Philosophica 77 (2006) pp. 45-68

CAUSATION, PLURALISM AND RESPONSIBILITY

Francis Longworth

ABSTRACT

Counterfactual theories of causation have had difficulty in delivering the intuitively correct verdicts for cases of causation involving preemption, without generating further counterexamples. Hall (2004) has offered a pluralistic theory of causation, according to which there are *two* concepts of causation: counterfactual dependence and production. Hall's theory does deliver the correct verdicts for many of the problematic kinds of preemption. It also deals successfully with cases of causation by omission, which have proved stubborn counterexamples to physical process theories of causation. Hall's theory therefore appears to be a significant improvement on extant univocal theories of causation, both physical and counterfactual. In this paper I present a series of counterexamples to Hall's theory. I also describe cases in which our causal judgments appear to be sensitive to moral considerations. It does not seem likely that conventional theories of causation, which attempt to situate causation in an objective metaphysical picture of the world, will ever accord with our intuitions in such cases. Finally, the notion of responsibility is considered, but rejected as an illuminating primitive for analyzing causation.

1. Introduction

Univocal theories of causation have struggled to account for cases of causation involving preemption, and cases of causation by omission. And attempts to refine the basic theories in order to give the right results in these cases frequently introduce new counterexamples. Hall (2004) has suggested that causation is not a univocal concept. According to his pluralistic theory, causation comes in two varieties: production and dependence. This account delivers the correct verdicts for the preemption cases and cases of causation by omission that have plagued extant univocal theories and therefore constitutes a major advance. While I am

broadly sympathetic to Hall's pluralistic approach, I will show that his theory faces a number of counterexamples.

I will proceed via an examination of a series of candidate counterexamples to Hall's analysis and relevant rival univocal analyses. In section 2, I begin with what has seemed to many philosophers to be the most promising approach to analyzing causation, based on the idea that effects depend counterfactually on their causes. I will show that such counterfactual theories of causation are subject to a variety of counterexamples. One response to these counterexamples would be to abandon the search for a counterfactual theory of causation and pursue some sort of local or intrinsic theory of causation. But such an approach also faces seemingly insurmountable difficulties: counterexamples involving causation by omission. In section 3, I provide a brief exposition of Hall's pluralistic theory, which attempts to circumvent these canonical counterexamples. In section 4, I present several counterexamples to Hall's theory. In section 5, I examine the thesis that our causal judgments take into account moral facts, and suggest that is unlikely that traditional metaphysical theories will ever deliver the intuitively correct results in such cases.

2. The failures of univocal theories of causation

2.1. Naive dependence

David Hume, in the *Enquiry Concerning Human Understanding* (1748), pointed out a link between causation and counterfactual dependence:

> [W]e 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*. (1748, Section VII, Part II).[1]

The first sentence in the above quotation expresses Hume's familiar constant conjunction theory of causation. The second (which, despite

[1] Hume's italics.

Hume's claim, is *not* logically equivalent to the first) expresses a counterfactual theory of causation: if, running counter to actual fact, the first object had not been, the second would not have existed. The second object therefore depends, counterfactually, on the first.

Following Hume, an initial counterfactual analysis (Naïve Dependence, ND) can be formulated as:

(ND) C is a cause of E if and only if E *counterfactually depends* on C. In other words, if C had not occurred, E would not have occurred.[2]

It is well known, however, that effects do not always depend counterfactually on their causes. Consider:

> Trainee and Supervisor: Trainee and Supervisor are on a mission to kill Victim. Trainee shoots first and Victim bleeds to death. Supervisor, observing that Trainee has fired, does not shoot. If Trainee hadn't shot, however, Supervisor would have stepped in and done so, again resulting in Victim's bleeding to death.[3]

Although Victim's bleeding to death would have depended on Trainee's shooting in the *absence* of Supervisor, Supervisor's presence breaks this dependence. Such cases, in which the actual cause *preempts* some redundant backup, are known as cases of 'preemption'. Preemption therefore presents a problem for those who wish to base an account of causation on counterfactual dependence. Let us call the lack of dependence in cases of preemption the "preemption problem". The preemption problem has caused great difficulties for the counterfactual analyst of causation; indeed much of the literature on counterfactual theories of causation is concerned with attempts to get around the preemption problem by adding further conditions that deliver the intuitively correct theoretical verdict that preemption *is bona fide* causation, but without thereby introducing any new counterexamples.

---

[2] In keeping with common current practice, let us take the relata in (ND) to be occurrent events rather than Hume's 'objects'.

[3] Adapted from Hitchcock (2001).

Let us look at the first notable attempt to solve the preemption problem: David Lewis's appeal to the thesis that causation is always *transitive*.

2.2. Counterfactual dependence and transitivity

Lewis's counterfactual theory, presented in his seminal "Causation" (1973), was the first significant advance on (ND). His theory relied heavily on the assumption that the causal relation is transitive. Call this thesis 'Transitivity'.

> Transitivity: Causation is a transitive relation; that is, if C causes D and D causes E, then C is also a cause of E.

