The metaphysics of causal models: Where's the *biff*?

Toby Handfield, Charles R. Twardy,
Kevin B. Korb, and Graham Oppy

June 2007

Abstract

This paper presents an attempt to integrate theories of causal processes – of the kind developed by Wesley Salmon and Phil Dowe – into a theory of causal models using Bayesian networks. We suggest that arcs in causal models must correspond to possible causal processes. Moreover, we suggest that when processes are rendered physically impossible by what occurs on distinct paths, the original model must be restricted by removing the relevant arc. These two techniques suffice to explain cases of late preëmption and other cases that have proved problematic for causal models.

1 Causal models and the problem of preëmption

An enduring difficulty for counterfactual analyses of causation is the phenomenon of late preëmption. A well-known, simple example is that of Billy and Suzy, each throwing a rock at a bottle. Both children are highly accurate, but Suzy throws faster. Suzy's rock smashes into the bottle, which then breaks, and Billy's rock passes through thin air as the shards of glass settle on the ground. Had Suzy not thrown, the bottle would still have smashed, so the counterfactual account suggests that Suzy's throw is not a cause of the bottle-smashing.[1]

This is a pre-print of a paper forthcoming in *Erkenntnis*.

[1] More sophisticated counterfactual accounts, such as David Lewis's influence theory (2000), do manage to account for cases such as Billy and Suzy, but this success appears to be in virtue of closely tracking the *symptoms* of causation, rather than the causal relation itself. Consequently, the account remains vulnerable to counterexamples (Schaffer 2001).

In sharp contrast to the counterfactual approach to causation, some theorists have thought that causation is at bottom a matter of a concrete relation – a physical process – between events. While no-one knows exactly what the relation is, it has come to be referred to as *biff*.[2]

Biff-theorists can easily distinguish between Billy's throw and Suzy's as causes of the bottle's smashing. Suzy's throw, but not Billy's, is connected to the event of the bottle-smashing by a physical process – the flying rock. Surely this is the crucial difference that leads us to conclude that Billy's throw is not a cause, while Suzy's is.

A naïve process theory, however, faces two serious objections. First, it massively overgenerates causes. There are physical processes everywhere. Between Billy's throw and the bottle smashing, for instance, light is transmitted. Why, then, is Billy's throw not a cause?

Secondly, process accounts struggle to account for causal prevention or for causation by absences. In causal statements such as "The driver's failure to brake caused the accident", it is difficult to identify a concrete event which is the driver's failure to brake. And without such a concrete event, it is difficult to see how the failure to brake could be connected to the accident by a concrete process.

Neither of these objections is an issue for a counterfactual account: first, because the light transmitted between Billy's throw and the bottle smash does not sustain any counterfactual dependence between these events; and second, because counterfactual dependence is not a concrete relation, hence there is no need for concrete causal relata.

In this paper, we use the framework of causal models, developed by Judea Pearl and others, in an effort to integrate the insights of the counterfactual and process-based theories of causation. In particular, we aim to show that supplementing the broadly counterfactual approach of causal models with some of the apparatus of a biff theory gives a promising way to deal with the problem of late preëmption.

Causal models incorporate insights from a number of earlier approaches to causation, including the counterfactual approach.[3] Each causal model is a set of variables and a set of structural equations involving those variables. It is frequently convenient to represent a model by means of a directed graph, where each node represents a variable, and the arcs represent the existence of relations between the

[2] It is not clear who coined the term, but it is strongly associated with D. M. Armstrong. The first occurrence in print of which we are aware is Lewis 2004. Leading biff theorists have been David Fair (1979), Wesley Salmon (1984), and Phil Dowe (2000).

[3] Causal models have been elegantly explained for a philosophical audience by Christopher Hitchcock (2001). Here we give only a very brief review of models, and direct unfamiliar readers to Hitchcock's paper.

variables, as given by the structural equations.

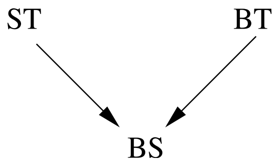
A variable in a model represents a factor with two or more possible states, such as the pressing of a button and the failure to press that same button. One great attraction of causal models for the purposes of our project is their seeming neutrality on the nature of the causal relata. The formal structure of a causal model does not require a variable to represent events, event-types, propositions, states-of-affairs, or any other particular members of the ontological zoo. Causal models, then, provide us with a blank screen onto which we can project our metaphysical claims.

The structural equations of a causal model give the counterfactual (probabilistic) dependencies between the factors. By definition, the variables which feature in the equation for a given variable are all and only the parents of that variable in the graph. Thus the graph of a causal model represents in an economical, visual fashion the pattern of dependencies between all the factors.

For the purposes of this paper, we restrict our attention to deterministic models and use only binary variables. As a heuristic device we will use "true" and "false" as the values for those variables, and operations from sentential logic in specifying the structural equations. It must be stressed, however, that the variables taking those values do not represent truth values directly. Rather, they represent propositions, facts, states-of-affairs, or suchlike entities. "$JoeWins = true$", for instance, represents the proposition that Joe wins.

Causal models appear to give us the means to analyse token causal relations between individual events. The advocates of such analyses have tackled an impressive range of examples that have proven difficult for other accounts to handle, and have offered intuitively plausible solutions. One reason to doubt the legitimacy of causal models as a means of understanding causation, however, is that those who advocate such techniques admit that there is no complete method for determining the correct model of a given situation (Halpern and Pearl 2005: 878; Hiddleston 2005: n. 16). Moreover, a different choice of model will often result in a different answer to the question: Do these events stand in a causal relationship? Thus, without a clear rationale for any particular model choice, causal models give only equivocal answers to questions about token causation.

