The Metaphysics of Interventionist Causal Modelling

Elliot Pine

University College London

MPhil Stud.

Abstract

This thesis critically examines a theory of causation called "interventionism", along a number of different dimensions. Interventionism is a manipulationist theory of causation, in that it seeks to give an account of what it is to cause something by reference to what it would take to bring about change in a given variable via a manipulation of some variable causally upstream from it. The manipulation in question happens via an intervention – an idealized "reaching in" to the system under investigation, in order to set the value of the cause variable artificially. The definition of causation thus relies on the notion of a causal system, or model.

The thesis explores what interventionism is a theory *of*. It turns out not to be giving a metaphysical account of causation, but is better seen as a theory of how to capture the meaning of causal statements. A number of criticisms of interventionism are deflected. One is that the theory is irredeemably circular. Another is that it is not equipped to deal with certain cases of pre-emption. A third criticism is that it cannot be a fully objective theory, since it defines causation relative to a set of variables, presumably chosen by a modeler. Finally, I show that, though interventionism tries to stay as metaphysically neutral as possible, this neutrality cannot be maintained for cases of genuine indeterministic causation. Specifically, I show that it is incompatible with singularism - the doctrine that, though causes can be indeterministic, they must always be determinate.

Table of Contents

Chapter 1 – Desiderata for a theory of causation

What should we expect from a theory of causation? What should be its desiderata? On the one hand, causation appears to be the quintessential metaphysical subject of enquiry – what *is* causation, and how best can we characterize it? Such an enquiry might look for definitions, or it might try to give necessary and sufficient conditions. Questions about realism might concern us. And indeed, much literature exists examining these questions. Of course causation also demands an epistemological enquiry; in fact this has often been the starting point for many writers, especially given the long shadow cast by Hume's scepticism concerning what we can know about natural necessity, and, by extension, causation. How far and into what realms Hume intended his skepticism to be taken is the matter of some debate, but even on a simple reading, one is struck by the tight interplay between the epistemology and metaphysics of causation. Perhaps this should come as no surprise – after all, it might be argued, the extent to which we can have epistemic access to whatever causation is, will constrain the ways we can characterize it when giving a metaphysical account.

But perhaps more obviously, the epistemological question (for Hume at least) seems to dominate the question of *meaning* in our causal talk. Again, the exact limits of Hume's thought here are not clear, suffice it to say that his empiricism is what drives this thought. Hume's so-called 'copy principle' says, roughly, that all our ideas concerning the external world must ultimately derive from information taken in via the senses; and since we can only have any kind of understanding of things we have ideas of, what we can mean by our utterances concerning the external world is thus delimited by the extent of our epistemic access *to* that world. In the case of causation, then, if one were a skeptic about what we take in via the senses, because all there is to see is just the movement of objects and not their 'necessary connexion', then unless we can re-engineer a way of finding meaning for causal talk, it will simply be meaningless.[1]

So a theory of causation encompasses metaphysics, epistemology, and derivedly, questions about meaning. Finally, there seems to be a question about the everyday, layman's concept of causation. It is not clear where this everyday conceptual question slots in. After all, it might be asked, why should we care about what the unreflective thinker takes to be the central connotations of the term? An important distinction thus opens up, especially when it comes to questions about meaning, between those who stress the importance of what we *do* mean and what we *can* mean in our causal talk.

1.1 Interventionism and its desiderata

This thesis will critically examine a theory of causation called *interventionism*. Interventionism can be used as something of an umbrella term, covering as it does a number of different theories, not all within philosophy. I focus on a recently defended account, spelled out in James Woodward's book "Making Things Happen." The core of the theory is that causal relationships are those relationships that are potentially exploitable for the purposes of manipulation and control. The term 'interventionism' derives from one of the theory's core ideas – that it is by means of an idealized intervention in a causal system that

[1] A common reading of Hume interprets him as saying that the idea of necessary connexion is generated internally, following repeated experiences of the cause-effect relation. Once the idea is generated, we are able to attach meaning to causal utterances. This would be to take a so-called 'projectivist' view. It could be claimed that Hume fails to show how this is possible. If one took Hume's empiricism seriously but accepted this failure, then causal talk would be left meaningless, or at the very least it would only capture the non-necessity involving features of causation, like contiguity and temporal priority.

such manipulation and control is achieved. This 'in a nutshell' explication elides much of the technical detail which will be explained in chapter 2. But I have also introduced some terms here which look to have certain metaphysical connotations. Most importantly, interventionism talks of causal relation*ships* and causal *systems*. This implies that it is primarily a theory of general causation, connecting event types, rather than tokens. Though it is not limited to a theory of general causation, this primacy of type over token causation is perhaps one way in which interventionism has metaphysical implications. This is important, because as we shall see, interventionism tries to stay as metaphysically neutral as possible. Much of this thesis will be devoted to drawing out some of these metaphysical implications. As a prelude, then, it is worth exploring some of the main themes within the metaphysics, epistemology, and questions about meaning in the topic of causation. Once we have laid down some of the more important views, we shall be in a position to assess whether and how interventionism is committed to any of them, and to what extent it can stay silent on some of the more metaphysical questions.

1.2 Metaphysics and Meaning

There may or may not be a thing(s) in the world which answers to the term 'cause'. The umbrella term of realism can be applied to those theories which hold that there is or are such things. A commonly held view is that there is a causal relation which holds between events. One event occurs, and then another one does; in between these events there is a relation which binds them together such that the first is said to cause, or bring about, or produce, the second. There is an intuitive pull towards the idea that, when one event causes another, the second event somehow *depends* on the first, such that were the first not to have taken place, the second would also not have. This 'counterfactual dependence' view is commonly held, and will be the subject of some discussion in chapter 3. But it is by no means the only way of specifying what the causal relation is; in fact from some of the causal terms we employ (cause, produce, bring about, dependence) we can get a handle on the various ways we could characterize the nature of the causal relation. Indeed, perhaps there is no single causal relation - we may have to make do with a multiplicity of relations.

On either of these two 'realist' positions (monism, pluralism), there is a simple link that holds between the relation itself as it exists in the world, and what we mean when we use causal language.[2] Simplest of all is the view that there is one single causal relation - in such cases whenever we use the term 'cause' or similar language, the referent of the term just is this relation, and on a simple truth-conditional view of semantics, we either say something true or false in a given situation depending on whether the causal relation holds between the events in question. Slightly more complicated is the multi-relational (or, pluralist) view; the term 'produce' might pick out a different relation from the term 'depend', (say) and we would then go wrong in describing the relation of dependence as one of production. (The term 'cause' would probably be ambiguous, or perhaps context sensitive, and it would not be obvious as to the truth or falsity of statements containing the term.)[3]

Furthermore, the realist has a fairly easy time explaining the relationship

[2] At least, as far as the *referent* of the terms go.

[3] Ned Hall (2004) strongly defends the notion that there are 2 concepts of causation, 'production' and 'dependence' which are incompatible with one another.

between the metaphysical question and the conceptual question. In point of fact, it is often argued that, methodologically speaking, the direction of 'discovery' should run from conceptual analysis to the world. On this view, we might notice that, for example, a monolithic concept has internal inconsistencies - leading one to posit a multiplicity of concepts. A realist who bought such an argument and who sympathized with this methodology would have to accept that each of the new concepts would then map onto a separate relation in the world.[4]

The realist camp has more to debate than just whether causation is a single relation. The nature of the relation itself, and the interplay between causation and causal laws have also been much discussed. For example, some argue that the causal relation is itself reducible to something else, whilst others claim that this is implausible and demand a more 'robust' notion. David Lewis is the most famous reductionist, and the recent literature on causation is dominated by his work. His programme of 'Humean Supervenience' was designed to remove any spooky natural necessity from his metaphysical picture. Instead, Lewis claimed that the relation of counterfactual dependence could itself be reduced to a closeness relation between worlds, and where the laws of nature in any world are nothing more than the regularities that hold there.[5] I shall return to Lewis' theory in chapter 3, but for now it is worth noting that this view is quite radical, and is at odds with another popular view, namely that there exists in the world a singular irreducible causal relation. These 'singularists' claim that the causal facts are not exhausted by what laws there are coupled with the initial conditions

[4] This is, in fact, how Hall argues for his dualist thesis.

[5] The 'regularity' view of laws introduces another dimension to the reduction. See Tooley (2004) who explains that causation can be reduced to non-causal facts, including what the laws are. But the laws of nature can further be reduced to mere local matters of fact – where the laws just systematize what goes on at a world. Lewis requires this further reduction in order to fully purge natural necessity from his ontology.

of a world, since we can imagine causation in a world in which laws are not fully deterministic (See e.g. Foster's "Ayer" p256), or even in entirely lawless worlds (Anscombe (1975)).[6]

We can most clearly contrast realists with nihilists. Russell is often touted as the arch nihilist, based on the arguments proposed in his 1912 paper "On the notion of Cause." I will not go into the details of Russell's arguments here; suffice it to say that Russell stresses the apparent lack of the use of the notion of causation within the scientific community. He thinks this is most starkly demonstrated in the way that mathematical equations which scientists posit to describe physical reality do not make reference to causation. Whatever the reasons for nihilism, such a theory faces some tricky questions. Firstly, it would need to propose some kind of error theory to explain why we think that there is such a thing. This is no mean feat - a major part of Hume's treatise is dedicated to this very task - whence the idea of necessary connexion? Furthermore, one would have to explain how the semantics of causal statements might work, would they be plain false or would they be elliptical for something else?

So the broadest distinction we can draw in the metaphysical debate is this 'realism versus nihilism' question. If we accept realism, we can ask two further questions. Firstly, is there just one causal relation or are there many? Secondly, can causation be reduced to something else, or must we accept irreducible causal

[6] A note of clarification is in order here – it is not that Lewis' theory cannot cope with indeterminism. Indeed he was at pains to spell out how his theory could be extended to cover indeterministic causation. Rather, the problem for Lewis is that there are imaginable circumstances in which *two* (or more) indeterministic causes can overlap. Lewis is committed to the view that there is no fact of the matter as to which of the causes acts to bring about the effect. The singularists claim this is implausible. See ch 5 for a discussion of singularism and more on probabilistic causality and indeterminism.

facts? However, we have still not covered all possible positions. We might wonder whether causation is fully objective, or whether it is mind dependent, (perspectival, subjective[7]) in some way. In fact this distinction is particularly relevant to interventionism, since a fore-runner to Woodward's theory is an anthropocentric version of interventionism, which tries to analyse causation in terms of *agency*. For example, Menzies and Price (1993) think that

"...an event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* would be an effective means by which a free agent could bring about the occurrence of *B*." (1993, p. 187)[8]

The reference to a free agent introduces a thoroughgoing anthropocentricity to their view of what causation ultimately is. They do not deny that causation is real, but argue that it can only be understood – and must be reduced to - facts about what would be involved in a minded agent bringing about some effect. Woodward's interventionism attempts to free itself from such anthropocentricity by introducing the notion of an idealized intervention – similar in nature to a human intervention, but not limited in the way that a human intervention is. But even if this strategy is successful, some subjectivity may still lurk. This is because Woodward's theory makes reference to *causal systems*. As it turns out, Woodward's definitions of causation necessitate reference to a model which represents some worldly system. But of course modelling is a human activity, and there may be many ways to correctly model a given system.[9] If this is so,

[7] Though each of these terms has different connotations, I shall not draw distinctions here. The point is just to contrast objectivity from these other positions.