It seems intuitively quite plausible that causation is transitive: think of a line of dominoes toppling one after the other: the first causes the second to fall, the second causes the third to fall, and it seems correct to say that the first domino's falling is also a cause of the third domino's falling. It doesn't seem unreasonable to expect that transitivity would hold generally. In fact, Transitivity may be one of our central 'platitudes' concerning causation. The core of Lewis's analysis of causation can be summarized as:

(L) C is a cause of E if and only if there is a chain of intermediate events D1…Dn between C and E such that E counterfactually depends upon Dn, Dn counterfactually depends upon Dn-1,… and D1 counterfactually depends upon C.

The truth conditions of the counterfactuals are given in terms of the similarity of possible worlds to the actual world. Lewis stipulates that counterfactuals must not *backtrack*: if we are considering a world in which some event D_n, in a chain of dependency $D_1 \ldots D_n$, did not occur, D_{n-1} would still have occurred; so too would D_{n-2}, and all the other intermediate events stretching back to (and including) C. This is because such a world is *closer* to the actual world than a possible world in which C, $D_1 \ldots D_{n-1}$ do not occur, according to Lewis's similarity metric for possible worlds. We are to understand the non-occurrence of D_n, Lewis

says, as a "minor miracle": D_n is to be cleanly excised from the causal history of E, with *no* disruption of prior events.

Transitivity enables (L) to get the right result for our case of preemption, Trainee and Supervisor. Trainee's shot *is* linked to Victim's dying by a chain of dependence. Consider the flight of Trainee's bullet through some particular intermediate point *en route* to Victim (call this event 'B'). Event B depends counterfactually on Trainee's firing; if Trainee had not shot, B would not have occurred. In addition, Victim's dying depends counterfactually on B. To see why this is so, note that if B had not occurred, *Trainee would still have shot* (the 'no backtracking' rule). Hence Supervisor would not have shot, and Victim would not have died. Hence Trainee's firing causes B and B causes Victim's death. Invoking Transitivity, Trainee caused Victim to die. Does this appeal to the supposed transitivity of causation solve the preemption problem? Unfortunately not. There are other varieties of preemption for which this strategy does not work. Consider:

> Billy and Suzy: Billy and Suzy each throw a rock at a bottle. Suzy's arrives first and the bottle shatters. Billy's rock arrives a split-second later, encountering only flying shards of glass.

It is intuitively obvious that Suzy's throwing rather than Billy's caused the bottle to shatter, but in this case, there is neither simple counterfactual dependence between Suzy's throwing and the bottle's shattering, nor a chain of counterfactual dependence between them. In contrast to early preemption, we *cannot* say that if Suzy's rock had not been at some intermediate position *en route* to the bottle, the bottle would not have shattered, because *Billy would still have thrown*. Billy's throwing is independent of Suzy's throwing. Hence we cannot use the 'no backtracking' rule to argue that if Suzy hadn't thrown, Billy would not have thrown. Billy and Suzy is therefore a counterexample to Lewis's theory (L).

In Trainee and Supervisor, the backup process is cut short by Trainee's shot (the actual cause), early on. In Billy and Suzy, however, the backup process (the approach of Billy's rock) is only terminated at a very late stage, by the occurrence of the effect itself (the bottle's shattering). For this reason, these two cases are instances of what are referred to as *early* and *late* preemption respectively. While Lewis was

able to deal with early preemption counterexamples, late preemption counterexamples stalled the counterfactual research program for many years.

There are a number of responses that one might make to the problem of late preemption.

2.3. Causation as a local intrinsic relation

One response has been to attempt to define causation as a spatiotemporally local or intrinsic relation. Lewis (1986a), in his discussion of 'quasi-dependence' focuses on this approach. Suzy's throw, according to this approach would count as a cause of the bottle's shattering in virtue of the spatiotemporally continuous local and intrinsic relation that exists between the two events (corresponding to the trajectory of Suzy's rock).[4] In a somewhat similar fashion, one might describe the relation between Suzy's throwing and the bottle's shattering in terms of physical processes, perhaps involving transfer or exchange of energy, momentum or some other physical quantity. Fair (1979), Sober (1984), Salmon (1984, 1994) and especially Dowe (1992, 2000) have explored such approaches. While this seems a very intuitive and promising solution to the problem of late preemption, as a general analysis it faces considerable difficulties. There appear to be many relationships that we intuitively call causal which are neither intrinsic nor local. The major class of counterexamples is causation by omission. Consider, for example:

> Gardener: My plants died when I was away on vacation. If my gardener had watered them, as he was supposed to have done, they would not have died.

It seems correct in this case to say that the gardener's failure to water the plants was a cause of their death – perhaps even *the* cause. Yet there is no obvious spatiotemporally continuous series of events that connects the gardener to the plants; we may assume that the gardener was never in the vicinity of the plants, and our intuition remains the same. Gardener is

[4] See Hall (2004, p.235-257) and Menzies (2001) for further discussion, and attendant problems, of this approach.

therefore a counterexample to univocal theories based on intrinsicness or locality.

After abandoning quasi-dependence, Lewis (2000, 2004) returned to a purely counterfactual approach, redefining causation as the ancestral of "influence", a more fine-grained version of counterfactual dependence. This approach offers a potential solution to the problem of late preemption. I will not discuss the details here, but see Hall (2004) and Menzies (2001) for convincing objections.