Causal modellers are happy to endorse some degree of pragmatic context-sensitivity for causal claims, and hope to show that there are correspondingly context-sensitive reasons for choosing a particular model (Hitchcock 2001: 287, 294–5). Others suggest that contextual factors be built directly into the analysis of causation within a model, rather than used to guide model choice (Hiddleston



$$\begin{aligned} \mathit{SuzyThrows} &= 1 \\ \mathit{BillyThrows} &= 1 \\ \mathit{BottleSmashes} &= \mathit{SuzyThrows} \vee \mathit{BillyThrows} \end{aligned}$$

Figure 1: *The simple model for Billy and Suzy.*

2005; Menzies 2004b). We suspect that both context-sensitive model-building *and* a context-sensitive analysis will ultimately be required. But until those context-sensitive reasons are made fully explicit, it remains unclear what insight has been added by the framework of causal models. At worst, we might suspect that modellers have merely provided a very flexible framework in which to re-cast our original intuitions about causation. Perhaps we can model any situation in such a way as to conserve our original intuitions, but only with the assistance of *ad hoc* decisions as to what model to use. If so, the development of causal models is no great philosophical achievement.

How then, do causal modellers handle late preëmption cases?

Example 1 (Billy and Suzy) *Billy and Suzy each throw a rock at a bottle. Both are highly accurate. Suzy's gets there first, and the bottle smashes before Billy's has time to hit. Had Suzy not thrown, however, Billy's would have hit.*

A straightforward model of Billy and Suzy is shown in Figure 1.

The naïve counterfactual dependence account wrongly entails that Suzy did not cause the smash, since had Suzy not thrown, the bottle would have smashed anyway. The graph-theoretic accounts of token causation given by Halpern and Pearl, and Hitchcock, however, applied to the simple causal model above, manage to avoid this conclusion in exchange for another: they entail that both throws caused the smash! This is most easily demonstrated by way of Hitchcock's account.[4]

For Hitchcock (2001: §5), $A = a$ causes $B = b$ in a model M – consisting of a set of variables V and a set of structural equations E – just in case (i) $A = a$ and $B = b$, and (ii) there is a weakly active path from A to B in M. For ease of understanding,

[4] We here ignore the subtle differences between Halpern and Pearl's and Hitchcock's analyses of token causation in causal models. See Korb, Twardy, Handfield, and Oppy 2005 for a more thoroughgoing attempt to analyse token causation in terms of processes. Here we aim simply to illustrate the benefits of incorporating the metaphysics of causal processes into our theory of modelling.

we first introduce the related notion of an *active path*, before proceeding to define a weakly active path.

A path ϕ from X to Z is ACTIVE in a model M iff Z depends counterfactually upon X within the new system of equations E' constructed from E as follows:

> for all $Y \in V$, if Y is intermediate between X and Z, but does not belong to the path ϕ, then replace the equation for Y with a new equation that sets Y equal to its actual value in E. (If there are no intermediate variables that do not belong to this path, then E' is just E.)

Very roughly, the active path analysis says: Holding fixed what happens on other causal paths, is X a difference-maker for Z along some path ϕ? If so, then $X = x$ is a cause of $Z = z$.

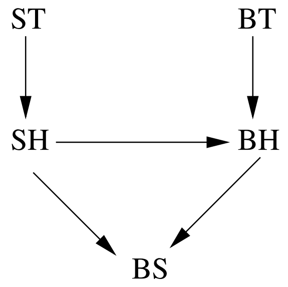
If we settled for the active path analysis of causation, we would get the conclusion that neither Billy nor Suzy cause the bottle to smash, because neither the path $\langle \mathit{BillyThrows}, \mathit{BottleSmashes} \rangle$ nor the path $\langle \mathit{SuzyThrows}, \mathit{BottleSmashes} \rangle$ is active.

To deal with some other problem cases, Hitchcock recommends a weaker analysis in terms of weakly active paths. To introduce the notion of a weakly active path, we need to define a *redundancy range* for a path (2001: §7). In some cases, the values on a path are counterfactually insensitive to various settings of other variables in the model. For instance some children are attempting to ruin a newspaper by making it wet. One child throws a bucket of water on the newspaper, while the other puts it into the bath. The model is simply: $\mathit{Newspaper} = \mathit{Bucket} \vee \mathit{Bath}$. Given $\mathit{Bucket} = \mathit{true}$ and $\mathit{Bath} = \mathit{true}$, it is irrelevant to the path $\langle \mathit{Bucket}, \mathit{Newspaper} \rangle$ whether or not the newspaper is dunked in the bath. Even if the newspaper had not been placed in the bath, it would still have been true that the bucket was tipped on the newspaper and that the newspaper is wet. Thus the values true and false for the *Bath* variable are both in the *redundancy range* for the instantiated path $\langle \mathit{Bucket} = \mathit{true}, \mathit{Newspaper} = \mathit{true} \rangle$. The values of all the variables on that path would remain unchanged, regardless of whether *Bath* is true or false.

A path is weakly active, then, just in case, for some redundancy-range setting of the variables elsewhere in the model, the path is active.

Relative to the instantiated path $\langle \mathit{SuzyThrows} = \mathit{true}, \mathit{BottleSmashes} = \mathit{true} \rangle$, $\mathit{BillyThrows} = \mathit{false}$ is in the redundancy range: it makes no difference whether or not Billy throws. And for the case where $\mathit{BillyThrows} = \mathit{false}$, the path from Suzy's throwing to the bottle smashing is indeed active. Therefore the path $\langle \mathit{SuzyThrows}, \mathit{BottleSmashes} \rangle$ is weakly active.

But precisely the same can be said for the path $\langle \mathit{BillyThrows}, \mathit{BottleSmashes} \rangle$. There is a value of *SuzyThrows* in the redundancy range (i.e. Suzy does not throw)



$$\begin{aligned} \mathit{SuzyThrows} &= 1 \\ \mathit{BillyThrows} &= 1 \\ \mathit{SuzyHits} &= \mathit{SuzyThrows} \\ \mathit{BillyHits} &= \mathit{BillyThrows} \wedge \neg \mathit{SuzyHits} \\ \mathit{BottleSmashes} &= \mathit{SuzyHits} \vee \mathit{BillyHits} \end{aligned}$$

Figure 2: *Billy and Suzy with additional variables to represent Billy's and Suzy's hitting the bottle. This model gets the correct result on Halpern and Pearl's analysis: Suzy's throw is a cause; Billy's is not.*

such that Billy's throwing makes a difference (given Suzy is not throwing, Billy's throwing makes a difference). Therefore, Billy's throwing is – incorrectly – deemed a cause of the bottle smashing.