[8] The theory is interventionist in nature since the agent must of course intervene in the world to bring about the desired effects.

[9] Of course there are also simply *wrong* ways. As an example of how two models might correctly capture a real world system, think of the way that a chemist and a physicist might represent a

then there might be a sense in which causation is somehow perspectival, or interest-relative, since two models could correctly represent some system, but give conflicting answers to whether something is a cause of something else.

Finally, there is a question about the relata (assuming causation exists and is a relation) of causation. The orthodoxy holds that it is events (including, when broadly construed, states) which cause and are caused. Some hold that, although causation *usually* holds between events, it does not always have to. Particularly salient here is the problem of omissions. Omissions don't seem to be 'things' in the world which can bring about other things - how could they? Under pressure to give an account of causation by omission, some philosophers have argued that sometimes causation can have missing relata (Lewis (2004b)), or that it is facts which should play the role of causal relata (See especially Mellor (1995)). Interventionism is rather funny in this regard, since it refuses to be drawn on this question. Again, the interventionist definitions of causation require reference to causal models. These models, as we shall see, use nodes and arcs to represent event types and the causal links between them. Because the definitions can 'go through' without specifying the underlying nature of what is being represented, we can arrive at the 'right' result without worrying about such metaphysical slipperiness as negative events, or how it is that something can be a cause without tracing a spatio-temporal path to its effect.

pressurized gas. Different levels of analysis require different representations. Higher level models necessitate the eliding of much causal information.

1.3 Epistemology

A theory of causation should have something to say about how we know about what causation is, or at least about what we *think* it is. It should also answer another epistemological/methodological question - namely, what is it about any methods we use for finding causal links in the world that makes them so successful? These are two distinct questions. One involves questions of psychological development. As already noted, this enquiry forms a substantial part of Hume's Treatise. Given his skepticism about the availability to the senses of any necessity, Hume was forced to retreat inwards, claiming that the repeated experience of one thing followed by another led to the feeling of necessity via habituation. This is a theory of how the concept is formed. A somewhat less sceptical position than Hume's might argue that we do perceive causality, and not infrequently. One interesting proposal is that we perceive causality when objects impinge on our bodies. Though Hume may have been correct in his scepticism concerning *visual* causal stimuli, we can perhaps gain a feeling of necessity through touch. The debate surrounding the epistemology of causation has taken something of a back-seat in recent years, following the dominance of 'speculative' metaphysics, for which it has been felt (for one reason or another) that no epistemological story need be told. Perhaps it is just assumed that, as realists, we must have some way of acquiring the concept, and that once we have it we can simply analyze it to get at metaphysical truths about the world. One need not quarrel fundamentally with this strategy to notice that, nevertheless, we should probably prefer a theory of causation for which there is a compatible and explanatorily relevant theory of developmental psychology. If it can be shown, for example, that children develop their causal concept by making interventions in the world to 'see what happens,' then this would certainly be a point in favour of interventionism, as compared with a theory whose psychology was entirely divorced from the concept. We might wonder, for example, how we could ever get knowledge of what goes on at other possible worlds, or how we were supposed to assess which worlds are the closest possible worlds to our world – a crucial element of Lewis' counterfactual theory. Nothing in our psychological make-up makes reference to other possible worlds or their similarity, and even if this is no reason to junk the theory (for it might have significant other advantages), it surely is a point against it.

Our other epistemological question concerns the way in which we get causal knowledge. This is particularly salient in scientific contexts. Though we may be sceptics concerning our knowledge of natural necessity, it surely is the case that we are incredibly successful in using causal knowledge gleaned via experimentation to make our way in the world. Despite Russell's claims about what use scientists may or may not have for the notion of causation, engineers (and many involved in the special sciences) certainly do have use for it. So, it could be argued that the actual way in which we infer, or test for causal links, might be an important guide to what causation is. We have something of a puzzle resembling that of the problem of induction. Inductive reasoning is apparently unsound because we have no way of rationally grounding it as a good way of achieving knowledge of the future. Yet at the same time it is so overwhelmingly successful that we cannot help but to assume that it does generate good predictions. Similarly, though there might be fundamental questions concerning how it is that we get knowledge of causation, we have been so successful in our methods concerning causal inference, that there must be something about those methods which should yield up information about what causation ultimately *is*. Thus interventionism, in the hands of Woodward, is a theory of causation which links itself to the concept of interventionism. It is only by intervening in some system that we are able to find out about how it works. Physiology is a great example of this. It is only through experimentation via intervention that we can determine what causal routes exist in the human or animal body. We reach into the circulatory or lymphatic system and tweak certain 'variables' in order to see what happens; it is these methods which yield up the causal information which we can then use in medicine or veterinary science.

In making this linkage, there is a danger that a charge of verificationism could be leveled against the theory. If interventionism ties the meaning of causal statements to the way that we can prove them to be the case, then this is potentially a bad move. This challenge will be dealt with in due course.

1.4 Thesis outline

This thesis is arranged as follows. Chapter two explains the fundamentals of interventionism. I will show how the definitions of causation make fundamental reference to a causal model. Also, it will be shown that interventionism is a non-reductive theory. This is because the definitions of causation also make reference to interventions. When we consider what an interventionism is, we find that it is itself a causal notion. The charge of circularity is answered by explaining that the circularity is not vicious. The definitions leave us with 'applicability conditions' which tell us when we are warranted in applying the term.

In chapter three I compare Interventionism with Lewis' counterfactual theory. I bring out the different metaphysical implications of each theory. This includes a careful look at the different priority given to type and token causation. I explain why, even though interventionism has a counterfactual flavour, the assessment of the counterfactuals is so different in each theory. Principally, it is because interventionism assumes that for every causal link, there must be a real-world closed causal system in which that link is embedded. As such, there is no requirement to condition on the entire history of the world up to the point at which the cause happens, in order to assess the truth of the counterfactual. We need only consider what would happen *within that closed system,* and so we are within our rights to, so to speak, 'demonstrate' the truth of the counterfactual by rigging up the very same system and running the experiment again.

Chapter four considers a recent critique of interventionism by the philosopher Michael Strevens. He first of all claims that there are certain cases of pre-emption for which interventionism is ill-equipped to deal. These cases involve pre-emptive causes which act along the same route as the actual cause. I show that Strevens is mistaken in his assumptions about how one must model a causal system. We are not necessarily forced to represent the activity of one worldly entity with just one variable.

He further claims that interventionism commits one to relativism, since the definitions of causation are made relative to a model. He tries to show that, using different models for a given system, we can get conflicting results concerning what causes what. Crucially, his counter-examples rely on what turn out to be faulty models – models which represent the world badly. I answer this charge by showing that although some perspectivalism is implied by interventionist definitions, this does not have radical implications for the objectivity of causation. I draw a parallel with the context-sensitive theory of knowledge employed by Lewis. On the question of how we can guarantee that we have the correct model, I answer that we surely cannot know for sure, but that this does not threaten the theory. It is a pre-condition of interventionism that there are pre-existing causal system out there in the world within which every causal truth is accounted for. We can best understand what the term 'cause' means, as given by interventionism, as something akin to a conceptual role. What are the implications of saying that *X* is a cause of *Y*? A central part of the answer is that, if we are lucky enough to have stumbled upon the correct causal model, we will be able to manipulate *Y* by intervening to change *X*. It should come as no surprise that when we have an incorrect model, we get the wrong results about what causes what.

Chapter five investigates a potential problem for the metaphysical neutrality of interventionism when it tries to give an account of indeterministic causality. The problem occurs because, while at the level of types we can represent the way in which a certain event type raises the probability of there being an event of another type, at the level of tokens, it is plausible to think that something can be a probability raiser without it being a cause. As it turns out, some (including Lewis) defend the idea that all probability raisers must be, in some sense, causes. I close by drawing the conclusion that, in the realm of indeterministic causation at least, interventionism is committed to this fairly controversial view.

Chapter 2 – Interventionism as a theory of causation

In his recent book, "Making Things Happen",[1] James Woodward defends a philosophical account of causation which relies on the notion of interventionism. By a 'philosophical' account, I mean to contrast his work with that of the likes of Spirtes, Scheines and Glymour (1993), and also Pearl (2000), from whom he borrows many significant results. Almost all of that literature has focussed on causal *inference* – the problem of inferring causation from a set of statistical data - a significantly different task from the one metaphysicians focus on.[2] Were he to simply replicate this work within the philosophical community, it might look plausible to defend only the claim that this is how we actually (or perhaps ought to) reason about or find out about causal links in the world. Such an account would do very well in answering at least part of the epistemological-conceptual enquiry discussed in chapter 1. It *might* describe the ways that children learn basic causal truths about the world, for example that solid objects cannot pass through one another, or that causes must precede their effects. Whether it did would be a matter of empirical research.[3] What such a theory would say would be that, via repeated interventions in the world, we learn that

[1] Much of my discussion revolves around Making Things Happen. Any references to Woodward, unless otherwise stated, refer to the book.

[2] Indeed, it is an entirely epistemological problem. Given a set of data regarding some variables in a given environment, how can we discern which variables are causally efficacious, and which are merely correlated?

[3] Such studies have indeed been carried out, with results which seem to (at least *prima facie*) favour the theory. See e.g. Gopnik et al (2004)

certain conditions have to be in place for effects to follow from their causes. For example, throwing a wooden block at a bell causes it to 'ping,' and this is an 'all or nothing' effect - the block must hit in order for the 'ping' to sound. Falling short is no good. Thus, contiguity is learned.