2.4. Holding fixed

Further attempts to solve the problem of late preemption within a counterfactual framework have recently been advanced, which involve the notion of *holding fixed* certain facts or events. Notice that in the cases of early and late preemption above (Trainee and Supervisor and Billy and Suzy respectively), while the effects do not depend on their causes *simpliciter*, they *do* depend on their causes if we hold fixed certain facts. Victim's bleeding to death *does* depend on Trainee's shooting if we hold fixed the (actual) fact that Supervisor doesn't fire. Similarly, the bottle's shattering *does* depend on Suzy's throwing, if we hold fixed the fact that Billy's rock does not hit the bottle. By holding the right facts fixed, we are thereby able to reveal the latent dependencies between cause and effect that are hidden by the presence of the preempted backups. One simple candidate formulation of a "holding-fixed" counterfactual theory is:

(HF) C is a cause of E if and only if E counterfactually depends on C, while holding fixed some fact G.

(HF) bears a close relation to familiar epistemic methods for discovering causes in science. The Galilean notion of experiment involves trying to reveal causal relationships by manipulating some candidate cause and looking for an anticipated effect, while holding fixed any potentially interfering factors. (HF) therefore has some initial plausibility, and is currently a popular strategy in the causation literature; accounts giving a central place to some version of (HF) have been proposed, most notably, by Hitchcock (2001) and Yablo (2002, 2004), and also by Pearl (2000), Halpern and Pearl (2001, 2005) and Woodward (2003).

The 'holding fixed' approach seems to introduce two particularly problematic new types of counterexample, however: 'switches' and 'self-canceling threats.' Consider the following example of switching:

> Two Trolleys: Two parallel rail tracks ('left' and 'right') run alongside one another towards a movable section of track that is connected to a single main track. The moveable section can be positioned so that it either connects the right or left subtrack to the main track (the movable section is initially connected to the right subtrack). Two trolleys are hurtling along (one on each subtrack) towards the movable section of track. If a switch is flipped, the left subtrack will be connected to the main track, and the trolley that was traveling down the left subtrack will continue its journey along the main track. If the switch is not flipped, the left trolley will derail, but the trolley that was traveling down the right subtrack will continue onto the main track. Victim is strapped to the main track just beyond the flipping point. As the trolleys are approaching the flipping point, Suzy flips the switch, which takes the left trolley onto the main track; the right trolley derails. The left trolley hits Victim, who is crushed. Had Suzy not flipped, the right trolley would have continued onto the main track and Victim would still have been crushed.

Intuitively, Suzy's flipping, which makes no difference whatsoever to Victim's fate, is not a cause of Victim's crushing. Yet according to (HF), Suzy's flipping *is* a cause. For if we hold fixed the actual fact that the trolley on the right subtrack does derail, Victim's crushing *does* depend on Suzy's flipping the switch, since if she does not do so, the trolley on the left subtrack would derail, and Victim would not be crushed.[5] Hence Two Trolleys is a counterexample to (HF).

The following case is an example of a self-canceling threat:

> Two Assassins: Captain and Assistant are on a mission to kill Victim. On spotting Victim, Captain yells "Fire!" and Assistant shoots at Victim. Victim overhears the order, and although the bullet almost hits him, he ducks just in time and survives

[5] Hall (2000) discusses some potential replies to a similar kind of switching counterexample. In my view, however, these replies are unconvincing; I do not have space to provide arguments here.

> unharmed... If Captain hadn't yelled "Fire!", Assistant would not have shot, and Victim would still have survived. If Victim had not ducked, however, he would have been hit by the bullet, and would not have survived.[6]

We do not intuitively feel that Captain's yelling "Fire!" is a *cause* of Victim's survival, yet holding fixed the fact that Assistant fired, if Captain hadn't yelled "Fire!", Victim would not have ducked, and consequently would not have survived. Hence, Victim's survival depends on Captain's yelling "Fire!", holding fixed Assistant's shooting, and therefore (HF) rules that Captain's yelling "Fire!" is a cause of Victim's survival.[7]

Self-canceling threats have the following structure: *C* introduces some threat to *E*, but at the same time also initiates some countermove that is successful in canceling the threat to *E*, and *E* consequently occurs. In Two Assassins, Captain's yelling "Fire!" poses a threat to Victim's survival, but at the same time, alerts Victim to the threat posed. Victim ducks, thus canceling the threat to his survival.[8]

Note incidentally, that the naïve dependence theory (ND) delivers the intuitively correct theoretical verdicts for Two Trolleys and Two

---

[6] Originally due to McDermott, but extensively discussed by Hitchcock (2003).

[7] Hitchcock (2003:9-11) reports (on the basis of informal surveys) that intuitions are either divided or unclear with regard to whether or not Captain's yelling "Fire!" is a *cause* of Victim's survival. I consider this intuition to be simply mistaken. In my experience, as soon as one reminds respondents that Assistant would not have fired had Captain not yelled "Fire!", they generally reverse their initial judgment. Such mistaken intuitions arise from a failure to take on board the stipulated facts of the case.