Halpern and Pearl argue that this problem is circumvented by adopting a different model, with an additional variable along each of the paths (see Figure 2). And indeed, they believe there is a general rule about modelling which they claim needs to be applied here.

Preemption modelling rule "If we want to argue in a case of preëmption that *c* is a cause of *e* rather than *d*, then there must be a random variable ... that takes on different values depending on whether *c* or *d* is the actual cause." (2005: 862)

This rule might be correct, but it is not an appropriate rule for a theory of model choice. In the first place, the rule requires us to have some pre-theoretical intuition as to whether or not the case at hand is one of preëmption: surely a causal intuition. Secondly, the rule requires us to be able to distinguish cases on the basis of *causal* differences between them. If we are trying to construct a model to give us an answer to the question: "Is *c* or *d* the cause of *e*?", then this rule will give us no advice until we have already answered that question by independent means.

Note that we are not making the (unreasonable) demand of Halpern and Pearl that they provide a *reductive* analysis of causation; one which at no point makes use of causal concepts in the *analysans*. Indeed, there is some reason to doubt that this is possible for anyone who adopts the methods of causal modelling. Causal model accounts require evaluations of counterfactuals about possible *interventions*, and it appears likely that interventions can be characterised only in causal terms.[5] But that does not prevent the theory of causation thus offered from being illuminating. The objection in the present instance is that the causal information required to apply the rule is too intimately related to the information that the analysis is supposed to supply. The circle of analysis is too tight. That vitiates the philosophical merits of the analysis, and at least threatens vicious circularity.

Halpern and Pearl do not seem entirely satisfied with this account anyway, and offer an alternative that requires distinguishing different bottle-smashing events on the basis of their time of occurrence (2005: 863–4). A bottle-smashing caused by Suzy at t has a causal influence on Billy's ability to hit the bottle at $t + 1$. This method of distinguishing between the events straightforwardly makes the smashing depend on Suzy's throw, because had Suzy not thrown, the event of «bottle-smashing-at-t» would not have happened at all.

But this more fine-grained approach to modelling the situation is – we believe – not a happy method of solving the problem. While we agree with Hitchcock that model choice will be in part governed by pragmatic considerations, and correspondingly there may be cases where we have a particular interest in the time at which the bottle shatters, Halpern and Pearl's introduction of this feature to the current case seems insufficiently motivated and *ad hoc*. *Prima facie*, there is nothing wrong with the model given in Figure 1, and the only justification offered by Halpern and Pearl for rejecting it is by appeal to a suspiciously circular rule.

Hitchcock (2001: 287) recommends three rules to govern the construction of causal models.

1. The equations must entail no false counterfactuals.

2. The equations must not represent counterfactual dependence relations between events that are not distinct.

3. We should not include variables whose values correspond to possibilities that we consider too remote.

But in the current instance, the model given in Figure 1 offends against none of these rules. So neither Hitchcock nor Halpern and Pearl have provided a sufficient rationale for the adoption of a more sophisticated model such as that in Figure 2.[6]

We believe that by paying attention to the underlying causal processes which are the basis of counterfactual dependence, we can select better causal models.

[5] See Woodward and Hitchcock 2003 and Korb et al. 2005: 5.

[6] Hitchcock himself does not offer a model for this case, but implies that he too would use a model like that given in Figure 2.

Indeed, we will show that at least some alleged problem cases for causal modellers are easily remedied, given our integration of process metaphysics into the task of causal modelling.

Our suggestions, however, are just a beginning. By no means will we offer an algorithm which decisively rules on correct model choice for every situation. We hope merely to lend plausibility to the claim that there are principled means of deciding whether a given model is apt. And consequently, we will indirectly be supporting the claim that causal models offer a genuine insight into the nature of causation.

1.1 A sceptical challenge

Before putting forward our positive proposal, we should consider an objection which has been made to a similar attempt at an integrated analysis, drawing upon both the process theory and counterfactual dependence theories.

It has been suggested that, ultimately, all causal facts *supervene* upon facts about a fundamental physical relation – the aforementioned biff. This supervenience claim pays heed to the idea that causation involves a concrete relation insofar as it suggests that we ought, ultimately, to be looking for instances of biff. But this does not entail that all causal relations are a matter of one event biffing another. Sometimes it is the *absence* of any biff relations that constitutes a causal relation between facts.

This suggestion has been made by Peter Menzies (1999) and subsequently has been criticised by David Lewis (2004). Lewis thinks successful conceptual analysis of causation will pick out an *a priori* concept of causation. According to Lewis, it is conceivable that causation is not subserved by biff at all (283–4). He therefore suggests that it is an *a posteriori*, contingent discovery that physical processes are the supervenience basis of causation, and that it could have been otherwise. Moreover, he points out, if you try to concoct a counterfactual definition in terms of biff-relations, you get a horribly disjunctive definition, as follows:

A causes *B* just in case:

1. *A* biffs *B*, *or*
2. Had *A* not occurred, A' would have biffed B', where B' is incompatible with *B*, *or*
3. Had *A* not occurred, A' would have biffed C', where C' is incompatible with *C*, and *C* biffs *B*, *or*

4. *A* biffs *C*, and had *C* not occurred, *C′* would have biffed *B′*, *or*

5. . . .

And this infinite disjunction does not seem to map our cognitive behaviour (285–6). When we entertain the thought that *A* causes *B*, we do not entertain a long list of biff-mediated connections that might obtain between *A* and *B*; we simply entertain a unified causal concept. Lewis thinks this concept should be given simply in terms of counterfactual dependence between events, without reference to biff.

We reply that: First, we are sceptical about the claim that, possibly, causation is not underwritten by biff. We think it arguable that it is an *a posteriori* necessity that causation supervenes upon biff. Lewis's claim that we can conceive of biff-less causation is, we might argue, a mistaken pseudo-conception, much like the alchemist who, upon discovering hydrogen, thinks that it might be a compound of water and air. We now know that the alchemist's hypothesis was not merely wrong, it was a metaphysical *impossibility*.

Secondly, our own integrated analysis, unlike the one suggested above, does not suffer from disjunctivitis. It is a suitably unified concept of causation. But this remains to be shown.