Notice that it would not touch the deeper epistemological question of whether we can really perceive causation. It could only go so far as to explain, epistemologically speaking, how it is that *given we have the concept that we do,* we go about looking for causes in the way that we do, and why those methods are successful. As such, then, it is fully compatible with full-blown Humean scepticism. Indeed, one might think that it is more than just compatible – since it plausibly requires the child to perform repeated interventions before it learns, we have a striking resemblance to Hume's habituation theory.

We must be careful, then, to distinguish between foundational epistemological enquiries on the one hand, and enquiries which focus on how the concept is formed, and the knowledge that we can glean once that concept is taken as a given. An analogy with other epistemological enquiries may be apt here. Foundational questions concerning whether we can know anything at all, given that it is conceivable that we are being deceived all the time, does not preclude us from wondering about the status of certain apparent ways of getting knowledge. We can set aside the foundational worries, and assume that we are, in fact, in touch with a real external world, and still question what right we have to use, say, inductive reasoning to achieve knowledge about the future. In fact the problem of induction provides a good comparator, since Woodward and others press into service the idea that *being successful* in the way we make our way in the world – which, he argues, relies heavily on the idea of gaining causal knowledge through interventions – is itself a justification for our claims on causal knowledge. The paradox of induction is a little like that – it is precisely a paradox because, although foundational questions threaten to undercut its ability to generate knowledge, we nevertheless recognise that we do get knowledge via inductive methods.

But what of the other enquiries? Can interventionism be seen as a theory which describes causation as it exists in the world? And can it provide us with a theory of meaning for causal statements? Woodward himself is somewhat difficult to pin down on these questions, although it is clear that his theory is designed to do more than just provide a philosopher's guide to causal inference as it exists in other disciplines.

This chapter is set out as follows. First I will lay down the foundational technical material which Woodward uses to build his theory. Most of this will be presented in the form of what appear to be definitions, or necessary and sufficient conditions for the various causal notions which Woodward uses. Following this, I shall critically examine the status of these definitions, questioning whether and how they can shed any light on the two unanswered questions – what *is* causation - and can we provide a theory of meaning for causal statements?

The key definitions, which Woodward gathers together in one big 'theory of causation', Woodward calls 'total cause', 'direct cause' and 'contributing cause'.[4] Spelling out what an intervention is will also require some care; specifically I will answer the charge of circularity which threatens to pull the rug from Woodward's theory.[5] I bring out the difference between two types of circularity - analytic and inferential[6] - and show that Woodward's definitions only suffer from the milder, analytic form. This will lead us into the discussion about metaphysics and meaning. Analytic circularity de-bars a definition from giving an account of causation which is fully reductive and explanatory – but it still allows the definition to be useful, since it gives the application conditions for the concept being defined. I show that if interventionism is only left with application conditions, then it cannot go very far in giving a metaphysics of causation. It does not, for example, as with related agency theories of causation, try to reduce causation to interventions based on the free actions of agents. But it can still provide a theory of meaning, since it can be claimed that what someone means when they say '*X* causes *Y*' is that 'were one to intervene to change the value of *X*, one would also change the value of *Y*'.

2.1 Preliminaries – Directed Acyclic Graphs

[4] Each of these has their analogues in the more traditional literature, and I flag this where appropriate.

[5] The circularity charge arises because, as it turns out, an intervention is a causal notion. One appears to argue in a circle if one defines causation in terms of interventions, which are themselves defined using causal terms.

[6] This distinction is found in Humberstone (1997)

Woodward introduces the notion of a *directed acyclic graph* (DAG) which serves a representative function. It models, by way of nodes and arcs, the causal structure of some worldly system. Nodes represent variables in the real world system; the specification of what a node can represent is broad. Typically, in scientific contexts, a node will be a determinable, and will be able to take a numeric value tracking some quantitative measure. Such a measure could be discrete or continuous, for example acceleration, force, or number of people with some disease. The arcs are a little trickier to specify at the outset, since they represent 'direct' causal links. This should raise an immediate red flag – surely such links represent, if anything does, what is at the heart of causation. Can we really 'take these for granted'? Woodward initially says that the point of direct causation is just that nothing mediates between cause and effect. For example, there are many ways one could bring about the acceleration of some body from rest. One could push directly on the body, one could blast air at it, or one could remove an obstacle from in front of something which would otherwise push it. In the case of a direct push, one might think, nothing mediates the cause-effect pair, unlike for the other scenarios described. The notion of direct cause gets a fuller treatment, to which I return below.

A graph is said to be *acyclic* if there is no way to get from a node, via the direct causal links, back to that same node. This is an important restriction – for it disallows backwards causation.[7] We can define certain features of the graph and specify certain

[7] It is by no means uncommon in the literature to simply stipulate backwards causation out of one's

relationships among the variables/nodes in the graph. A sequence of variables $\{V_1 \ldots V_n\}$ is a *directed path* or *route* from V_1 to V_n iff for all $i(1<i<n)$ there is an arc from V_i to V_{i+1}. *Y* is a *descendent* of *X* if there is a route from *X* to *Y*. In such a case, *X* is said to be an *ancestor* of *Y*. Finally, the direct causes of *X* are said to be the *parents* of *X*.

Each DAG has associated with it a set of *structural equations*[8] which encode information about how the variables would change under interventions on their parents. Such information is counterfactual information, which gives interventionism a distinctly counterfactual flavour. But the equations allow for much richer counterfactual information to be specified that simply 'had *x* not happened, *y* wouldn't have,' since the equations can encode information about mathematical relationships such that we can predict not only that some event will happen, but also that it will happen in a specific way. One thing to note here is that interventionism is principally a theory of causation at the level of types. Variables can take a number of different values, and the structural equations specify the way that each of those variables interacts using something akin to a law based generalisation.[9] The usefulness of these equations becomes apparent when one compares interventionism with the simple counterfactual approach; all sorts of problems concerning things like pre-emption, transitivity, and omissions have beset that view, and as we shall see later

account. By limiting oneself to acyclic graphs, this is in effect what Woodward and others do.

[8] This is not the case in the causal inference literature, at least not at the outset. Rather, a probability distribution is applied to the DAG, and algorithms run to attempt to find the causal links, and the relationships which govern their behaviour.

[9] Woodward calls the relationship 'invariance'. Chapter 6 of the book is devoted to spelling out this notion.

on,[10] some of the solutions that have been proposed in the literature come close to the way in which interventionism, at a stroke, manages to do away with them by simply allowing for this kind of richness.

2.2 Direct Causation

So far we have only roughly characterised interventionism as saying that *X* is a cause of *Y* if there is some intervention on *X* that will change *Y*. When this is the case, Woodward calls *X* a *total cause* of *Y*.[11] But, notoriously, there are some situations for which, though *X* is a cause of *Y*, changing *X* by itself will not always bring about a change in *Y*. For example, it has been discovered that the ingesting of birth control pills is a positive causal factor for thrombosis. However, pregnancy is itself a positive causal factor for thrombosis, and so by preventing pregnancy, the taking of birth control pills is also a negative causal factor for thrombosis, since it prevents something which would otherwise have caused it. We can imagine that the positive causal influence that the pills have *directly* on the chance of getting thrombosis is exactly matched by the negative influence exerted *indirectly* by the prevention of pregnancy. In these kinds of circumstances, making an intervention on the variable concerning the taking of the pills will *not change* the variable concerning the likelihood of getting

[10] Chapter 3

[11] In Reichenbachian terms, *X* is a *prima facie* cause of *Y*.

thrombosis.[12] The notion of *direct* cause is therefore necessary. Woodward (p52) identifies the importance of *distinctness of mechanism* for causal models. The idea is that for any way in which one variable causes change in another variable, there will be a specific mechanism which effects that change. Direct causation captures this idea; something is a direct cause of something else if it works via a *single* mechanism only, and, crucially, we misrepresent reality if we run one or more mechanism together in our model. This is fairly clear in the case of the birth control pills; whilst there is some mechanism which leads from the chemical composition of the pills themselves to some physiological variable affecting the likelihood of thrombosis, there is another *indirect* pathway which leads, via the prevention of pregnancy, to that same variable.[13,14]

The reason why the notion of direct causation (distinctness of mechanism, modularity) is so important is because it allows for a much richer description of causal systems. The problem posed above was that in certain cases, intervening on the putative cause will not effect a change in the putative effect. But we can now see that this objection can be defeated if we are careful to distinguish *direct* cause from what Woodward calls *total* cause. Total causes must, for at least some interventions on them, change their effects. But direct causation is defined more strictly, as follows:

[12] What Spirtes *et al.* (1993) call a failure of *faithfulness*.

[13] Actually, this elides some trickiness. Preventions seem to be precisely *non*-mechanisms.

[14] The equations which govern the causal links should also reflect this distinctness, a thesis that Woodward calls *modularity*. (ibid. p48) This will become important when we consider Strevens' critique of Woodward's programme in chapter 4.

> "A necessary and sufficient condition for *X* to be a direct cause of *Y* with respect to some variable set **V** is that there be a possible intervention on *X* that will change *Y* [...] when all other variables in **V** besides *X* and *Y* are held fixed at some value by interventions." (ibid. p55)

It is clear that this will solve the birth control pills case. Since, were we to hold steady the chance of pregnancy (for example by experimenting on a menopausal woman), by administering the pills we increase the chance of thrombosis.

2.3 Contributing Cause

Woodward notes that a further notion of causation is needed to cover situations for which something can be a cause, but where it is neither a total nor a direct cause. Situations like the Hesslow example come close to this. If we imagine that the positive causal influence that birth control pills exert on thrombosis happens via another variable (lets say because it induces one to sit in crouched positions for long periods), then the taking of the pills can no longer be considered a direct cause of the thrombosis. As stipulated, however, it is also not a total cause since we still have the negative influence exerted via the prevention of pregnancy. But, since taking the pills are clearly causally connected to thrombosis, any theory should have it come out as a cause (at least at the level of types). Thus, Woodward introduces the notion of a contributing cause:

> "If *X* is a contributing type-level cause of *Y* with respect to the variable set **V**, then there is a directed path from *X* to *Y* such that each link in this path is a direct causal relationship; [...]. Put differently, if *X* causes *Y*, then *X* must either be a direct cause of *Y* or there must be a causal chain, each link of which involves a relationship of direct causation, extending from *X* to *Y*." (p57)

Finally, Woodward states his manipulability theory thus:

> "(**M**) A necessary and sufficient condition for *X* to be a (type-level) direct cause of *Y* with respect to a variable set **V** is that there be a possible intervention on *X* that will change *Y* [...] when one holds fixed at some value all other variables Z_i in **V**. A necessary and sufficient condition for *X* to be a (type-level) *contributing cause* of *Y* with respect to variable set **V** is that (i) there be a directed path from *X* to *Y* such that each link in this path is a direct causal relationship [...], and that (ii) there be some intervention on *X* that will change *Y* when all other variables in **V** that are not on this path are fixed at some value. If there is only one path *P* from *X* to *Y* or if the only alternative path from *X* to *Y* besides *P* contains no intermediate variables (i.e. is direct), then *X* is a contributing cause of *Y* as long as there is some intervention on *X* that will change the value of *Y*, for some values of the other variables in **V**."