[8] One might attempt to reply to this counterexample by arguing that the intuition that Captain's yelling "Fire!" does not cause Victim's survival is mistaken. One could suggest, as Lewis (2004) has done, that *in general* assassination orders do not cause survivals, but that in this *particular* case, the order (the yell) did cause the survival. The mistake, Lewis argues, is a confusion of singular causation with general causation. I do not find this objection convincing; my intuition with regard to this particular case is still firm, even when taking note explicitly of Lewis's warning.

Assassins. Suzy's flipping is not a cause since Victim's crushing doesn't depend on the switching. Similarly, Victim's survival doesn't depend on Captain's yelling "Fire!". As is often the case, introduction of further technical conditions introduces new counterexamples that the simpler theory already dealt with satisfactorily.

To summarize, (ND) falls to cases of early preemption. (L), while delivering the correct verdict for early preemption, delivers the wrong verdict for late preemption. (HF), while delivering the correct verdicts for early and late preemption, introduces new counterexamples involving switches and self-canceling threats. Lastly, attempts to characterize the causal relation in intrinsic or local terms fall to cases of causation by omission.

Given the repeated failures of these initially promising univocal theories, a few philosophers, such as Hall (2004), Godfrey-Smith (forthcoming), Hitchcock (2003), and Cartwright (1999) have recently begun to explore pluralistic approaches to causation. In the next section, I focus on Hall's dualistic theory, which I consider to be the most fully developed and best defended of these approaches.

3. Hall's two concepts of causation

Hall (2004) proposes that there are *two* concepts of causation: production and dependence:

(TC) C is a cause of E if and only if (E depends on C) or (C produces E).[9]

Each disjunct is given a different analysis. Dependence is just counterfactual dependence (though without Lewis's addition of Transitivity). Hall does not attempt a definitive analysis of production, but says that "we evoke it when we say of an event *C* that it helps generate or bring about or produce another event E." Whatever production is, it is a local, intrinsic relation, which, Hall claims, will also

[9] Godfrey-Smith (forthcoming) has also emphasized the distinction between the 'difference-making' and productive aspects of causation.

turn out to be transitive. Hall tentatively advances the hypothesis that the producers of E are those events that are *minimally sufficient* for *E*, in appropriate circumstances, given the laws of nature. Sober (1984) suggests that this sort of productive relation might be usefully analyzed in terms of energy-momentum transfer. Dowe's 'conserved quantity exchange' is an alternative candidate for production. I will not pursue these possibilities here, but will assume that the notion of production is sufficiently intuitive for the purposes of this paper, and that we can recognize it when we see it.

Both disjuncts of (TC) are *sufficient* for *C* to be a cause of E. Note that dependence and production are frequently *co-instantiated*. For example, in paradigmatically causal billiard ball collisions, the motion of the second ball is both produced by the motion of the first *and* depends on it. We might call this relation 'productive dependence.'

It is worth noting that causation, as defined by Hall, is not ambiguous in the sense in which words like 'bat' and 'bank' are ambiguous. In these cases, the two disjuncts (e.g. river bank and savings bank) are generally not co-instantiated in the same particular; their extensions do not overlap: there are no individuals that are both river banks and savings banks. Production and dependence, on the other hand, very often *are* co-instantiated in the same particular, as in the billiard ball case. It appears to be an accident that we use the same word 'bat' for both the nocturnal flying mammal and a piece of sports equipment. The two meanings of word 'bat' are not related in any interesting way, and the two types of bat share few significant properties. 'Bat_1' and 'bat_2' are merely homonyms.[10] In Dutch, two different words are used: 'vleermuis' and 'knuppel' respectively.[11] In the case of causation, however, the production and dependence senses *do* seem to be closely related, and related in interesting ways. For example, they are both able to play similar roles in explanation, prediction, agential control, and so on. It is no accident that the same word 'causation' is used for both production and dependence. Causation exhibits *polysemy* rather than *homonymy*.

[10] Their shared properties are rather uninformative and do not seem central to the meaning of either homonym. For example, they both share the property of being physical objects.

[11] Thanks to an anonymous referee for this point.

The senses are so closely related that only trained philosophers, such as Hall, might think to tease them apart.

How does (TC) fare will regard to the canonical counterexamples to the major univocal theories? I will not attempt to provide a comprehensive survey here and will instead restrict myself to pointing out some of the major advantages of Hall's dualistic theory. (TC) takes care of both early and late preemption with impressive ease. In Trainee and Supervisor, Trainee's shot is a cause of Victim's death because of its local, productive relationship (via Trainee's bullet) with Victim's death, despite the absence of dependence. It is extremely plausible that when making intuitive judgments about preemption, it *is* this local productive relation that we pay attention to. This is a very natural psychological diagnosis of our intuition-forming process. Early and late preemption count as cases of causation *not* in virtue of any dependence of the effect on the cause (while holding fixed the redundant backup) as the counterfactualist would have it; rather, C is a cause of E in virtue of the *productive relation* between the two. (TC) handles our late preemption counterexample Billy and Suzy in exactly the same fashion.