2 Processes underlying causal models

Our guiding idea is that each arc of a model must correspond to a salient *possible* connecting process. Models which fail to meet this criterion are in some way defective. Remedying this sort of defect will – we argue – give a partial solution to the problem of late preëmption.

Putting the requirement of salience to one side for the moment, our master rule for causal models is then:

> For every arc $A \to B$, there must be a value of *A* in a possible state of the model such that a connecting process can exist between an *A*-grounding event and a *B*-grounding event.

In order to unpack the meaning of this rule, we shall need to explicate the concepts of: (*i*) a possible state of a model, (*ii*) an *X*-grounding event for a variable *X*, and (*iii*) a connecting process between two such events.

A POSSIBLE STATE of a model is simply a state represented by an assignment of values to the variables which satisfies the structural equations of that model. The remaining two concepts are explained below.

2.1 Connecting processes

By a connecting process, we mean something like biff. That is, something like Wesley Salmon's (1984) or Phil Dowe's (2000) accounts of a causal process: a physical four-dimensional entity that has some distinguishing feature which makes it apt to be a means by which causal influence can be propagated. For instance, Salmon suggested that a process must be capable of transmitting a mark to be causal. Dowe's definition of a causal process is: a worldline that possesses a conserved quantity – a quantity governed by a conservation law.

Processes stand in very intimate relations to events. One might think that processes are *composed* of events – that is, have events as parts. Alternatively, you might reject a part–whole analysis of events and processes, while acknowledging that events in some fashion *constitute* processes. We shall remain uncommitted on this point, and simply refer to events as "constituents" of processes without implying that events are not simply parts of processes.

A straightfoward hypothesis then suggests itself as to how events might be connected by a process: Two events are CONNECTED by a process P if and only if they are both constituents of P.

Consider a simple scenario which illustrates this variety of connection between events.

Example 2 (Suzy alone) *Suzy throws a rock at a bottle. The rock hits, and the bottle smashes.*

The events, «Suzy throwing at t» and «the bottle smashing at $t + 1$» are both constituents of a process. That process is simply the four-dimensional spacetime worm of the flying rock, from throw to collision.

An obvious model for this scenario is simply two binary variables, one of which represents whether or not Suzy throws, and the other represents whether or not the bottle smashes. The graph is simply: *SuzyThrows* → *BottleSmashes*, with structural equation *BottleSmashes* = *SuzyThrows*. We have not yet defined a grounding event for a variable, so it is not evident whether this model satisfies our modelling rule. But we do know that the model will satisfy the rule, *if* Suzy's throw and the bottle's smashing are events which ground the variables *SuzyThrows* and *BottleSmashes*.

2.2 Grounding events

Having already suggested that the relevant connecting process in this model is the process of the rock flying from Suzy's hand until it collides with the bottle, it is easy to see what we intend the grounding events to be. For *SuzyThrows*, the grounding event is Suzy's throwing, and for *BottleSmashes* the grounding event is the bottle's smashing.

In straightforward cases like this, the grounding relation between the event in question and the value of the variable is, we conjecture, simply the relation of *truthmaking*. That is, an event e grounds fact X if: if e exists then X obtains.[7] An event grounds an instantiated variable then, if it grounds the fact which the instantiated variable represents.

Not all cases of grounding are so straightforward. In the alternative state of this simple model, where *SuzyThrows = false* and *BottleSmashes = false*, there are no localized events which are truthmakers for the facts that Suzy did not throw and that the bottle did not smash. This is not a problem for this model, for our rule directs us only to find a connecting process between grounding events for *one* possible state of the model.

Consider, however, a case of causation by omission.

Example 3 (Suzy Catching) *Suzy stands in front of the window, ready to stop the flying ball. There is no obstacle between Suzy and the window. If Suzy catches the ball, the window will not break. If she misses, it will break. The simplest appropriate model is* SuzyCatches → WindowSmashes, *with structural equation* $WS = \neg SC$.

The ball is on its way, but on a whim, Suzy decides not to catch it. The ball crashes into the window, which immediately shatters. Suzy's failure to catch caused the window to break. Thus the actual state of the model represents:

> The window broke because Suzy did not catch the ball.

The alternative possibility was that Suzy might have caught the ball, and the window not have broken:

> The window did not break because Suzy caught the ball.

In each of these scenarios, it is difficult to find a truthmaker for both the cause and the effect. In the first scenario, the smashing of the window is a clear truthmaker, but there is no similarly localised event which necessitates the truth of the proposition that Suzy did not catch. In the second scenario, Suzy's catching the ball is a convenient truthmaker for the fact that Suzy caught the ball, but there is no similar truthmaker for the fact that the window did not smash.

[7] Note we are using "fact" to refer to something like *truths*, and thus as things that can have *truthmakers*. Not everyone adopts this terminology, as for some – such as Russell – "fact" refers to the truthmakers themselves.

In the literature on truthmaking, these truths that seem to lack truthmakers are known as "negative truths". Note that negative truths are not identified by linguistic form, but by their apparent metaphysical properties: they appear to lack truthmakers.[8]

How then ought we to model this case, given the lack of truthmakers to connect up with a physical process? We believe that, by paying attention to what an observer would actually see when looking for causal connections, the resources for solving this problem are made clear. What do we observe when we perceive that Suzy has not caught the ball? Certainly not a mere absence! Rather, we observe at least one event: *the ball's flying past Suzy*. This event (among others) is at least physically incompatible with it being true that Suzy catches the ball (at that time). The negative fact that Suzy didn't catch the ball is not necessitated by this event alone, so we cannot strictly say that this event is a *truthmaker* for the negative fact, but we nonetheless shall call such events *grounds* of the fact.

Generally then,

> An event *e* GROUNDS fact *X* just in case:
>
> 1. *e* is a truthmaker for *X*, or
> 2. the existence of *e* is physically incompatible with *X*'s being false – perhaps with the tacit assumption of some background facts.
>
> An event grounds an instantiated variable just in case it grounds the fact which the instantiated variable represents.