Though this formulation looks complex, the underlying idea is fairly simple. As Woodward says, it embodies the idea that there is *"No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference"* (ibid. p61). In other words, everything about causation can be captured in terms of what would happen under hypothetical interventions on variables in the system we are interested in. We need the three notions of causation (direct, contributing, total) in order to fully capture the ways in which something can be a cause.

2.4 Contrastive focus

It is also important to note that the interventionist account replaces talk of causation with that of causal relevance. In setting up our models and associated structural equations, we 'establish' the general claim that, say, smoking is relevant, in some way, to lung-cancer. This is important, since it has been noted (especially by Hitchcock, (1995)) that binary causation does not give us the full story about what happened even in a given (token) case. This is because of what Eells (1988) calls the problem of 'disjunctive causal factors'. The problem is basically that, given a simple counterfactual, there are many ways that one can read the antecedent, many ways that we can depart from the actual world to get to the counterfactual world. In quantitative situations, this can be disastrous when trying to assess how to fill in the antecedent. If we consider the level of smoking in a particular country, we can say both that 'smoking a pack a day caused lung cancer rates to increase' and also 'smoking a pack a day caused lung cancer rates to decrease' because it depends on what the background circumstances are. As Hitchcock (ibid. p262) notes, we can imagine a country in which everyone smokes two packs a day. For such a population, smoking (only) a pack a day would reduce the rate of lung-cancer. Saying that 'smoking causes cancer' is then somewhat ambiguous, or smuggles in assumptions about what the background situation is. Causal relevance captures this ambiguity rather well, since it says that smoking and lung cancer co-vary according to some equation, and say not that 'smoking causes cancer', but rather that 'cancer is causally relevant to cancer'. Of course Hitchcock's point really applies at the token level – any given causal claim is relevant to only some backgrounds – only to some ways of filling in what happened in the counterfactual scenario[15]. As we are about to see, actual causation in the interventionist framework does well in capturing this important feature.

2.5 Reproducibility

Another important aspect of the theory is that it assumes that the relationships between the variables are such that they are not only applicable at just a single moment in time. The relationships are invariance relationships which are rather like laws. Just as laws hold over time, so too invariance. The importance of this point comes out in the special way that interventionism evaluates the counterfactuals associated with its causal claims. This reproducibility assumption is what ultimately grounds the interventionist counterfactuals, and it figures prominently in the debate between Lewis and Woodward concerning how to evaluate counterfactuals. Because Lewis reduces causation to counterfactual dependence, he cannot tolerate the notion

[15] This is brought out especially well when considering the stress we can put on different parts of a causal statement. "Boris was arrested because he *stole* the bike" has very different counterfactual implications than does "Boris was arrested because he stole the *bike*". Whilst the former implies that the causation involved runs from Boris' stealing something (not necessarily the bike), the latter implies that what caused the arrest necessarily involved the stealing of the bike (perhaps because it was expensive). Causal claims thus have 'contrastive focus'.

that causal systems are part of a theory of causation. He would no doubt have argued that such a move is circular, and would undermine the reductionism inherent to the theory. He therefore was forced to use a notion of similarity between worlds, and where this similarity goes 'all the way down', without using any causal information to discriminate between worlds. Woodward allows that causal information is allowed to discriminate, and it is this move which allows him to say that the counterfactuals can be assessed relative only to a certain model. This result comes at a cost – the charge of circularity. This charge shall be considered below.

2.6 Actual Causation

Woodward notes in the introduction to his section on actual (or, token) causation that philosophers of science, and those theorists working in the structural equations paradigm have tended not to consider actual causation. Rather than think that this points to some kind of fundamental divide (as Sober (1985) and Eells (1991) do) between token and type causation, Woodward assures us that 'the ideas about causal relationships between variables [...] can be extended to capture important features of token causation.' (ibid. p75) In light of the discussion in chapter 1, it is important to point out that Woodward, even at the level of tokens, thinks that 'we can capture the *content* of many token-causal claims.' (ibid., emphasis added). He thus sets out his stall as attempting to give a theory of *meaning* for both type and token causal statements.

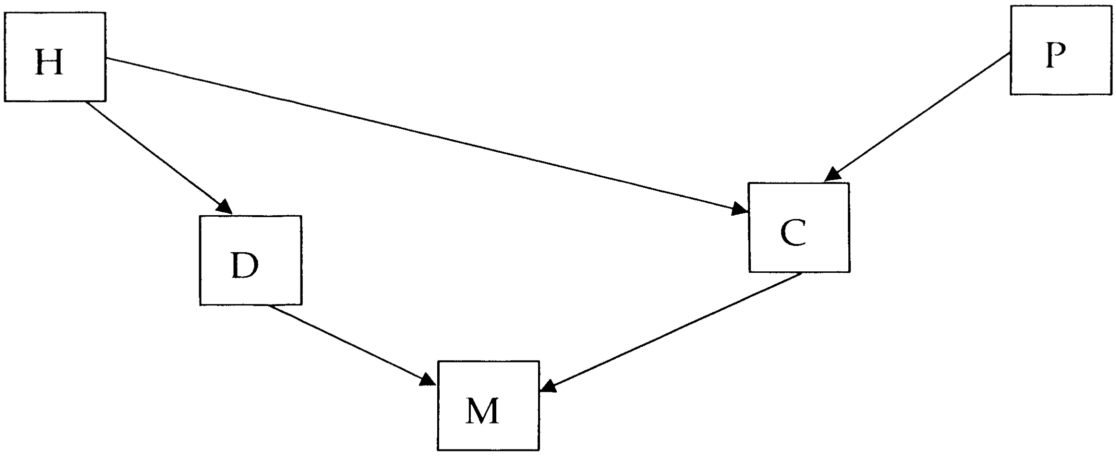
In Woodward's account, token causal claims are parasitic, for their meaning, on their type level parents, whose structural equations capture law-like generalisations between event *types*. Woodward makes this explicit in his rejection of the views of Anscombe (1971) when he says that "...token or singular causal claims should always be understood as committing us to the truth of some type level causal generalizations" (ibid. p72). So for example, in an actual case of a short circuit causing a fire, the causal claim 'has its truth' somehow borrowed from a general causal structure for which there is an invariant relationship between the binary variables 'short circuit' and 'fire'. In the general case, the variables are determinables, in the token case they are those determinables' determinates. Woodward couches it in epistemic terms:

> "Is there a directed path from *X* [putative cause] to *Y* [putative effect] such that some intervention that changes *X* from its actual value would change *Y* from its actual value when (a) other variables along all other directed paths from *X* to *Y* are fixed by interventions at their actual values and (b) other direct causes of *Y* that are not on any directed path from *X* to *Y* remain at their actual values?" (ibid. p76)

This would be a *test* for whether some variable having the actual value it had was the cause of some other variable having the actual value that it had – but we can easily recast the epistemic language into metaphysical language. What we cannot do in making this transition, however, is get rid of mention of the graph itself, including its mention of paths, interventions and direct causes. We will consider the ramifications of this relativization in chapter 4. In order to define actual causation, Woodward uses a token level analogue of his definitions at the level of types:

> "**(AC)**
> (AC1) The actual value of $X = x$ and the actual value of $Y = y$
> (AC2) There is at least one route R from X to Y for which an intervention on X will change the value of Y, given that other direct causes Z_i of Y that are not on this route have been fixed at their actual values.
>
> Then $X = x$ is an actual cause of $Y = y$ iff both conditions (AC1) and (AC2) are satisfied." (ibid. p77)

As an example, Woodward uses the case of the traveller, T, who dies in the desert. The traveller has two enemies, the first, A, fills his water bottle with cyanide, and the second, B, punctures the water bottle, thus depriving the traveller of any liquid that might be in there (unbeknownst to B, he merely drains the bottle of poison). What was the cause of death? We can set up the following causal graph and associated set of equations:



(1) $C = P.-H$

(2) $D = H$

(3) $M = C\mathit{v}D$

In this graph, each of the variables are binary, taking the value 'true' or 'false'. P represents whether A put poison in T's water bottle. H represents whether B punctured a hole in T's water bottle. D represents whether T was dehydrated, C whether T ingested cyanide, and M whether T died. The equations stipulate (1) that ingestion of cyanide by T requires both that B placed the poison in the water bottle and that A did not puncture the bottle, (2) that T dehydrates only if A punctures T's water bottle, and (3) that T dies either if he ingests cyanide or if he dehydrates. Of course we could quarrel with this graph, by claiming that it does not represent reality properly. For example, it could be clamed that T has the ability to detect cyanide, and so would not have drunk it under any circumstances. In such a case, T would die of dehydration *whether or not* A put a hole in the bottle. It would also render (1) false, since the ingestion of cyanide would require more than just that someone placed it in T's bottle. Though there might be legitimate concerns in relation to how to model, for the purpose at hand let us just assume that we have the correct model in place. Now by stipulation, in actuality each of the variables had the value true except C. Though cyanide was placed in T's water bottle, T did not ingest the cyanide. So, claims Woodward, the cause of death was that H punctured a hole in T's bottle. Why is this? Because it meets the two conditions for actual causation.

2.7 What is an intervention? - The 'circularity' objection

So far we have taken the notion of interventions at face value. They are events, broadly construed, which reach in to a causal system to change the value of a variable in a surgical manner, so as not to disrupt anything in the system except the mechanism controlling that very variable. Whilst I think this is an understandable notion (and is also shared by the counterfactual view, though it goes under the moniker of 'miracle' in that thesis), the fact that the idea of change – a causal notion – appears as part of the definiens in our definitions of causation, has left those definitions open to a charge of circularity. Circularity is a charge to be levelled against a definition; if the concept or term being defined turns up in the definiens, this would seem to be problematic. After all, to what extent can we really explain something by reference to some other, more understandable things, if those 'more understandable things' presuppose an understanding of what is to be explained? The matter is a little more complicated, however. Humberstone (1997) considers two types of circularity.

