also brought to attention, which means that the destruction is disregarded when we check for dependence. As a result of this, the ceteris paribus of CA holds, *Restricted Dependence* is again violated, and my action fails as a cause of the train's arrival. Again, we get the right result.

As I mentioned when introducing *Plain Switching*, Ned Hall claims that we are wrong in denying that my moving the switch was among what caused the train to arrive at the station. In particular, he argues that *Plain Switching* has the same structure as the following case, and that our causal judgments should be the same in both:

> *The Kiss*: Billy and Suzy have grown up. One day, they meet for coffee. Instead of greeting Billy with her usual formal handshake, however, Suzy embraces him and kisses him passionately, confessing that she is in love with him. Billy is thrilled – for he has long been secretly in love with Suzy, as well. Much later, as he is giddily walking home, he whistles a certain tune. What would have happened had she not kissed him? Well, they would have had their usual pleasant coffee together, and afterward Billy would have taken care of various errands, and it just so happens that in one of the stores he would have visited, he would have heard that very tune, and it would have stuck in his head, and consequently he would have whistled it on his way home [just as he actually did].[24]

As Hall points out, there is no question but that the kiss was among what caused Billy to whistle that tune. Furthermore, Hall claims that there is *no relevant difference* between this case and *Plain Switching* – the structure of counterfactual dependence seems to be the same, and both seem to be cases where one event interferes with a process that would have led to the effect in any case – and proceeds to argue that we should accept that my switching did cause the arrival, initial appearances to the contrary. But Hall is wrong in his claim that there is no relevant difference between the cases: they differ with respect to *Restricted Dependence*. In *Plain Switching*, the fact that the train was heading in a certain direction did the work of forcing attention to the counterfactual journey. But there is no sense in which, ceteris paribus, Billy was *heading* to a store where the tune was played such that this fact about Billy was a supporting condition in virtue of which Billy's whistling followed lawfully from Suzy's kissing him. Hence, there is nothing that forces attention to the supporting conditions for such a counterfactual history, which means that *Restricted Dependence* is preserved.

Contrast this with a case that *does* share the structure of *Plain Switching*:

> *The Dog Bite*: A terrorist plans to detonate a bomb. The day before, his dog bites off his right forefinger, so when he goes to press the button he uses his left forefinger instead.

It seems clear that the dog bite did not cause the subsequent explosion, and I would suggest that the *intention* to detonate the bomb is the crucial element, intentions being intuitively understood as *headings-to-the-intended-state-of-affairs*, and this brings the assumption that the intended state-of-affairs will be realized, ceteris paribus.

Two Servants is interestingly different. Here, *Restricted Dependence* fails because considerations of what was *already* the case are forced into the picture in a rather direct way. The reasoning seems to be as follows:

(1) I order my two servants to push in opposite directions against a moveable object, M.

(2) If I order my two servants to push in opposite directions against M, they will do so.

(3) follows from (1) and (2):

(3) My two servants push in opposite directions against M.

(4) If my two servants push in opposite directions against M, M stands still.

(5) follows from (3) and (4):

(5) M stands still.

My order caused my servant to the right to push, and his pushing is among what causes M to stand still, but my order did not cause M to stand still: this lead to problems for Hitchcock's theory. However, by the *Constancy Hypothesis*, the fact that M stands still brings attention to the fact that M *was* standing still before my order, which brings attention to the fact that it would continue to stand still, ceteris paribus. Given this and the assumption that the ceteris paribus clause holds absent my order, *Restricted Dependence* is violated. My order didn't cause M to stand still.

Finally, return to *Cory's Scurvy*. It seemed quite plausible that the loss of the vitamin pills in the storm was a cause of Cory's contraction of scurvy. It also seemed quite plausible that bringing the vitamin pills on board was a cause of her losing them in the storm. Nonetheless, it seemed very implausible that bringing the vitamin pills was a cause of her contraction of scurvy. Consider how (in rough outline) we might come to think that bringing the vitamin pills was lawfully sufficient for Cory's contraction of the disease:

(1) Cory brings her vitamin pills.

(2) *If* Cory brings her vitamin pills *then* Cory loses her vitamin pills in the early storm.

(3) follows from (1) and (2):

(3) Cory loses her vitamin pills in the early storm.

(4) *If* Cory loses her vitamin pills in the early storm *then* Cory doesn't have her vitamin pills during most of the trip.

(5) follows from (3) and (4):

(5) Cory doesn't have her vitamin pills during most of the trip.

(6) Cory has no good source of vitamin C except for the pills.

(7) *If* Cory doesn't have her vitamin pills during most of the trip *and* Cory has no good source of vitamin C except for the pills *then* Cory contracts scurvy.

(8) follows from (5), (6) and (7):

(8) Cory contracts scurvy.