Note then, there is a connecting physical process between «the ball's flying past Suzy» and the event which is the truthmaker of the fact that the window smashed. It is in this sense that a process may connect to a negative fact: *by connecting to an event which is a ground of the negative fact.*

2.3 Problems involving negative causation

The possibility of finding connecting processes via events that are not truthmakers greatly expands the range of possible connecting processes. Thus, the account is threatened by one of the problems which afflicted earlier process accounts: it is not sufficiently discriminating, and will find too many processes. An important element of our account of causation, however, is that some processes are more *salient* than others.

[8] Important historical examples of the struggle to account for negative truths are Russell 1972: 67–72 and Wittgenstein 1922: §2.06. More recently, see Molnar 2000.

In models which have only binary variables and deterministic causal relations – which is the sort we have been considering – there are two types of arc, corresponding to the different types of fact which are represented by the arcs in the various states of the model. A "positive" arc has the following space of possibilities:

	Parent	*Child*
State 1	Positive	Positive
	(throw)	(smash)
State 2	Negative	Negative
	(non-throw)	(non-smash)

For example, the model suggested for the case of Suzy Alone has a positive arc. In one state, the fact that Suzy threw connects via a process to the fact that the bottle smashed. In the other state of the model, the facts represented by the variables are both negative, and there is no salient process connecting them.

The model is adequate, then, because in one state of the model there is a connecting process. In the other state, there is no connecting process – or if there is such a process, it is a far less salient one, relying upon unintuitive grounds of the relevant facts.[9]

This sharp difference between the two states and the relative simplicity of the grounding relations involved suggests that only the throwing–smashing state of the model has a highly salient connecting process.

A "negative" arc – or a "prevention–omission" arc – in contrast, is as follows:

	Parent	*Child*
State 1	Positive	Negative
	(catch)	(non-smash)
State 2	Negative	Positive
	(non-catch)	(smash)

Suzy Catching exemplifies this type of arc. In one state, the positive fact is that Suzy catches the ball, and this prevents the window from breaking. In the other state, Suzy's omission – her failure to catch – causes the window to smash. In both states of the model, it is possible to find something that might meet our criterion of

[9] For instance, it could be said that the event of the rock being in Suzy's hand at t' is incompatible with Suzy's having thrown the rock at t. And moreover, the event of the rock being in Suzy's hand at t'' is incompatible with the bottle's having smashed, because given the laws and facts about Suzy's throwing powers, she could not have thrown the rock so as to smash the bottle, and retrieve the rock again in such a short time.

a connecting process.[10] Unlike the positive arc, however, both states of the model have similarly indirect, somewhat unintuitive connecting processes. Thus, it is not clear which state of the model includes the more salient connecting process.

This observation is generally applicable to prevention–omission arcs. In such "negative arcs", it is frequently not obvious which state has the more salient connecting process. Consequently, we predict that – as a rough generalisation – causal judgments regarding models that involve such arcs will be less stable, and more prone to context-sensitivity.

One way for a model to be inadequate is for there to be no process which is sufficiently eligible for salience. This is exemplified – generally – where complex chains of causation are modelled with insufficiently many arcs, such as the following double-prevention case, due to Ned Hall (2004: 241):

Example 4 (Bomber and fighter) *A bomber is heading towards its target, preparing to drop its explosives. An enemy plane, Enemy, approaches the bomber, attempting to destroy it. A third plane, a fighter assigned to defend the bomber, shoots down Enemy, thus preventing Enemy from destroying the bomber. The bomber proceeds to the target, and accomplishes the mission.*

It is plausible to suggest in such a case that the fighter's action caused the mission to succeed. One might model this very simply, as follows: *FighterShoots* → *MissionSucceeds*, with an equation *MissionSucceeds* = *FighterShoots*.

A number of possible and actual processes are clearly relevant to the relationship between these variables. In particular: the process by which the bomb falls from the bomber to the target; the process by which Enemy's projectiles travel towards and collide with the bomber; and the process by which the fighter's projectiles travel towards and collide with Enemy. Despite this abundance of relevant processes, however, the model is inadequate, because in no state of the model does a process manage to connect from the fact that the fighter shoots (or does not) to the fact that the mission succeeds (or does not). Thus the model violates the master modelling rule.

It is not difficult to remedy the above model. We simply need to interpolate additional variables, to highlight the sorts of events which mediate the dependency between the fighter's shooting and the mission succeeding. A model: *FighterShoots* → *EnemyAttacks* → *BomberDrops* → *MissionSucceeds*, for instance, would satisfy the rule. There is a possible process from the Fighter's shooting to an event

[10] In the first, it is a process that runs from Suzy's catch to the ball being in her hand at a later time. In the second, a process that runs from the ball flying past Suzy at t to the later smash.

which is a ground of the fact that the Enemy does not fire. There is a possible process from the event of the Enemy's attacking to an event which would be a ground of the fact that the Bomber does not release its bombs. And finally, there is a possible process from the event of the Bomber dropping its weapons to an event which is the truthmaker of the fact that the mission succeeds.

Note, we do not claim that violating the master rule will always lead to incorrect causal judgments. Indeed, in the case described above, the simple model does just as well as our more complex model at delivering the verdict that the fighter's shooting caused the mission to succeed. We claim, rather, that the modelling rule is an attractive way of integrating process theories into the technique of causal modelling, and that it is the basis of a refinement in causal model analyses which allows them to be successfully applied to late preëmption cases, as we shall show in sections 3 and 4 below.

2.4 Context sensitivity and spoils to the victor

Salience, clearly, is a partly psychological phenomenon. Thus, the account of causation we are offering is in some sense not objective. For instance, people considering causal scenarios are, we imagine, frequently ignoring – or are outright ignorant of – some of the relevant processes. This will lead to causal judgments which differ – perhaps only in certain relatively rare circumstances – from those who are more attentive to the relevant processes.

In general, if certain processes are highly salient to a wide range of subjects, this will result in stable judgments about causation. If, however, there are a number of processes of similar eligibility for salience, then contextual factors might influence judgments of causation.

Therefore, the theory we develop below has the resources to explain a certain sort of context sensitivity in causal judgments, and in particular why some judgments are more susceptible to contextual influence than others.

When discussing a case of symmetric overdetermination, David Lewis famously said that he had no clear intuitions about such cases, and consequently the case was one of "spoils to the victor" (1973: 194). Whichever analysis delivers a better account of cases where intuitions are clear may be declared the victor, and we should accept the conclusions of that analysis on the cases where intuitions are less clear.