(TC), since it does not need to appeal to the holding fixed strategy in order to deliver the intuitively correct verdicts in cases of preemption, has the great advantage of not thereby ruling in switching and self-canceling threats such as Two Trolleys and Two Assassins respectively. In switching and self-canceling threats, there is no dependence *simpliciter* between the putative cause and effect, and the latent dependencies that would be revealed by holding fixed certain facts *remain* hidden, as we desire.

(TC) also deals straightforwardly with causation by omission counterexamples such as Gardener that beset the physical process theories of Salmon (1984, 1994) and Dowe (1992, 2000). My gardener's not watering my plants caused their death in virtue of the *dependence relation* that links the two, despite the absence of any local physical process linking the two events.[12]

---

[12] Strictly, the death of the plants does not depend on the gardener's not watering them. The plants, being mortal, would have died sooner or later. In order to make the counterexample work, some other detrimental effect on the plants (due to their not being watered) should be chosen as the effect. Alternatively, we could precisify the effect (e.g. the plants' death at time t).

(TC) is thus an important advance on univocal counterfactual theories. In addition to (TC)'s success with recalcitrant counterexamples to univocal theories, Hall provides a more general argument in favor of splitting our concept of causation in two: it enables us to preserve what he takes to be several of our important platitudes about causation: locality, intrinsicness and transitivity. While locality, intrinsicness and transitivity do not apply in cases of omission, they always apply in cases of production.

4. Counterexamples to (TC)

It appears, however, that there are several counterexamples to Hall's theory; there are cases that exhibit neither production nor dependence, but which we intuitively judge to be causation, and cases that exhibit production and/or dependence, which we intuitively judge not to be causation. We therefore have reason to suspect that if causation *is* a non-univocal concept, Hall's non-univocal analysis is not quite the right one.

4.1. Causation with neither production nor dependence

Hall himself provides the following counterexample to (TC):

> Second Escort: Suzy is piloting a bomber on a mission to bomb a particular target. She is escorted on this mission by Billy in a second plane, and Mary in a third plane. Enemy's fighter approaches, intending to shoot down Suzy's bomber. Billy shoots before Enemy does, however, and Enemy's plane goes down in flames. Suzy proceeds to the target and completes the mission. If Billy hadn't shot down Enemy, Mary would have. (Hall, 2004).

Billy's action prevents Enemy from preventing Suzy's bombing (which Hall calls 'double prevention'). Even though no local process connects Billy's shooting with Suzy's Bombing, Hall claims that Billy's shooting

is intuitively a cause of the bombing.[13] This example is an instance of 'preempted double prevention'; Billy's prevention of Enemy's attempted prevention preempts Mary's prevention of it. In virtue of the lack of locality, there is no production, and, in addition, the introduction of the backup preventer Mary breaks the dependence between Billy's shooting and the bombing. Hence we have causation with neither production nor dependence, and Second Escort is therefore a counterexample to (TC). Hall leaves this type of example as important 'unfinished business' for his account.

Second Escort illustrates an important general point. If we make just *one* link in a transitive causal chain non-local, there will be no productive relation between C and E. By adding in a redundant backup, we can also remove any dependence. We can thus generate counterexamples to (TC) at will. Here is a similar counterexample:

> Victim's Plants: Trainee shoots Victim, who bleeds to death. If Trainee hadn't shot Victim, Supervisor would have done. Victim, having bled to death at the hands of Trainee, is now unable to water his plants, which subsequently die.

The death of the plants does not depend on Trainee's shot, since Supervisor would have shot Victim had Trainee not done so. There is no productive relation between Trainee's shot and the death of the plants either, since there is no local connection between them. Yet Trainee's shooting seems intuitively to be a cause of the plants' death.

It seems plausible that what we are doing psychologically when we make our intuitive judgments in these cases is the following: we naturally break the cases down into two discrete steps. In Second Escort, the first step consists of Billy's shooting down Enemy's plane. The second step consists of the absence of Enemy's attack on Suzy and the subsequent bombing. Intuitively, each of these constituent steps is clearly causal. We then implicitly link these two causal steps together in a chain and conclude that Billy's shooting was a cause of the bombing. We can tell a

---

[13] I must confess that I am not entirely confident of Hall's intuition regarding this case. If one is skeptical, the force of this counterexample is reduced. There are, however, several other clear counterexamples of similar form such as Victim's Plants.

similar story for Victim's Plants. This may seem a plausible account of our judgment processes in these cases, but it will not do as an analysis. Transitivity is *false*; there are counterexamples in which C causes D, and D causes E, but C is *not* a cause of E. In Two Trolleys, for example, Suzy's switching causes the left trolley to travel down the main track, and that trolley's travelling down the main track causes Victim's crushing. Yet we do not want to say that Suzy's switching causes Victim's crushing. Hence Transitivity fails.

Similarly, in Two Assassins, Captain's yelling "Fire!" causes Victim to duck, and his ducking causes him to survive. Hence, according to Transitivity, Captain's yelling "Fire!" causes Victim to survive. Again, this verdict is highly counterintuitive.

Moreover, even if there were some way of rescuing Transitivity, we could easily generate related counterexamples in which we could not appeal to this chaining strategy. This could be done by starting from an ordinary case of early preemption such as Trainee and Supervisor, and making the productive link non-local. For instance:

> Action at a Distance Guns: Trainee and Supervisor are armed with action-at-a-distance guns. Trainee shoots first and Victim vaporizes. If Trainee hadn't shot, Supervisor would have, and Victim would have been vaporized in exactly the same manner.