One type involves what he calls 'analytic' circularity, which comes close to the kind of circularity we have just been considering. Analytic circularity involves circularity at the level of *explanation* – it is clearly circular to try to explain what a certain concept involves by reference to that very concept (even if the reference to it is not explicit). Circularity of this kind is often found in definitions of dispositional concepts,[16] but it may not prevent a definition being of at least some use. To see why not, we need to consider the other type of circularity, which Humberstone calls 'inferential'. This type of circularity not only leaves us devoid of understanding, but also without a way of even being able to say when the concept applies. A simple example would be a definition for which the definiens simply repeats what is to be defined[17]. As Humberstone says:

> "Analytical circularity is a fault, then, when and because it obstructs the transfer of understanding an account of the application conditions of a concept may be designed to effect: from understanding of the terms in which the account is couched to understanding the concept being analysed. Inferential circularity, on the other hand, is a fault to the extent that what is obstructed is the transfer, not of understanding, but of knowledge. Here we envisage using the account of the concept's application conditions not so much as a way of getting the concept across to someone not familiar with it, but as a recipe for telling us when it applies: if we had already to know about this before we could

[16] For example, we could define 'beautiful' as those items deemed beautiful by beings in suitable conditions. The term beautiful appears on both sides of the definition.

[17] E.g., "red is red".

> employ the account to that purpose, the recipe would not yield the desired knowledge." (ibid. p251)

'Knowledge' here refers to knowledge of when we have a case of *xyz*, or of when we can say that we have a case of *xyz*. In a sense, then, this is an epistemic notion – a definition which is only analytically circular may not tell us anything concerning the metaphysics of the concept being defined, but it will still allow us knowledge of when the concept applies – it can still give us the *application conditions*.

Can it do anything more? Appeal is sometimes made to *illumination* of one concept by another. By mapping a group of concepts, so the argument goes, we might be able to shed light on each of the concepts, even though none are being reductively defined. In fact this is a fairly common line in relation to the family of nomic concepts that includes, amongst other things, causation, laws of nature, probabilistic dependence, and counterfactual reasoning (see e.g. Carroll (1990)).[18,19] But leaving this idea of illumination aside for the moment, let us now turn to consider Humberstone's

[18] Although it is not the only place where we find such arguments. For example Mark Johnston (1989) says about response dependent concepts that "it may be that sometimes the biconditional of the relevant form which shows the concept to be response-dependent is strictly speaking circular. Circularity would be a vice if our aim were reductive definition. However, our aim is not reductive definition but the exhibition of conceptual connections. In such an endeavour, circularity is a defect only if it implies the triviality of the biconditional." It appears that Johnston is aiming for more than just an epistemic notion – the 'exhibition of conceptual connections' sounds like it is doing more than just that.

[19] Igal Kvart (1986) has defended the view that we cannot even make sense of counterfactuals without recourse to the concept of causation. This argument, if successful, obviously scuppers any attempt to reduce causation to counterfactual dependence.

distinction in relation to interventionism. First, we shall have to consider in finer detail what an intervention is, and why it leads to the charge of circularity.

2.8 Interventions in more detail

In chapter three of his book, Woodward spends some time detailing how we should conceive of an intervention. He first of all says that an (actual) intervention is a token level causal notion, which must itself be characterized by reference to an intervention variable, a type level notion. A variable *I* is an intervention variable on *X* with respect to *Y* iff it meets the following conditions (**IV**)[20]:

I1. *X* causes *Y*

I2. *I* acts as a 'switch' for all other variables that cause *X*. This is the 'surgicality', or 'arrow breaking' requirement, such that what value *X* takes is determined solely by *I*.

I3. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y*, and is not a cause of any causes of *Y* which do not pass through *X*.

I4. *I* is independent of any variable *Z*, that causes *Y* and that is on a directed path that does not go through *X*.

Woodward points out that "cause" here means contributing cause, rather than total cause. Then, Woodward defines an intervention as follows:

(**IN**) *I*'s assuming some value $I = z_i$, is an intervention on *X* with respect to *Y* iff *I* is an intervention variable for *X* with respect to *Y* and $I = z_i$ is an actual cause of the value

[20] The following is a paraphrase of Woodward p98

taken by *X*.

We are now in a position to critically assess whether and how these definitions can provide any kind of metaphysics or theory of meaning for causal statements, and crucially to consider the charge of circularity.

2.9 Metaphysics

What is it to give a metaphysical account of something? And how could something like a set of definitions or necessary and sufficient conditions provide such an account? We said in chapter 1 that many options are open to the metaphysician when considering how to give an account of causation. Which, if any, is the interventionist committed to?

It should be fairly clear that interventionism embraces realism over nihilism. It does not try to explain away why we think there is such a thing as causation, or argue that the notion of cause is incoherent or unnecessary. On the question of reductionism, it should be fairly clear by now that interventionism does not have reductionist ambitions. For one thing, the definitions of causation disallow any form of reduction. This is the metaphysical flip side of the (linguistic) circularity charge raised above. A reductive theory aims to spell out what some concept is by reference to other, more basic concepts. That interventionism deals in the currency of circular definitions should by now be obvious. The various definitions of causation all require reference to what would happen under hypothetical interventions. But interventions are themselves spelled out in terms of causation – the very first of the conditions for one variable being an intervener with respect to another variable is that the former is a direct cause of the latter. So the question can be raised, if there is such obvious circularity within the network of interconnecting definitions, in what way can they be said to be at all illuminating?

Woodward, though he does not mention Humberstone explicitly, seems to rely on his ideas about the different notions of circularity. He says, for example,

> "...it is crucially important to understand that [the definitions of an intervention] are not viciously circular in the sense that the characterization of an intervention on *X* with respect to *Y* itself makes reference to the presence or absence of a causal relationship between *X* and *Y*." (p104)

Woodward admits that various pieces of causal information are required, and are 'drawn into' the definition of whether some variable *X* qualifies as an intervener with respect to some other variable *Y*, but he stresses that, crucially, this does not include

> *"information about the presence or absence of a causal relationship between X and Y."* (ibid. p1p5)

Why is this so crucial? Precisely to avoid the more vicious strain of circularity. As it stands, the various definitions of causation and intervention allow that, subject to the system under investigation being stable and providing reproducible results,[21]we can get causal knowledge. We test, using idealized interventions, some system. Should we get the result that a change in one variable leads to a change in another variable, and where certain constraints are in place, we can infer that a causal relationship holds between those two variables. So the theory is not inferentially circular. But the analytic circularity does mean that we are stuck with causation being, in some sense, unanalyzed. The metaphysical flipside of this is that interventionism has nothing to say about what underlies, or what plays the role of 'truthmaker' for causal claims. This, in fact, may be no bad thing. One might think that a certain metaphysical neutrality is actually quite welcome. This will be especially pleasing to those philosophers who think that speculative metaphysics, on offer in many quarters, outruns our epistemic access to such outlandish ontological posits. What reason should we have to believe that the causal relation is reducible to counterfactuals? Rather, it might be argued, our only access to causal truths is via empirical means. The best we can do, on this view, is to spell out as best we can what the various *implications* are in saying that something is causally relevant to something else. We cannot go beyond these impicational or inferential links. As such, then, interventionism can be seen as a realist theory but one which refuses to say anything deep about what causation is. It is best seen as implementing a theory of meaning for causal statements, in that it insists on a certain number of important truths that are

[21] a not insignificant caveat. See chapter 5 for the way in which this reproducibility constraint can produce unwanted results.

connected with the statement that X causes Y. Most important is that, at least under certain circumstances, intervening on X will bring about a change at Y.

We have not yet considered whether interventionism is in some way subjective, or whether it can outrun this charge. The subjectivity involved is based on the idea that the definitions are all made relative to some causal model or DAG. But of course there may be different ways to model a system. Modelling is a human activity, and so there will be some work to do in explaining how to ground a model. That discussion must wait until chapter 4.

Chapter 3 – Interventionism and the counterfactual theory

In order to bring out more clearly what the metaphysical implications of interventionism are, and how it provides a theory of meaning for causal statements, it will be instructive to compare it with a leading competitor – David Lewis' counterfactual theory. There are two important differences between the theories. Firstly, Lewis aims at giving an account of token causation only, whilst interventionism aims principally to capture type level causal claims, and only to capture token causal claims in a derived fashion. Secondly, Lewis tries to reduce causation to something else – counterfactual dependence. Woodward rejects this for one main reason. Since the interventionist definitions of causation make reference to other causal notions – interventions – they cannot be reductive.

This chapter is structured as follows. First I shall give a rough outline of Lewis' theory. Following this, I shall have some brief comments to make on the different metaphysical commitments that each of the theories have. Specifically, I shall show how, although both theories are counterfactual-involving, they evaluate them in very different ways. Following this, I shall present two problems which have dogged Lewis' account – transitivity and omissions. In showing how Lewis and his followers have attempted to patch up the theory in order to solve these problems, I shall also demonstrate how interventionism deals with each of these. The discussion of causation by omission will bring out two further important points. Firstly, the metaphysical neutrality of interventionism allows it the luxury of ignoring the thorny issue of negative events. Secondly, we shall introduce a mild perspectivalism to the theory. One problem with omissions is that it is rather difficult to demarcate which omissions should count as causes. The lack of all sorts of wacky scenarios and events could be said to be among the causes of many other things, yet if we want to discount them, it should be required that this be done in a principled, rather than ad-hoc, manner.

Finally, it will become apparent that interventionism's ability to call on a rich representative structure like a DAG allows it the flexibility to get the right results in each of these types of situation. I shall explain why the assuming of this kind of apparatus, though extravagant, still allows the notion of cause to be of use, because interventionism is first and foremost a theory of meaning for causal statements.

3.1 Causation and counterfactuals

We can summarize Lewis' basic position thus: "*C* causes *E* iff both *C* and *E* occur, and if the following counterfactual is true: if *C* had not occurred, then *E* would not have occurred."

As we noted in chapter 1, this idea is intuitive. Often, in causal discourse, we find that the identification of what was the cause of something else relies on assessing what would have happened had the cause not happened. Though Lewis thinks that causation is a relation between events, it is important to note that the relation of counterfactual dependence is a semantic relation. This relation provides the reductive base on which the causal relation rests. The relation that holds between events gets reduced to a counterfactual relation between 'event occurrences'.