This methodology has some appeal: if an analysis solves all the problems where we are sure we know the answers, then all the more reason to trust its verdict where we don't feel so sure that we know the answers. But this methodology sits uncomfortably with the conservative aims of conceptual analysis. If our folk-concepts are unclear or disputed, then successful conceptual analysis should arguably explain why they are unclear or disputed (as Lewis recognised, *ibid.*). We believe that the salience requirement for connecting processes has this sort of explanatory power: it explains why intuitions are conflicted, fuzzy, contextually sensitive, or disputed in the relevant sorts of case.

On the other hand, one might complain that the sort of explanation we are offering is too psychological. Perhaps it is wrong to think that causation is partly an epistemic or subjective phenomenon. Perhaps a case can be made to eliminate the subjective aspect of salience from the correct analysis of causation. We cannot begin to address these matters satisfactorily here. Rather, we wish to make the more modest claim that judgments of process-salience appear to play a role in judgments of token causation. This might be because of cognitive error on our part, or it might be because it is an essential part of the causal concept. We remain agnostic on that question, though we shall talk for convenience, as though process-salience is indeed part of the correct conception of causation.

2.5 Is salience itself a causal concept?

We have suggested that, in explaining causal judgments, we must make some reference to the salience of possible connecting processes. What makes a process salient, however? Moreover, we have complained against Halpern and Pearl that the principle they offer to guide model choice in preëmption cases appears to suffer from an excessive degree of circularity. What makes us confident that an appeal to salience does not itself involve reliance upon causal concepts, and therefore fall victim to the same criticism?

With respect to the first question, we think that we do best to introduce the concept of salience by way of paradigmatic cases, rather than to attempt an analysis of a concept that is largely psychological in its nature. *A priori*, it is hard to say what makes something salient. We can say this much: it seems likely that salience will be *relative* to the causal claim in question. A baseball careening through a wall is salient, relative to causal claims about what caused the window to smash. A photon travelling from the ball to the window is much less salient, relative to the same class of causal claims. Relative to a causal claim about what caused the window to *remain transparent* from the time of the throw till the time of the smash, however, both of these processes are similarly low in their degree of salience.

Why do facts such as these – facts regarding salience – hold? Perhaps it is because of prior causal assumptions we hold regarding, for instance, the ability of massive objects like balls to shatter objects like windows, and a corresponding inability of photons to shatter windows. Or perhaps these are brute psychological facts not susceptible of further analysis. We think it wisest to leave this for those adopting a more empirical approach to this question. In this paper, we submit, the claims we make regarding salience are plausible enough to render the program worthy of further examination. And if the approach recommended in this paper proves to have promise, it should be a topic of future interest to psychologists and philosophers to inquire further into the nature of salience, and to examine the hypotheses we have very crudely raised here.

With respect to the second question, then, we cannot blithely dismiss out of hand the concern that we might be involved in some degree of circularity. Until we know exactly what salience is, we cannot be sure. However, we think at this early stage it needs to be stressed that we have made some progress, relative to the approach of Halpern and Pearl, by broadening the circle of analysis. Rather than asking directly: "is *c* or *d* a cause of *e*?", as their heuristic suggests, we ask a much less direct question about the possible processes involved. We ask with respect to each arc, *is there a possible state of the model in which a salient process connects the variables along this arc?* If not, the model is problematic. In the context of Billy and Suzy, for instance, our concept of salience does not require that the possible process by which Billy's rock smashes the bottle be in any way deprecated as "non-salient". That process can occur, and is just as salient as the process from Suzy to the bottle. As we will show below, however, information about salient possible processes can be used in an analysis of Billy and Suzy which vindicates the judgment that Suzy, but not Billy, is a cause of the smash.

3 Vulnerable arcs

A double-prevention case with only a single arc suffers from having no salient process for any state of the model. This is the most fundamental sort of model failure, and is proscribed by our master modelling rule. A more specific way for a model to fail is to have no salient possible process corresponding to an arc, *given the actual state* of distinct causal pathways on the model.

This is what occurs in Billy and Suzy. If *SuzyThrows* = true, no connecting process can exist between *BillyThrows* and *BottleSmashes*. If Billy throws, the process of Billy's rock flying through the air exists, but this process clearly does not connect to an event which grounds the fact that the bottle smashes. And if Billy does not throw, then it would appear no relevant process exists whatever.

We say that an arc $A \rightarrow B$ is VULNERABLE just in case, for some variable X, where X is not A or B, nor an ancestor of A or a descendant of B, for some value χ of X, it is physically impossible that $X = \chi$ and a connecting process exists between an A-grounding event and a B-grounding event.

A stronger modelling rule which might seem attractive then, is:

ANTI-VULNERABILITY RULE Causal models must contain no vulnerable arcs.

The thought behind such a rule might be: arcs must represent potential causal influence. Influence is mediated by the possibility of a connecting process. But if an arc is vulnerable, then it it possible for the parent variable of that arc to be rendered impotent by a factor on a distinct causal path. Possibly impotent variables should not be treated as potential causes, so avoid introducing vulnerable arcs.

This rule, if accepted, would give us a principled reason to reject the simple model of Billy and Suzy. Unfortunately, however, this requirement appears to be too strong. It rules out a great number of models that are intuitively very plausible. And we are aware of *no* intuitively plausible model of Billy and Suzy which respects this rule.[11] Instead, we propose a modification of the analysis of token causation, where that modification pays crucial attention to the status of vulnerable arcs.

3.1 Wounded arcs

If we are analysing token causation in a model with vulnerable arcs, if any of the vulnerability-inducing variables X actually instantiate χ, then we say the relevant vulnerable arcs have been WOUNDED. Thus, in the simple model of Billy and Suzy, if Suzy throws, the arc *BillyThrows* $\rightarrow$ *BottleSmashes* is a wounded arc.

Our suggestion is that, when considering a token case in which wounding has occurred, then wounded arcs must be removed from the model. Thus if the original model had a structure $A \rightarrow C \leftarrow B$, and the arc from B is wounded, the new model should have the structure: $A \rightarrow C$.