In this case there is neither production nor dependence, yet our intuition that Trainee's shooting is the cause of Victim's vaporization remains solid. Neither can we point to a chain of intuitively causal links.

A second general method of generating such counterexamples is to begin with an omission, and add in a redundant backup omission:

> Patricidal Brothers: Jack and Bobby are tired of waiting to inherit their father Joe's money and independently decide to do away with him. They both decide that the best way to kill Joe is to withhold his medication. Joe must take two pills every day (one red, one green) in order to keep him alive. Every evening, before going to bed, Jack leaves a red pill on the kitchen table for Joe, and Bobby leaves a green pill. One evening, Jack, unable to wait any longer for his inheritance, decides not to leave his red pill on the table, and retires for the evening. A few minutes later, Bobby, who has decided on the same course of action, notices that his brother Jack

> has not left his pill on the table. Bobby, not wanting to risk being incriminated for his father's death, leaves his green pill on the table as usual. But if Jack had left his red pill on the table, Bobby, wanting to guarantee Joe's demise, would not have left his green pill. Joe, deprived of his full dosage, dies shortly thereafter.[14]

It is perfectly clear that Jack's omission is a cause of Joe's death. Yet there is no productive relation between these two events. There is also no dependence, since Jack's omission merely preempted Bobby's omission. If Jack had left the red pill, Bobby would have withheld the green pill, and Joe would still have died. This case is interesting in that it challenges Hall's hunch that "there could be nothing more to causation by omission than counterfactual dependence."[15] In virtue of what then, does Jack's omission count as a cause, if not dependence? One is initially tempted to answer that it is counterfactual dependence, but holding fixed the fact that Bobby *did* deliver the green pill. Note that (HF) would also deliver the intuitively correct verdicts for Second Escort, Victim's Plants, Action at a Distance Guns. If we hold fixed the fact that Mary (the Second Escort) doesn't shoot, then the Bombing depends on Billy's shooting; if we hold fixed the fact that Supervisor doesn't shoot, the death of Victim's plants depends on Trainee's shooting; if we hold fixed the fact that Supervisor doesn't shoot his action-at-a-distance gun, Victim's being vaporized depends on Trainee's shooting. Yet we have seen that (HF) drags in switching and self-canceling threats as *bona fide* types of causation (e.g. Two Trolleys and Two Assassins). So where does this leave us? It appears that there is *something* right about (HF), yet as a univocal theory, it falls frustratingly short of universality.

4.2. Production and/or dependence without causation

Two Trolleys, as we have seen is a counterexample to Transitivity; Suzy's flipping of the switch is not a cause of Victim's crushing. However, it seems clear that there is an intrinsic, spatiotemporally

---

[14] Cases with an analogous structure can be constructed in which an individual refuses to vote for a certain proposition for which unanimity is required.

[15] Hall (2004).

continuous relation between Suzy's flipping and Victim's crushing, corresponding to the motion of the trolley from switch to Victim. Hence Two Trolleys is a counterexample to the sufficiency of production for causation. Note also that this example indicates that production (conceived of as a local, intrinsic relation) is not transitive, contrary to Hall's claim. Suzy's flipping causes the left trolley to travel down the main track, and this trolley's traveling down the main track causes Victim's crushing, yet Suzy's flipping is intuitively not a cause of Victim's crushing.

Dependence does not seem to be sufficient for causation either. Consider:

> Queen Elizabeth: My plants died when I was away on vacation. If Queen Elizabeth had watered them, they would not have died.

Clearly Queen Elizabeth's failure to water the plants is not a cause of their death. Yet if she had watered the plants, they would not have died. Hence the plants' death does depend counterfactually on whether or not the Queen waters them. Woodward (2003) has suggested that we do not judge Queen Elizabeth's omission to be a cause of the plants' death because we do not take it to be a *serious possibility* that (counterfactually) she would have watered them. This seems too strong, however. Consider the case of the chronically unreliable gardener, who has *never* remembered to water my plants. We would still want to say that his failure to water my plants was a cause of their death, even though, after a certain point, we would no longer take serious the possibility of his watering them.[16]

5. Causal judgments, moral facts and responsibility

Perhaps we judge the unreliable gardener's failure to water the plants to be a cause of their death because we consider him to be *morally to blame* for their death, perhaps in virtue of having violated some gardener's

[16] Beebee (2004) makes a similar point.

contract, for example. The idea that causal facts might be judged partly on the basis of moral facts seems to be supported by examples such as:

> Automobile Accident: Billy is driving steadily down a deserted highway when suddenly, without warning, a truck ploughs into the side of his car. It is later revealed that the driver of the truck was heavily intoxicated and had run a red light.