Here, it seems that (8) follows from the negation of (1) together with (6) and (7). Or rather, (8) follows given the further assumption that if Cory does not bring her vitamin pills, then she doesn't have them during most of the trip. But this extra assumption is forced by attention to the negation of (1). If we think that the vitamin pills are not brought on board, then – by the *Constancy Hypothesis* – we will make the default assumption that they will continue not to be on board. Since nothing violates this assumption, we will conclude that (5) would hold if Cory had not brought the pills, and further – attending to (6) and (7) – conclude that she would have contracted scurvy. Hence, we can see that dependence fails without unforced consideration of factors outside the above deduction. By contrast, the fact that Cory brought the vitamin pills on board qualifies as a cause of her losing them in the storm, and her losing them qualifies as a cause of her contracting scurvy.

Variations in causal judgment

In my discussions of various puzzle cases, I have worked under the assumption that for each case, there is one typical and fairly clear pre-theoretic causal intuition to be explained. But not only do we find deviant causal judgments; the model of causal thinking proposed in this paper predicts or leaves room for considerable variation. Of course, on any model of causal judgment, judgments about a certain case will depend on how we conceive of that case. What I will do in this section, then, is to illustrate mechanisms that will yield variations given this particular model, focusing on the case of *Plain Switching*.

In *Plain Switching*, I suggested, attention to the fact that the train was *heading* towards the station invoked the following constancy assumption:

CA The train will arrive at the station, ceteris paribus.

Moreover, the assumption that the ceteris paribus clause was in fact satisfied in the counterfactual scenario needed no further cognitive work, that being the nature of ceteris paribus assumptions. Hence the conclusion that the train would arrive at the station without my action.

If this account of our causal judgments about *Plain Switching* is correct, one thing that could affect our judgments is the *strength* of CA. For example, suppose that, in *Plain Switching*, the train reaches the station along track B using only its momentum – track B is flat or slightly downhill, say – whereas it would have reached the station along track A only because the train driver would have given extra power to overcome a hill. Given this specification of *Plain Switching,* we can see that my switching was sufficient for the arrival assuming merely that the train was *rolling* towards the station. For a train to be rolling towards some point is for it to arrive at that point, ceteris paribus, and these ceteris paribus conditions holds in this version of *Plain Switching*. But notice that the ceteris paribus conditions here would be much stronger than those involved in the train's *heading* towards a certain point, given how I believe most people conceived of *Plain Switching*. A train heading towards a station will normally put its engines to use to overcome gravity, accelerate after having made stops for meeting trains, etcetera, before reaching its destination. We think that it is heading towards the station even so, for there are mechanisms in place that normally overcome gravity when going uphill, or get the train moving again when it has stopped. But an object that is *rolling* in a certain direction ceases to roll when it makes a stop, and there are no mechanisms that normally get objects rolling again in the same direction after a stop. What we have, then, are two different constancy assumptions:

CA^{H} The train will arrive at the station, ceteris paribus (Heading).

CA^{R} The train will arrive at the station, ceteris paribus (Rolling).

So, suppose that we conceive of the case in terms of the train's rolling towards the station. Then the constancy assumption made for the actual journey – CA^{R} – is no longer strong enough to take us all the way to the station in the counterfactual scenario. According to *SRDH*, we will now think that the switching *was* a cause of the train's arrival at the station. And I believe that this is exactly what happens if we focus on this version of *Plain Switching* and think of the train as a rolling object.

Making it clear that weaker assumptions are enough to ensure *Sufficiency* is one way to save *Restricted Dependence.* Here is another. Ceteris paribus assumptions are assumptions that something will continue to be the case unless things depart from the normal, tied to the assumption that things *are* normal unless there is evidence to the contrary. Given this, we should be able to change people's causal intuitions by making salient various threats to the constancy hypotheses, threats that remove the assumption of normalcy and force a more guarded conception of the case in question. And we are: the phenomenon is neatly illustrated by Stephen Yablo's variation of *Plain Switching,* in which track A is damaged as the train approaches and the repair team, finding itself short of time, cries out for me to switch the train over to B. I do so, and the train arrives safely at the station. However, by a miracle – let us say by hitherto unknown magical intervention – track A is fixed, and the train would have arrived at the station without my action.[25] Even though the explicit characterization of *Plain Switching* is consistent with this case, and even though both actual and counterfactual journeys can be exactly as we originally conceived of them when we denied that the switching was a cause of the arrival, these added aspects might change the way we think about how counterfactual dependence is violated. What attention to the damaged track in Yablo's case might do is to transform the reasoning by supplanting for CA^H the more explicit:

CA^{HI} If track A were intact, the train would arrive at the station, ceteris paribus (Heading on Intact track).