The new equation for C therefore will make no mention of B, and it is not possible to get the result that B is a cause of C.

We do not offer an algorithm to then "repair" the modified model and write a new equation for C – this is work that remains to be done.[12] What we *are* offering

[11] We leave it as an exercise for the reader to confirm that both the model in Figure 2 and the time-indexed model preferred by Halpern and Pearl contain vulnerable arcs.

[12] It might be wondered why anything needs to be done more than simply setting C to its actual value. This will then get the result, however, on most analyses, that Suzy's throw is not a cause of the bottle smashing. See Korb et al. 2005 for further discussion of a strategy to proceed beyond a

is a principled reason to *eliminate* potential causes. To solve the problem of late preëmption, this is an important first step. Late preëmpted potential causes are not actual causes, because they are connected to the effect via *wounded* arcs.

Given the modesty of what our account achieves, it might be wondered whether it is worth the metaphysical investment in truthmaking, processes, and so forth. We believe it is, precisely because the phenomenon of wounding – even without a full analysis of causation – seems to explain a variety of puzzling cases.

4 Applying the account

4.1 Preëmptive prevention

Example 5 (Suzy and the Backup) *A ball zooms towards a window, but is caught by Suzy* (*SuzyCatches = true*)*. But if Suzy had not caught it, it would have struck a brick wall* (*BrickWallPresent = true*) *which stands in front of the window. We want to know if Suzy's catch prevented the window from breaking* (*WindowSmashes = false*).

$$\begin{aligned} \mathit{SuzyCatches} &= \mathit{true} \\ \mathit{BrickWallPresent} &= \mathit{true} \\ \mathit{WindowSmashes} &= \neg \mathit{SuzyCatches} \wedge \neg \mathit{BrickWallPresent} \end{aligned}$$

This example is derived from one first suggested by Michael McDermott (1995: 525), and subsequently discussed by John Collins (2004). It is an interesting case precisely because it evokes somewhat unstable intuitions. A typical initial response is that Suzy did not prevent the window from smashing, because the presence of the wall made her redundant. McDermott then suggests we ask: Given that together Suzy and the wall prevented the window from smashing, *which one* was the cause (or was it only both)? To this, it is not uncommon for intuitions to flip over, and to endorse Suzy as a cause, because it seems like the brick wall made no contribution.

Our account explains why people are clear that the wall was not a cause, but ambivalent about Suzy's catch: the arc *BrickWallPresent* → *WindowSmashes* is clearly wounded, but the arc *SuzyCatches* → *WindowSmashes* is wounded on one plausible reading, and intact on another. To see this, consider what the possible connecting processes are, corresponding to the arc from *SuzyCatches* to *WindowSmashes*. This is a negative arc, as described above, so in no state of the model is one connecting process obviously salient. In one possible state, a process connects

wounded model to a complete analysis.

from the ball flying past Suzy to the window's smashing. In another possible state, a process connects from Suzy's catching the ball to the ball being in Suzy's hand at some later time – an event which is incompatible with the ball smashing into the window. Suppose that only the first process is salient to a subject making a causal judgment about the case. Given the presence of the wall, it is not physically possible for this process to exist. Therefore, relative to this assignment of salience, the arc is wounded, and the subject should judge that Suzy did not prevent the window from smashing. Suppose, however, that the other possible process becomes salient for the subject – perhaps because, on further reflection, they are attempting to find some eligible process which explains the apparent causal relation between either Suzy or the wall and the prevention of the window's smashing. This process, however, is still possible *even with* the wall present. Therefore the arc is not wounded, and the subject may judge that Suzy is indeed a cause.

Note, in contrast, while two possible processes are eligible for the arc from the brick wall to the Window – one a process involving a ball flying past the absent backup, and one involving the ball being blocked by the wall – *both* of these processes are wounded by the fact that Suzy catches. So there should be no inclination to identify the presence of the wall as a cause. This arc is unequivocally wounded.

Therefore our account predicts the observed instability in our judgments about Suzy as a preventer of the window smashing. And moreover, it predicts that there will not be a similar instability in intuitions about the efficacy of the wall. This is precisely the pattern of intuitions that McDermott reports.[13]

4.2 Other cases

We have now shown how the wounding analysis assists in discussing cases of late preëmption and preëmptive prevention. We conclude by briefly reviewing two other problem cases where the wounding analysis yields a promising result.

4.2.1 Context sensitive cases

Example 6 (Deadly antidote) *An assassin gives the king some poison* (*Poison = true*)*. The king's bodyguard notices this, and quickly administers an antidote* (*Antidote = true*)*. The king survives* (*Survives = true*)*. But note, the antidote, if given*

[13] Note that our account does not, on its own, explain the tendency to rule Suzy as a cause more often for some backups than others. For instance, if the backup is a human, rather than a brick wall, it is more common to judge that Suzy is a cause. However, strategies that have been suggested by others to explain such phenomena (e.g. Collins's "dependence prevention" account (2004)), appear to be compatible with our preferred approach.

in the absence of poison, is deadly. The straightforward model is Survives $= (P\&A) \vee (\neg P\&\neg A)$.

This has been suggested as a counterexample to the analyses of Hitchcock, Halpern and Pearl (Menzies 2004a), because those analyses yield the dubious conclusion that giving the poison caused the king's survival. This is because, holding fixed the fact that the antidote was given, whether or not the poison is delivered makes a difference to the outcome. Hence the path is active.

On reflection, one can see what underwrites this result, since, relative to the substance delivered by the bodyguard, the "Poison" is also an antidote. The causal roles of the two substances are strikingly symmetric, but our judgment is non-symmetric, presumably in virtue of contextual information, regarding the different times at which each was administered, and perhaps the different intentions with which each was administered. None of this contextual information is included in the above model, however.

We observe that each substance has two possible connecting processes: one when acting alone, and one when the other substance is also present.

When acting alone, either substance damages internal organs, grounding death – call this a poisoning process. When either acts in the presence of the other, it meets with and neutralizes the other substance (and vice versa) – call this a neutralizing process.

So we suggest that to model the influence of contextual information, we need to treat only one of the possible connecting processes for each arc as salient. The "Poison" is administered *alone*, so we disregard the neutralizing process on its arc. The "Antidote", however, is administered in the presence of the "Poison", so it can only neutralize: the possible poisoning process for this second substance is not salient.