Was Billy's driving steadily down the deserted highway a cause of the accident? One instinctively replies 'no.' (Note that this example demonstrates that Hall's production and dependence are *not even jointly sufficient* for causation). If we change the example and replace the car and the truck by billiard balls, however, we *would* say that each ball's motion was a cause of the collision. Hence our causal judgments appear to be sensitive to the presence or absence of moral agents. It might be objected that billiard ball variant merely indicates that we have conflated causation with moral responsibility in Automobile Accident: that although Billy is not *morally* responsible for the accident, his driving steadily down the highway *is* a cause of the accident. Perhaps this is the correct judgment, although the intuition that Billy's driving down the highway is *not* a cause of the accident does appear to be quite robust. Why not, therefore, take this example to indicate that (human) causation is in part a moral concept?

Others have also proposed that our causal intuitions are sensitive to considerations of moral responsibility. For example, psychologists have found empirically that our causal judgments are influenced by moral facts (Alicke, 1992; Knobe, MS). And for certain kinds of omission, Beebee (2004) has argued convincingly that moral responsibility plays a role in our causal judgments. Obviously causation cannot involve moral responsibility if moral agents (humans) are not involved. For example:

> Automatic Watering Machine: My plants died when I was away on vacation. If my automatic watering machine hadn't broken down, they would not have died.

In this example, it is perfectly acceptable to say that the machine's breaking down is a cause of the plants' death, yet we cannot attribute any *moral* responsibility to the machine (at least not in the strict sense). Note,

however, that it does seem acceptable to claim that the machine's breaking down was *to blame* for the plants' death, at least in some loose sense. On this subject, Hitchcock (MS) has recently written:

> [I]n fault analysis in engineering, or in performing an autopsy, one is trying to discover which component of a complex system is *responsible* for the failure of that system to function; this type of responsibility is not literally moral, but is broadly normative in character. Given this role, it is not altogether surprising that our judgments of token causation are influenced by normative considerations.[17]

Theories of causation that attempt to fit causation into an objective metaphysics will struggle to deliver the intuitively correct results in cases where judgments of (moral) responsibility seem influential. It is hard to see what objective metaphysical difference there could be, for example, between my gardener's failure to water my plants and the Queen's. Hence it would seem that theories with this goal will have to be revisionary to some extent.

The examples discussed in this paper are summarized in the table below. For each case, the table displays whether or not the putative cause in question is intuitively a genuine cause (Int), whether or not the putative cause produces its putative effect (Prod), whether or not the putative effect depends on its putative cause (Dep), and the respective theoretical verdicts of (TC) and (HF). In addition, whether the putative cause is responsible (Resp) for the relevant effect, in the sense described above, is also presented. Hall's (TC) only delivers intuitively correct theoretical verdicts for Trainee and Supervisor, Billy and Suzy, Gardener, Two Assassins, Unreliable Gardener and Automatic Watering Machine. The other examples are all counterexamples to (TC). Counterexamples to the various theories are indicated by asterisks.

[17] My italics.

Ex	Int	Prod	Dep	(TC)	(HF)	Resp
Trainee & Supervisor	Yes	Yes	No*	Yes	Yes	Yes
Billy & Suzy	Yes	Yes	No*	Yes	Yes	Yes
Gardener	Yes	No*	Yes	Yes	Yes	Yes
Two Trolleys	No	Yes*	No	Yes*	Yes*	No
Two Assassins	No	No	No	No	Yes*	No
Second Escort	Yes	No*	No*	No*	Yes	Yes
Victim's Plants	Yes	No*	No*	No*	Yes	Yes
Action at a Distance Guns	Yes	No*	No*	No*	Yes	Yes
Patricidal Brothers	Yes	No*	No*	No*	Yes	Yes
Queen Elizabeth	No	No	Yes*	Yes*	Yes*	No
Unreliable Gardener	Yes	No*	Yes	Yes	Yes	Yes
Automatic Watering Machine	Yes	No*	Yes	Yes	Yes	Yes
Automobile Accident	No	Yes*	Yes*	Yes*	Yes*	No

Given the failure of (ND), (L), (HF), local process theories and (TC), what further options are available to the causal analyst? One might try to retain (TC) and attempt to find alternative means of ruling out Two Trolleys, Queen Elizabeth and Automobile Accident, and ruling *in* the *bona fide* cases of causation that display neither production nor dependence (Second Escort, Victim's Plants, Action at a Distance Guns and Patricidal Brothers). But it is not at all obvious how one might go about doing this.

A second option would be to try to persist with (HF) and to try to rule out switching, self-canceling threats, Queen Elizabeth and Automobile Accident as genuine cases of causation. This option would obviously lead us away from pluralism, however, and back towards a univocal counterfactual analysis. One might try to go the Lewisian route and argue that ordinary intuitions are simply mistaken in all of these cases; but that seems like an uphill battle.

Third, notice from the rightmost column of the table above that responsibility is, *by itself*, sufficient to distinguish between the genuinely causal examples above, and the non-causal. Might then the following 'analysis' of causation be satisfactory?

(R) C is a cause of E iff C is responsible for E.