Notice that CA^{HI} seems to be just as strong as CA^{H}. The only difference is that the latter has the explicit condition of CA^{HI} hidden in its ceteris paribus clause. But that difference can make a big difference to our causal judgment. To arrive at the conclusion that the train would arrive at the station along the counterfactual route, we now need more than the constancy assumption and the assumption that ceteris paribus holds: we need the *explicit* assumption that the track would have been intact. But since this assumption is not part of what we were forced to consider in seeing that the switching was sufficient for the arrival, so *Restricted Dependence* is preserved and the switching now qualifies as a cause.

Concluding remarks

When David Lewis introduced his counterfactual theory of causation, he noted that:

> Hume defined causation twice over. He wrote "we may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second.* Or, in other words, *where, if the first object had not been, the second never had existed.*"[26]

Lewis's hope was that some version of Hume's latter definition would encounter fewer problems than various regularity theories of causation that build on the first definition and take causes to be lawfully sufficient conditions for their effects. If I have been correct, the situation might be almost the opposite: the best account of what is right about the second definition will start from exactly the idea that causes are lawfully sufficient conditions for their effects. Of course, I have only proposed a model of *causal judgment*, and said nothing explicit about *causation*. But if we take causation to be the relation between cause and effect that our capacity for causal judgment is designed to keep track of and to guide our action and inferences relative to, then it *seems* clear that causation is largely a matter of one event's following from another as a matter of law or basic constancy.

Admittedly, though, the model of causal thinking offered here falls short of a complete general account of causal judgment. For example, I have said nothing about how notions of probability or uncertainty enter causal reasoning, and I have said nothing about how judgments about type causation ('smoking causes cancer') are related to the judgments about token causation discussed here, and those are both issues that should be addressed in a complete account. Moreover, I have been satisfied with an intuitive understanding of various central notions. For example, I have not said much about what it is for something to be a *basic* assumption of constancy of the kind needed to establish *Sufficiency*, and I have said nothing about what kind of real constancies such assumptions correspond to when they are correct. It is tempting to say that they correspond to *lawful* correlations such that effects follow by law from their causes, given the circumstances, but then again I have said nothing about what a law is.

Lack of definitions of central notions might not be a comparative drawback to the model suggested here, however. As we have seen, counterfactual theories of causation tend to involve notions that have to be taken intuitively by demanding *substantial* ranges of *not-too-distant alterations of events* or by excluding *far-fetched* possibilities. Moreover, most accounts of how we go about assessing counterfactual conditionals, including Lewis's, suggest that we take into account lawful correlations among events or facts. This suggests that our grasp of lawful correlations is more fundamental in our thinking than our grasp of counterfactual conditionals. Regularity-based analyses of our concept of causation, such as suggested by the model of causal thinking presented here, thus seems to tie causation to a more fundamental aspect of cognition than do counterfactual analyses.

But what I want to stress is this: Assuming an intuitive understanding of the model, we got the right results in a variety of cases of redundant causation and causal intransitivity that have provided problems for theories of causation that take counterfactual dependence as their basic notion, and we got these results for what seemed like the right reasons. Moreover, if the model is correct, it is quite intelligible why we have a concept of causation that yields the puzzling variety of causal judgments discussed in the literature, and why it is so frequently employed. Causal judgments are made according to principles that are fundamental in instrumental reasoning and thus help us grasp opportunities for action afforded by the circumstances.

Notice what this gives us. *The Sufficiency and Restricted Dependence Hypothesis* is supported both from below and from above. It is supported by its success in predicting our judgments in the problem cases discussed here, but also by the arguments suggesting that finite beings capable of instrumental reasoning would be sensitive to sufficiency and restricted dependence in their everyday dealings with the world. This two-way support, moreover, is good methodological news. First, it means that we can hope for non-*ad hoc* modifications in the face of seeming anomalies, for there are no doubt further aspects of instrumental reasoning and general cognition that could shape our concept of causation. Second, since the concept of causation has been given a definite place in human cognition, it means that these two fields of research can start to inform each other. It is not so clear that counterfactual analyses have anything similar to offer.[27]

[1] See for example Phil Dowe's "What's Right and What's Wrong with Transference Theories", *Erkenntnis* 42 1995, pp. 363-74.

[2] "Causation", *Journal of Philosophy* 70 1973, pp. 556-67. Reprinted in *Philosophical Papers*, Volume II, Oxford: Oxford University Press 1986, pp. 159-72.

[3] "Postscripts to 'Causation'", *Philosophical Papers*, Volume II, Oxford: Oxford University Press 1986, pp. 172-213.

[4] "Trumping Preemption", *Journal of Philosophy* 97:4 2000, pp. 165-81.

[5] Schaffer's 'law of magic' says that it is the *first* spell of the day that matches the enchantment irrespective of other spells. I have changed the example to get rid of some considerations having to do with time, considerations that are irrelevant to the structure of the example.