Lewis has a further, more general theory about how the semantics of counterfactuals work. Counterfactuals are modal statements – their truth depends on what happens in un-actualised situations. Famously, Lewis deploys realism about possible worlds on order to cash out modal statements. So, counterfactuals have their truth conditions set down by what happens at other possible worlds –but which ones? After all, there are many possible worlds in which the antecedent of the counterfactual is actualised. Lewis uses the notion of similarity between worlds to decide which is the correct world, or set of worlds, to determine this. This similarity/closeness relation is tricky to spell out, and I shall not go into the details here. Suffice it to say that Lewis employs a ranking of factors to decide this question. For a world to be similar to ours it must not have "big, widespread, diverse violations" (Lewis 1979) of the laws of nature which hold here in our world. Secondly, match of matters of fact also must not diverge too much. In order to assess the truth of a counterfactual we must consider a world which has (almost) exactly the same laws as the actual world, and where the particular matters of fact match exactly until the point of departure – the possible world diverges from our world in the most minimal way possible to make the antecedent true. How does the antecedent come about if the laws are identical? Lewis uses the notion of a 'miracle' - this just means that viewed from *our* world, the diversion required to make the antecedent true is to be viewed as a miracle, because it violates the laws of *our* world. In the possible world where this occurs, it happens because the laws are ever so slightly different, and allow for this one diversion – the first of its kind in the history of that world. We can see immediately the similarities between Lewis' miracles and Woodward's interventions. But whereas Lewis requires us to condition on (or, hold fixed) the entire history of the world up until the time of the antecedent of the counterfactual, Woodward requires only that we intervene into a causal system which is already assumed to exist, and where the causal relationships are assumed to hold in a reproducible way – that is, making the same identical experimental interventions will always result in each of the variables taking the same value. The counterfactuals employed by interventionism are not threatened by irrelevant happenings outside that causal system, unlike Lewis'.

For Lewis, since differences in events which happen before the antecedent takes place – even those which have nothing to do with the particular cause-effect pair in question – take the possible world in which those different events take place further away than a world in which *only* the antecedent is made true, Lewis ends up disallowing all antecedent strengthening as an inference pattern. For example, in Lewisian semantics, if I say about a plane which crashed yesterday that "had I got on that plane I would have died", and this is true, I cannot assume that it is true that "had I got on that plane and someone in Australia had sneezed, I would have died."[1] For Woodward, since we are not required to condition on the entire history of the world, we are entitled to make such inferences, because only those factors affecting the plane crash and my getting on it have any bearing on the counterfactual. The counterfactuals are only assessed relative to a particular causal structure. If the model does not contain information about what would happen if a man in Australia sneezed, then this has no bearing on the counterfactual - and therefore causal - truth. Lewis was forced to accept this consequence, since in the reduction of causal to non-causal facts, we cannot discriminate which particular matters of fact will be relevant, since that is *causal* information. No doubt Lewis would charge Woodward with assuming what had to be defined – that the definition of causation is circular. We have briefly

[1] This latter claim might be true independently, however. The point is that one cannot derive its truth from the simpler non-sneeze involving world.

considered why this charge does not stick[2]. Still, Lewis' programme, designed to purge any kind of natural necessity from his ontology, would not tolerate non-reduced causal facts, and so he is unable to use causal information to discriminate between which matters of fact affect the closeness relation and which do not.

3.2 Transitivity

It is important to note that Lewis takes causation to be the ancestral of counterfactual dependency. A chain of dependent events implies causation, even if it is the case that a certain event would have taken place without the purported cause. This means that Lewis is committed to the *transitivity* of causation. If *A* causes *B*, and *B* causes *C*, then it must be the case that *A* causes *C*. This is because there will necessarily be a chain of counterfactual dependency which holds between *A* and *C*. This causes trouble for Lewis, since a number of counter-examples to transitivity have been presented in the literature. McDermott (1995) provides the following example. Imagine a right handed terrorist is going to detonate a bomb, using his right hand. Just before he manages to do this, a dog bites his right hand, severely injuring it. So, the terrorist detonates the bomb with his left hand instead. But now, it looks as if we

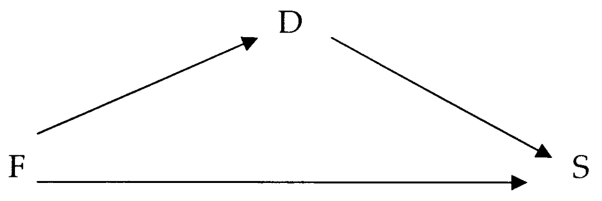
[2] Since the definitions are only *analytically* circular.

have the following chain of dependency. The dog caused the terrorist to detonate the bomb with his left hand, and the detonation caused the explosion. But it seems odd to say that the dog caused the explosion, especially as the dog bite would seem to be a hindrance to the terrorist. Lewis ends up biting the transitivity bullet. Even in his refined theory as laid down in his (2004a) Lewis says,

> "In rejecting the counterexamples [...] I think I am doing what historians do. They trace causal chains and, [...] they conclude that what comes at the end of the chain was caused by what went before. [...] And every historian knows that actions often have unintended and unwanted consequences. It would be perfectly ordinary for a move [...] to backfire disastrously." (p195 of Collins *et al.*)

I think we can do better than this. The many counterexamples to transitivity suggest that causation is, at least sometimes, intransitive. But at the same time, there are plenty of ordinary cases in which we would want to say that causation is transitive. If I use a snooker cue to hit a ball *A*, which hits another ball *B*, and ball *B* ends up dropping into the corner pocket, it just seems obvious that the my action of thrusting the snooker cue caused ball *B* to drop into the pocket. Interventionism is able to cope with both kinds of cases.

We have seen already how interventionism allows for transitivity, for the kinds of cases where we want it. The relation of contributing cause captures this idea. If there is a path through a DAG from one variable to another, then we can demonstrate that intervening on the causally upstream variable to change it will have an effect on variables at the other end of the path. Even if there are intermediate variables, this does not matter. But how do we block the 'unwanted' transitive cases? This all rests on how the model is set up. We can use another example, this time from Ned Hall. Imagine a boulder falls from a ledge, causing a hiker to duck. If he had not ducked, he would not have survived. So although the boulder falling caused the hiker to duck, and furthermore this ducking activity caused the hiker's continued survival, we do not want to say that the boulder caused the hiker to survive. We can model this scenario as follows:[3]



[3] This handling of the case is drawn from Woodward p79

Here, F D and S are each binary variables representing, respectively, whether the boulder falls, whether the hiker ducks and whether or not the hiker survives. The structural equations are as follows:

D = F (If the boulder falls, then the hiker ducks)

S = ~F *v* D (Survival happens if either the boulder does not fall, or if the hiker ducks)

The above picture is a type level graph, with the structural equations capturing the type level relationships between variables. We can see from this graph that, even if we do not want to say that the falling boulder is a cause of continued survival *in the actual case,* nevertheless the falling boulder is causally relevant to the continued survival of the hiker. This must be true, since if the hiker is going to die, it will be because the boulder hits him. Boulder falls also cause duckings, though, so we must represent, using a separate arrow, this piece of causal information. Finally, we also must represent the fact that ducking does cause continued survival. So in the actual case, each of the variables takes a value. F = true, D = true, and S = true. Using our definitions of actual causation from the previous chapter, we can assess whether the falling boulder was a cause of the continued survival of the hiker. Condition (AC1) is satisfied – both the cause and effect variables are true. What about (AC2)? (AC2) requires that we intervene on the putative cause to see whether we get a change in the effect variable of interest.

The influence of this variable on S can be along any route, but, crucially, we hold fixed the values of any off-path variables at their actual values when we do this. So we have a path that exists that goes directly from F to S. In order to test whether this path is effective, we must hold fixed F at its actual value (True). Because the hiker always survives as long as he is ducking, changing the value of F makes no difference. What about the other route through the graph, via D? The test cannot be completed, since we have no intermediate variables along the F-S route to hold fixed. So as expected, F is not, in the *actual* situation, a cause of S. One might object that, although we have not represented explicitly on the graph any intermediate variables along the F-S route, these surely do exist. This is because the boulder traces a spatio—temporal path from the cliff-side to hiker – why is it ok to omit the 'intermediate' states of the boulder? This provides a good example of the idea which Woodward frequently refers to, that of only representing *serious possibilities*. The idea here is that we must consider the way that the falling boulder and hiker's ducking interact as well as how the boulder might kill the hiker. If we interpolate one (or more than one) variable between F and S to represent the path traced by the boulder, we must be careful to include the ducking which any of these intermediate positions would cause. As Woodward notes (p81), if we posit the boulder at any point prior to which the hiker will notice and have time to duck, then we will just be left with the same causal structure as before. The only way interpolating an additional node will make a difference is if it represents the boulder appearing close enough to Hiker so that he will not have time to duck. But then, in order to test the causal route from F via D to S, we will end up with rather a strange counterfactual. We need to hold fixed the 'fact' that the boulder appeared very close to Hiker's head, close enough so that he cannot notice it. Then we have to intervene to change the value of F, to see whether that makes a difference to Hiker's survival. Very well, it clearly will not. But this is not a situation we should really be recognising as possible. For how can the boulder not fall, yet appear out of the blue very close to the Hiker's head? This is one way in which causal modelling is constrained by what is possible. The required notion of possibility here is somewhat flexible, but it will depend on the features of the case. When we are trying to model a real world system, it is a given that *nomic* possibility is what must be in play. We do not recognise events which are contrary to the laws of nature.

Other transitivity cases are solved not by considering the causal graph, but rather the structural equations which accompany them. McDermott's example is handled in this way. The example involved a case in which the dog bites the hand of the terrorist, who then switches the hand with which he detonates the bomb. The interventionist blocks the transitivity, denying that the dog bite was the cause of the blast, because the structural equation representing the connection between the dog bite and the detonation only allows that it will change the *way* in which the detonation will occur. It cannot prevent that event coming about *altogether*. This richness of causal representation is a powerful tool. One important feature is the way in which interventionism is not forced to take a stance on what the nodes represent. In other words, interventionism is not metaphysically committed to any particular picture of what the causal relata are. It simply glides over this metaphysical messiness, since the definitions only talk in abstract terms about the relationships among variables on the graph. As long as the definitions 'go through' and get the right result, that is all that matters. Lewis' followers have to make all sorts of twists and turns to escape the transitivity problem. L. A. Paul (2003), for example, claims that problematic cases of transitivity force us to rethink the metaphysics of causal relata, so that it is *aspects* of events, rather than events themselves which play this role. This positing of rather unusual entities might be successful in accounting for the different cases we have considered, but it appears to be a rather clunky and ad-hoc way out. By putting the modeller 'in charge', interventionism can bypass all of these cases. That does not mean that anything goes, as we have seen there are restrictions based on appropriate notions of possibility. It is also imperative that we model accurately, a point which will be made salient in responding to a major criticism from Michael Strevens, in chapter 4.