Therefore, the arc *Poison* ⟶ *Survival* is wounded because *Antidote* = *true* disrupts the salient connecting process: poisoning. Conversely, *Poison* = *true* fails to wound the arc *Antidote* ⟶ *Survival*, because it is compatible with the salient neutralizing process.

Thus there is the potential for a sort of context-sensitivity in our intuitions about this sort of case. If one of the substances' causal roles is emphasised at the expense of the other, it may lead to our ignoring a possible connecting process, thus leading to different patterns of wounding, and leading to different judgments about causation.

And this appears to be vindicated by the following modification of the example. Suppose that, instead of a faithful bodyguard who is administering an antidote, the second substance is delivered by a second assassin, operating in ignorance of the first. The second assassin gives the substance in the belief that it is a poison which will cause death. Having described the case this way, we suspect some people will endorse the conclusion which was raised as an objection above: that *neither* substance has caused the king to survive. And this is plausibly thought to be because, described in this fashion, the example obscures the antidote-role of the second poison.

Alternatively, the temporal order of the administration might be doing the work of determining salience here. So it might be that people continue to endorse the second assassin's deed as the cause of survival, because the prior administration of the first substance establishes a threat, making the *neutralising* process more salient.

4.2.2 Problems with redundancy ranges

Example 7 (Hiddleston's Terror alert) *A terrorist alert goes out that a poisonous gas will soon be released into the atmosphere. Suzy takes the precaution of consuming an antidote that will protect her. The alert, however, was a false alarm, and no poisonous gas is released. The antidote is otherwise harmless, and Suzy survives. The model is:* $Survives = \neg Poison \vee Antidote$.

Eric Hiddleston (2005) raises this as a counterexample to Hitchcock's weak path analysis, and also to Halpern and Pearl's similar analysis. Given the antidote has been taken, $Poison = true$ is in the redundancy range for the path *Antidote* $\twoheadrightarrow$ *Survival*. And for that value in the redundancy range, the path from antidote to survival is active. Therefore taking the antidote is held – implausibly – to be a cause of Suzy's survival.

If the process from ingestion of antidote to interaction with the poison is salient, however, the arc from *Antidote* to *Survival* is wounded by the absence of the poison. Thus we avoid the result that the presence of antidote is a cause.

5 Conclusion

We have shown that the method of wounding causal models is a powerful means of incorporating the insights of process theories of causation into an essentially counterfactual approach to causation. Moreover, it shows how some of the context-sensitivity of our intuitions about causation can be accounted for in terms of processes. While we have not addressed all existing problems, we think the elegance and power of this technique makes it worthy of serious consideration.[14]

TOBY HANDFIELD
School of Philosophy & Bioethics, Monash University
CHARLES R. TWARDY
Information Extraction & Transport, Inc. (IET)
KEVIN B. KORB
Clayton School of Information Technology, Monash University
GRAHAM OPPY
School of Philosophy & Bioethics, Monash University

References

Collins, John. 2004. "Preëmptive Prevention". In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul. Cambridge, Ma.: M.I.T. Press, 107–17.

Dowe, Phil. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Fair, David. 1979. "Causation and the Flow of Energy". *Erkenntnis* 14: 219–250.

Hall, Ned. 2004. "Two Concepts of Causation". In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul. Cambridge, Ma.: M.I.T. Press, 225–76.

Halpern, Joseph Y., and Judea Pearl. 2005. "Causes and Explanations: A Structural–Model Approach. Part I: Causes". *British Journal for the Philosophy of Science* 56: 843–87.

Hiddleston, Eric. 2005. "Causal Powers". *British Journal for the Philosophy of Science* 56: 27–59.

Hitchcock, Christopher. 2001. "The intransitivity of causation revealed in equations and graphs". *Journal of Philosophy* 98: 273–99.

Korb, Kevin B., Charles R. Twardy, Toby Handfield, and Graham Oppy. 2005. *Causal Reasoning with Causal Models*. Tech. Rep. 2005/183, Monash University, Faculty of Information Technology.

Lewis, David. 1973. "Causation". *The Journal of Philosophy* 70: 556–67. Reprinted (plus postscripts) in Lewis 1986, 159–240.

———. 1986. *Philosophical Papers*, vol. 2. New York: Oxford University Press.

———. 2000. "Causation as Influence". *The Journal of Philosophy* 97: 182–97.

[14] This research was supported by a Monash University Arts–IT Grant. Handfield's work on this paper was supported, in part, by a Monash University Research Fund Postdoctoral Fellowship and, in part, by an Australian Research Council APD Fellowship. Thanks for helpful comments or suggestions to two anonymous referees and to audiences at Oxford and at the Australasian Association of Philosophy Conference, Sydney, 2005.

———. 2004. "Void and Object". In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul. Cambridge, Ma.: M.I.T. Press, 277–90.

McDermott, Michael. 1995. "Redundant Causation". *British Journal for the Philosophy of Science* 46: 523–44.

Menzies, Peter. 1999. "Intrinsic Versus Extrinsic Conceptions of Causation". In *Causation and Laws of Nature*, edited by Howard Sankey. Dordrecht: Reidel, 313–29.

———. 2004a. "Causal Models, Token Causation, and Processes". *Philosophy of Science* 71: 820–32.

———. 2004b. "Difference making in context". In *Causation and Counterfactuals*, edited by John Collins, Ned Hall, and L. A. Paul. Cambridge, Ma.: M.I.T. Press, 139–80.

Molnar, George. 2000. "Truthmakers for Negative Truths". *Australasian Journal of Philosophy* 78: 72–86.

Russell, Bertrand A. W. 1972. "The Philosophy of Logical Atomism". In *Russell's Logical Atomism*, edited by David Pears. London: Collins. First published 1918.

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

Schaffer, Jonathan. 2001. "Causation, influence, and effluence". *Analysis* 61: 11–19.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul. Translated by C. K. Ogden.

Woodward, James, and Christopher Hitchcock. 2003. "Explanatory Generalizations, Part I: A Counterfactual Account". *Noûs* 37: 1–24.