[6] "Causation: Influence versus Sufficiency", *Journal of Philosophy* 99:2 2002, pp. 84-101.

[7] "Causation as Influence", *Journal of Philosophy* 97:4 2000, pp. 182-97.

[8] See Peter Menzies' "Counterfactual Theories of Causation" in *The Stanford Encyclopedia of Philosophy* (Winter 2004 Edition), Edward N. Zalta (ed.), URL=http://plato.stanford.edu/archives/win2004/entries/causation-counterfactual/.

[9] "Causation as Influence", p. 191.

[10] "Causation and the Price of Transitivity", *Journal of Philosophy* 97:4 2000, pp. 198-222.

[11] "Causation and the Price of Transitivity", p. 207

[12] "Aspect Causation", *Journal of Philosophy* 97:4 2000, pp. 235-56.

[13] Typical intuitions about *Reversed Contrived Switching* also contradicts John L. Mackie's suggestion that an effect is a condition that is necessary in

the circumstance for the effect *as it came about*. See pp. 44-6 of *The Cement of The Universe*, Oxford: Oxford U. P. 1974.

[14] "Causation and the Price of Transitivity", p. 201.

[15] "Causation as Influence", pp. 194-5.

[16] "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy* 98:6 2001, pp 273-99.

[17] "The Intransitivity of Causation Revealed in Equations and Graphs", pp. 286-7, e.g.. Notice that *Hitchcock-causation* is relative to a choice of alternatives to *c*, much as *Lewis-causation IV* is relative to an interpretation of "not-too-distant alterations" of *c*. Unlike Lewis, however, Hitchcock might not demand that alternatives come in "substantial" ranges.

[18] "The Intransitivity of Causation Revealed in Equations and Graphs", pp. 288-9.

[19] "Causation: Influence versus Sufficiency", p 95.

[20] In "De Facto Dependence", *Journal of Philosophy* 99:3 2002, pp. 130-48, Stephen Yablo makes an interesting suggestion that seems to avoid many of the problems suggested here. I believe that his account fails to account for the difference between Merlin's and Morgana's spells, but reasons of space prevent me from pursuing that argument here.

[21] Early attempts to define causation in terms of action are provided by Douglas Gaskins in "Causation and Recipes", *Mind*, 64 1955, pp. 479-487, and Frank Ramsey in "General Propositions and Causality" in *The*

Foundations of Mathematics, pp. 237-55, Totowa, NJ: Littlefield, Adams & Co. 1965. But most famous, perhaps, is Georg Henrik von Wright's argument that the concept of cause involves that of action. See his *Explanation and Understanding*, London: Routledge & Kegan Paul 1971, chapter 2. More recent attempts include that by Peter Menzies and Huw Price in "Causation as a Secondary Quality", *The British Journal for the Philosophy of Science* 44 1993, pp. 187-203. Criticism has been leveled by Jaegwon Kim in his "Review of *Explanation and Understanding*", *Philosophical Review* 82, 1973, pp. 380-8; by Judith Jarvis Thomson in *Acts and Other Events*, Ithaca: Cornell U. P. 1977; by Paul Horwich in *Asymmetries in Time*, Cambridge, MA: The MIT Press 1987, pp. 139-40; and by Ernest Sosa and Michael Tooley in their introduction to *Causation*, Oxford U. P. 1993

[22] The principle doesn't hold where there is failure of content identification. Knowing that Mr. Stevens is at home but unaware that Todd is no other than Mr. Stevens, there is clearly a sense in which one might employ a representation of Todd's being at home without activating one's belief that Stevens is at home. However, such failures play no role in our judgments about the cases discussed in this paper.

[23] This is how one might counter McDermott's objection to attempts to make a difference between Merlin's and Morgana's spells in terms of lawful sufficiency. See "Causation: Influence versus Sufficiency", pp. 89-90.

[24] "Causation and the Price of Transitivity", pp. 209-10.

[25] "De Facto Dependence", p. 146.

[26] "Causation", p. 159. The quotes are from Hume's *Enquiries concerning Human Understanding*, Section VII, Part II.

[27] Versions of this paper have been presented at the philosophy departments of Stockholm University and the University of Connecticut, and at CENSS 2002 in Ghent, Belgium. On these and other occasions, many people have provided useful comments, but I want to give special thanks to Ned Hall, Christopher Hitchcock, Lars-Göran Johansson, Ruth Millikan, Dugald Murdoch, Paul Needham, Jun Olivier, Johannes Persson and Stephen Yablo. The writing of this paper was made possible by grants from *The Swedish Foundation for International Cooperation in Research and Higher Education* and from *The Bank of Sweden Tercentenary Foundation.*