3.3 Omissions

Causation of and by omissions has been one of the thorniest issues in the metaphysics of causation in the past few years. The problems stem from the fact that omissions don't seem to be things in the world that have the required features to do the 'pushings and pullings' required to be the truth-makers of causal claims. Yet, there doesn't seem to be anything obviously wrong with statements asserting that the lack or absence of something caused something else or that something caused the absence of something else. We frequently make such assertions, and it would require quite a strong error theory, postulating a serious reengineering of our words, in order simply to make sense of these kinds of claims. I will consider one such attempt.

The problem of omissions has been a major motivation for the idea that it is *facts* rather than *events* which should be considered the causal relata. This is the strategy adopted by D. H. Mellor (1995). This strategy allows for a unified account of causal relata, but at the expense of the seemingly counter-intuitive position of having rather dubious entities play the role of the causal relata.

Only by taking Mellor's route can one short-circuit the problem of omissions entirely, and those within the event causation mainstream need other strategies.

I discuss two - David Lewis' and Helen Beebee's. Beebee (2004) builds on seminal work by Davidson, using his causation/causal explanation divide to explain causation by omission statements. Davidson noted that our causal locutions sometimes seem to pick out events, and sometimes facts. But it seemed to him odd to entertain the possibility of there being two concepts of causation; rather what is going on is that people often conflate causation with causal explanation. Causation is a relation that holds between events only, but of course there are facts that obtain (or fail to obtain) which go into the explanation of why the cause caused the effect. If a screw snaps, causing a cabinet to come crashing down, the causal relation holds between those two events, even though we may not know anything about why the screw snapped. It seems natural to say, "The fact that the screw was warped caused the crash" but this is a conflation on Davidson's view, since facts don't cause things. The warped nature of the screw helps us to understand why the screw snapped, but it was the screw snapping as an event in a region of space and time which caused the crash.

There is a good reason why such conflation exists. As human beings, we seek explanations that are most salient in order to make good sense of what goes on in the world. So it comes as no surprise that if there is a feature of an event which is most explanatorily relevant, we will simply assert that it was that feature which *was* the cause. One has to know quite a bit of philosophy in order to realise that it is very counter-intuitive to say that facts are causes.

A natural extension of the Davidsonian picture is to say that the same sort of thing is going on with omissions. Beebee marks a difference between *relationists* and *non-relationists* about causation. The relationists hold on to the network (neuron diagram) model that David Lewis introduced, and insists that the causal relation can only hold between events represented as neurons in the diagram. The question about absences within this picture is then of course the same question about whether there are negative events.

Beebee herself wants to deny that such events exist, and so she relegates causation by omission to explanations, claiming that causal statements cast in event form are strictly speaking wrong, and are elliptical for some kind of explanation.[4] There doesn't seem to be anything particularly implausible about this suggestion, given that Davidson showed us how easy it is to conflate causation with explanation even in the regular case. As spelled out above,

[4] This is the 'reengineering of our words' adverted to above.

although our everyday locutions often pick out facts as causal relata, this is a mistake, although a very natural one.

Evidence for Beebee's position is provided for by a powerful argument concerning the difficulty involved in deciding which things to rule in and which to rule out in cases of causation by omission. For example, it is plausible to claim that my failing to water my plant whilst away on holiday caused it to die – but Mick Jagger's failure to water it clearly did not. It looks like norms (or expectability) will be the deciding factor here, yet norms don't appear to have anything to do with causation – how could they? Rather, she claims that in picking out some omissions and not others, we pick out explanatorily salient features of the environment that, had things been different, would have prevented whatever it was that actually occurred from happening. This is to provide some kind of modal information as part of the explanation. Such explanations can be sensitive to norms, and hence we can understand the very natural way in which people rule out certain omissions from the causal story.

Lewis takes a different tack. He discusses causation by omission in "Postscripts to Causation", but overhauls his thinking in one of his last papers on the subject – "Void and Object" (2004b). Lewis fundamentally disagrees with Beebee about whether omissions or voids can cause things. He raises the question of the missing relatum and states that:

> "We could deny [...] that absences ever cause anything. (Likewise, we could deny that anything ever causes an absence. In other words, we could deny that there is any such thing as prevention.) Simply to state this response is to complete the reductio against it." (p281 of Collins *et* al. (2004))

Clearly then, Lewis believes wholeheartedly in causation by and of omissions. In order to escape the problem of the missing relata, Lewis is forced to accept that causation does not always require a relation, but that this is unproblematic since the counterfactual theory doesn't always require one:

> "We do not need to reify the void in order to ask what would have happened if the void had not been there. The void causes death to one who is cast into it because if, instead, he had been surrounded by suitable objects, he would not have died. [...] Absences are spooky things, [...] but absences of absences are no problem. [...]
> The counterfactual analysis escapes the problem [of missing causal relata] because, when the relata go missing, it can do without any causal relata at all." (ibid. p283)

This is all very well, and it is worth noting that only a full reductivist like Lewis could say this – since he reduces causation to counterfactual dependence, what matters is not whether there is anything answering to the relata – that is immaterial – but rather whether the reduction works. And Lewis is surely right in saying that "absences of absences" pose no special problem as far as counterfactual dependence is concerned. The problem that does arise for Lewis, however, is that of explaining why we judge some absences as causes, and others not (like in the plant watering case for example). Lewis deals with this question in his (2004a), in which he is forced to say that, in truth, all of the apparently spurious causes are really causes (such as Mick Jagger's failure to water my plant). It's just that it is inappropriate to say such things. Of course 'appropriateness' will be relative to context, and Lewis gives some fairly wacky examples to demonstrate this point – which I think is well taken.

What does the interventionist have to say about causation by and of absences? Again, we need to view this from the point of view of causal modelling. Is there anything special about variables that can take a range of values including *nothing at all*? – As far as the *model* is concerned, no. Why should it be ruled out? What does matter is which variables ought to be represented in the model in the first place. Things become a little tricky here. We saw above, in relation to one of the problems of transitivity, that far fetched possibilities – possibilities which are not appropriate to represent are uncontroversially not to be included in the model. But plenty of omissions which we may not consider relevant are not far-fetched in this way. Often, a modeller just does not want the added complication of having to represent many irrelevant causes. For example, an airline engineer might be modelling the ways in which mechanical failure might bring down a plane. Now it is of course true that the continued flight of aircraft is 'caused' by the lack of surface to air missiles being launched, say. So does the modeller 'go wrong' in not representing this possibility?

This points, I think, to the fact that often causal claims are *interest-relative*, or *perspectival* in some way. We have already noted that interventionism is best thought of as a theory of meaning for causal statements, and where the theory gives 'truth conditions' based on the outcome of hypothetical interventions on the system under investigation. Now if surface to air missiles were being explicitly represented as a node in the model, it would surely be true that were you to intervene so as to change the value of that variable, that would change the variable representing whether the airplane continues to fly. But, I claim, we can tolerate some context sensitivity. This is not to deny that the omission of surface to air missiles causes the continued flight of airplanes, but rather recognises that causal *claims* are made relative to a particular background. This position is not, in fact, uncommon. But it is important to bring out the way in which, because interventionism has all its definitions rely on a *model,* the particular model chosen will ensure that only certain causes are recognised.

A problem with this idea might be that it makes causation subjective, since modelling is a human activity. This worry will be addressed in the next chapter.

Chapter 4 – Modelling, Realism and Objectivity

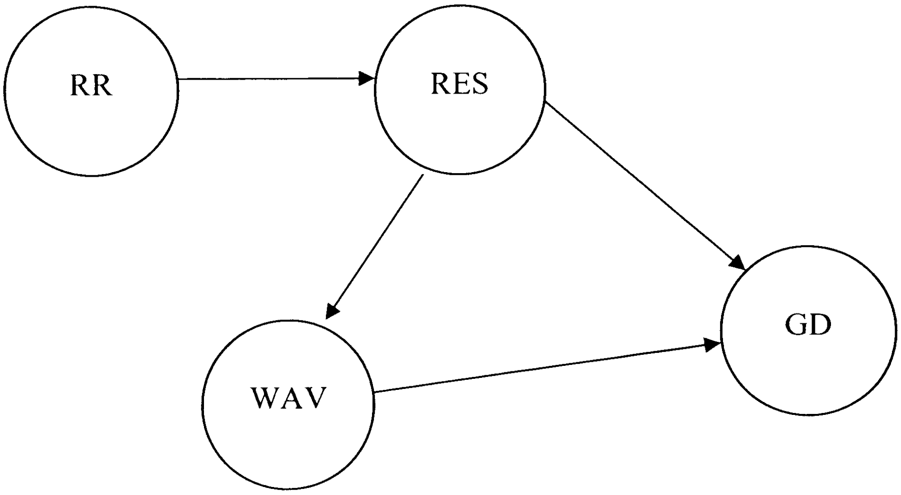
In this chapter I turn to some recent critiques of interventionism. It has been claimed that interventionism fails, at least to some degree, in its attempt at a theory of causation, and for a number of reasons. I shall be investigating two such attacks. The first claims that, on purely technical grounds, the interventionist cannot always model reality in order to get the results that reflect seemingly unassailable intuitions about certain cases. The cases involve pre-emptive causes, but where we cannot hold these fixed because they lie on the actual causal route. Though we judge the actual cause (and not the pre-emptive cause) to be the cause, this result is not supported by interventionism, because were we to hold fixed the pre-emptive cause, changing the actual cause does not produce change in the effect. I will show that the counter-example relies on a certain tension between two ideals of how we model. The assumption made, in presentation of the counter-example, is that differences occurring within a single entity in the world must be modelled by a single node taking different values. But this, I suspect, is false, although I concede that using two nodes on a graph to represent different states of a single item might present a significant problem when it comes to drawing arcs. This is because there may be *logical* connections between nodes on the graph, and so far we have seen that arcs represent only *causal* links. I suspect, however, that with some even more careful attention to the modelling techniques we ought to employ, we might be able to get away without the need to have logically related entities on one DAG.