Is (R), finally, an analysis of causation that evades all of the canonical counterexamples presented above? Or have we 'cheated' in some way? Is responsibility *really* a suitable primitive on which to base an analysis of causation? Or is responsibility too closely synonymous with causation to provide any real illumination of what causation is? Perhaps all we have really done when asking "Is C responsible for *E*?" is ask "Is *C* a cause of *E*?" *Moral* responsibility is clearly distinct from the broader normative notion of responsibility, and from causation: one can be responsible for setting off a booby-trapped bomb (and can cause it to explode) without being *morally* responsible for its exploding. But it is not clear that such a weakening of the notion of moral responsibility to responsibility *simpliciter* (as we would *need* to do to all for non-human cases of causation such as Automatic Watering Machine) leaves us with anything more than a mere synonym for causation. Indeed, it is hard to think of any cases in which *C* can be responsible for *E* without C causing *E*, and *vice-versa*. To a degree, of course, this is what we want from an analysis.

But one suspects that responsibility just falls into the same category as very near-synonyms for causation such as 'bringing about', and consequently does not really provide an illuminating reductive analysis of the concept.

6. Conclusion

Hall's pluralistic theory of causation appears to be a significant improvement on univocal theories of causation, handling several canonical counterexamples with ease. There are, however, several clear counterexamples to Hall's theory. In addition, our causal judgments appear to be sensitive to considerations of moral responsibility, and it does not seem likely that objective metaphysical theories of causation will ever accord with our intuitions in such cases. Such theories will therefore need to be somewhat revisionary. Lastly, the notion of responsibility is considered, but rejected, as an illuminating primitive for analyzing causation, since it appears to provide only an unenlightening synonym.

University of Birmingham
f.longworth@bham.ac.uk
Ohio University
longwort@ohio.du

REFERENCES

Alicke, M. (1992), "Culpable Causation," *Journal of Personality and Social Psychology*, 63, pp. 368-378.

Beebee, H. (2004), "Causing and Nothingness," in: Collins, J., Hall, N. and Paul, L. (eds.), *Causation and Counterfactuals* (pp. 291-308), Cambridge (MA): MIT Press.

Cartwright, Nancy (1999), *The Dappled World*, Oxford: Oxford University Press.

Dowe, P. (1992), "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science,* 59, pp. 195-216.

Dowe, P. (2000), *Physical Causation*, New York: Cambridge University Press.

Fair, D. (1979), "Causation and the Flow of Energy," *Erkenntnis,* 14, pp. 219-250.

Godfrey-Smith, P. (forthcoming), "Causal Pluralism," in: Hitchcock, C., Beebee, H. and Menzies, P. (eds.), *Oxford Handbook of Causation*.

Hall, N. (2000) "Causation and the Price of Transitivity," *Journal of Philosophy,* 97, pp. 198-222.

Hall, N. (2004) "Two Concepts of Causation," in: Collins, J., Hall, N., and Paul, L.A. (eds.), *Causation and Counterfactuals* (pp. 225-276), Cambridge: MIT Press.

Halpern, J. and Pearl, J. (2001), "Causes and Explanations: A Structural-model Approach — Part I: Causes," *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 194-202), San Francisco: Morgan Kaufmann.

Halpern, J. and Pearl, J. (2005), "Causes and Explanations: A Structural-model Approach — Part I: Causes (expanded version)," *British Journal for the Philosophy of Science*, 56, pp. 843-887.

Hitchcock, C. (2001), "The Intransitivity of Causation Revealed in Equations and Graphs," *Journal of Philosophy,* 98, pp. 273-299.

Hitchcock, C. (2003), "Of Humean Bondage," *British Journal for the Philosophy of Science,* 54, pp. 1-25.

Hitchcock, C. (MS), "Prevention, Preemption, and the Principle of Sufficient Reason."

Hume, D. (1902), *An Enquiry Concerning Human Understanding*, ed. L.A. Selby-Bigge, Oxford: Clarendon Press. First published 1748.

Knobe, J. (MS), "Attribution and Normativity: A Problem in the Philosophy of Social Psychology."

Lewis, D. (1973), "Causation," *Journal of Philosophy,* 70, pp. 556-567. Reprinted in: Lewis (1986b), pp. 159-172.

Lewis, D. (1986a), "Postscripts to 'Causation'," in: Lewis (1986b), pp. 159-213.

Lewis, D. (1986b), *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

Lewis, D. (2000), "Causation as Influence," *Journal of Philosophy,* 97, pp. 182-197. Expanded version appears in Collins, J., Hall, N. and Paul, L.A. (eds.) (2004), *Causation and Counterfactuals* (pp. 75–106), Cambridge (MA): MIT Press.

Longworth, F. (2006), *Causation, Counterfactual Dependence and Pluralism*, Ph.D. Dissertation, University of Pittsburgh.

Menzies, P. (2001), "Counterfactual Theories of Causation," in: *Stanford Encyclopedia of Philosophy*.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Salmon, W.C. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Salmon, W. (1994), "Causality Without Counterfactuals," *Philosophy of Science,* 61, pp. 297-312.

Sober E. (1984), "Two Concepts of Cause," *PSA 1984*, vol. 2, pp. 405-424.

Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

Yablo, S. (2002), "De Facto Dependence," *Journal of Philosophy,* 99, pp. 130-148.

Yablo, S. (2004), "Advertisement for a Sketch of an Outline of a Prototheory of Causation," in: Collins, J., Hall, N., and Paul, L.A. (eds.), *Causation and Counterfactuals* (pp. 119-137), Cambridge: MIT Press.