The second critique calls into question how realist and objective interventionism can be, given that the definitions of causation are relative to a variable set. Now to say that one is a realist is, in Wright's words, to do nothing more than to 'clear [one's] throat' ("Truth & Objectivity, p1), since that term covers a myriad of positions in many different areas of philosophy. It will be easier to bring into focus exactly what the worry is once we look at some cases, but the rough idea seems to be that, whilst pre-theoretically we tend to think that a realist/objectivist about causation should hold that *the world* should ultimately 'decide' whether *X* is a cause of *Y*, interventionism seems to suggest otherwise. If the *truth* of causal statements is always relative to some model, and where this model might not be representing *all* of causal reality, then a fully realist/objectivist notion of causation would seem not to be possible. Now we have already seen that interventionism is little bothered about what the metaphysical underpinnings of causation are. This was touted as something of a strength, but only as long as it was tacitly assumed that there was some mind-independent entity which played this role. Interventionism can be seen more as a theory of meaning for causal locutions than as a theory of metaphysical underpinning. The charge of relativism is thus a serious one, because it calls into question this mind-independent, objective nature of the causal relation.

4.1.1 Modelling - Nodes

In a recent review of Woodward's book, Michael Strevens presents a challenge to the interventionist programme via the use of a case of pre-emption. In the example, Strevens asks us to imagine that two rebels, members of a gang intent on killing a general, have set up an ambush. The first rebel, "Waverer" is instructed to detonate a roadside bomb when the General's motorcade passes. The second rebel, "Resolute" is instructed to launch a rocket at the General, should Waverer not manage to detonate the bomb. We assume determinism, so that should either the bomb or the rocket launcher be detonated, the general is sure to die. Also, should Waverer not detonate, Resolute surely will. Now as it turns out, Waverer does detonate the bomb, killing General. This looks like a typical case of pre-emption, of a kind with the cases we considered in chapter 3. Interventionism demands that we hold fixed any 'off path' variable fixed at their actual values, and then change the actual cause to assess whether the effect variable in question changes. So, by holding fixed Resolute's action, we correctly conclude that Waverer's action was the cause of the General's death. But Strevens introduces a twist. What if we can rig it so that both Waverer and

Resolute lie on a *single* causal pathway? Strevens tells the following story. Suppose that Radio Rebellion issues the order to attack General. Resolute, on hearing the order, and knowing that Waverer might indeed waver, shows herself. Waverer, on seeing Resolute, reasons that the General will die whether or not he detonates his bomb, and so decides to go ahead and detonate. So there are two counterfactual dependencies in play. Had Resolute not heard the order, she would not have shown herself. And had Waverer not seen Resolute ready with her rocket launcher, he would not have detonated. But then, since elements of the alternate pathway (Resolute and her rocket launcher) lie on the actual causal pathway (Waverer and his detonator) it turns out that, were we to hold fixed, very broadly speaking, *what Resolute did*, and only consider Waverer, then changing the behaviour of Waverer will not bring about a change in whether the General died, because Waverer only detonated because of Resolute's action in showing herself and her rocket launcher. It is not entirely clear how Strevens imagines one ought to be modelling such a case, but I suggest the following. Radio Rebellion is modelled as a binary variable, RR, representing whether or not it issues the order to attack. Another variable, Wav, represents whether or not Waverer detonates his bomb. GD, a third variable, represents whether General dies. Finally, there is a sort of hybrid variable, RES, which represents a number of possibilities for Resolute's actions. It can represent Resolute showing herself to Waverer (or not), and can also represent Waverer launching her rocket (or not).

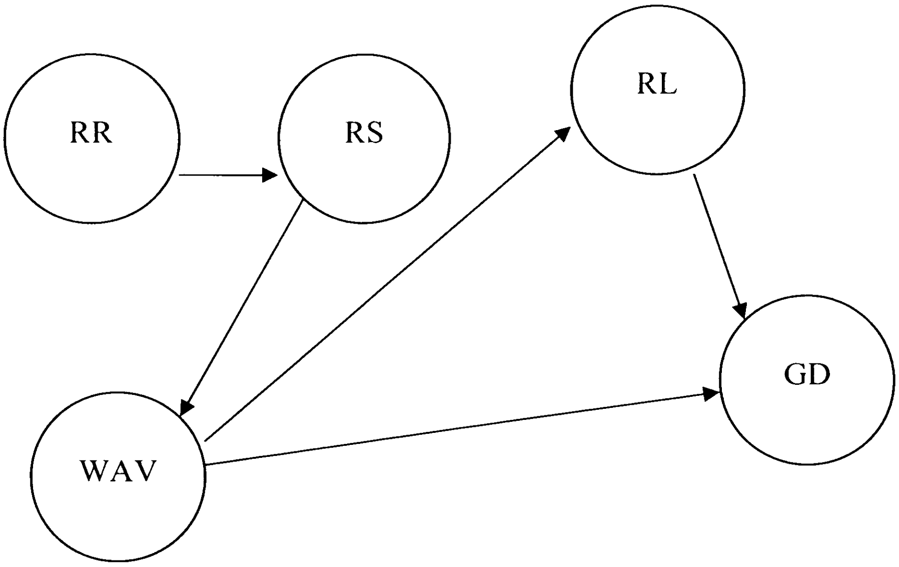


It is unclear, to say the least, how we are to construct the structural equations in such a case. We could try to write something like:D = WAV *v* RES ("The General dies if either Waverer or Resolute act")

WAV = RES ("Waverer fires if Resolute shows herself")

RES = RR ("Resolute acts if she hears the order over the radio")

But this is problematic, because RES does not here represent a single binary variable. Rather, the graph should look like this:



Here, RR, WAV, and GD represent as they did before, but now RES has been replaced by two further variables, RS, standing for whether Resolute shows herself to Waverer, and RL standing for whether Resloute launches her rocket. The structural equations are as follows:

GD = WAV *v* RL

WAV = RS

RL = ~WAV

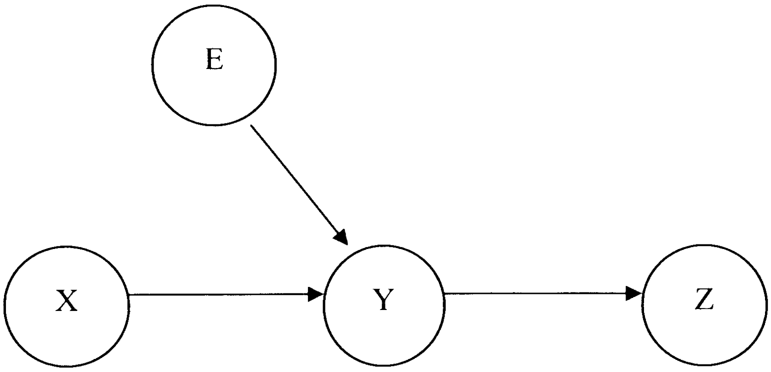
RS = RR

The important lesson here is that, even in the actual circumstance (in which Waverer detonates) though it is true that Resolute does have something to do with the death of the General, this is via something she does that does not involve the alternate causal pathway that leads to the General's death. Strevens complains that actions of Resolute (broadly construed) lie on the actual causal pathway. But it is not as if we should judge otherwise. As stipulated, Waverer *needs* the encouragement of Resolute else he will not detonate. Strevens is trying to have his cake and eat it!

4.1.2 Modelling – Arcs

In a forthcoming symposium on Woodward's book[1], Strevens presents a more sophisticated version of the pre-emption worry. In order to simplify matters, he makes use of a number of switches. X, Y, and Z are each variables. For Z to take some value z, Y must take either of two values, 1 or 2. Y can also be in state 0, where it has no causal consequences. X can be in either of 2 states, 0 or 1. X being in state 1 causes Y to be in state 2, X being in state 0 has no causal consequences. A constraint on the whole system is that electricity is necessary for any of the variables to change value, and we are told that, initially, $\sim(Z = z)$. We are also told that, before the electricity is switched on, X is in state 1 and Y is in state 0. Finally, it is stipulated that Y is unstable in its 0 state when the electricity is switched on, and, absent the causal influence flowing from X, Y would take value 1, causing Z to take value z. Herein lies the redundancy. We can model this scenario, at least at the level of types, as follows:

[1] Strevens' paper is available in draft form at http://www.strevens.org/research/expln/WoodRiposte.pdf



Again it is not obvious how to draw up our equations, because the variables take specific values based on specific values of other variables. The best we can do is to formulate some "if-then" conditions, together with some simple binary variable formulations:[2]

if $Y = \{1 | 2\}$ then $Z = z$ else $\sim(Z = z)$

if $(E \,\&\, X)$ then $Y = 2$ else if $(E \,\&\, \sim X)$ then $Y = 1$ else $Y = 0$

The problem is that in an actual case where the electricity is turned on, two causal routes exist, both of which suffice to bring about the effect we are interested in. Yet, they both go through Y, and again the challenge is that we cannot hold fixed any off-path variable at its actual value, since the alternate path lies along the very same path as the actual cause. Strevens focuses on the causal status of X. Is X a cause of Z? It seems not, since even if were we to intervene to change X from 1 to 0, the value of Z would not change (for any given value of Y).

[2] Since E, X and Z are, in effect, binary.

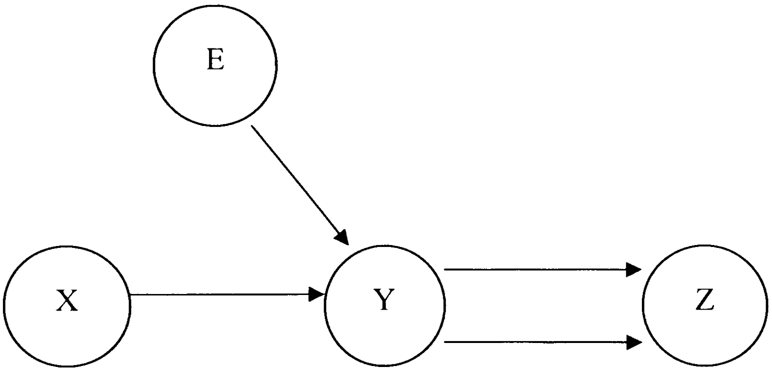
Couldn't the interventionist just bite the bullet here, and admit that X is not causally relevant to Z? When we consider that, ultimately, Z never takes value z unless the electricity is on, and that this is an enabler of Z doing *anything at all*, it might actually seem quite plausible to accept this. What difference does it make, as far as Z is concerned, whether Y is 1 or 2? X fails to be a cause of Y, in other words, because one cannot bring about change in Z by changing X. Strevens counters this move by suggesting that there might be different ways in which Z happens:

> "...stipulate that Y's being 1 causes [$Z = z$] in a very different way from Y's being 2 – perhaps because different fundamental laws apply in each case." (ibid. p3)

As an example, Strevens borrows from a case in which a vase is falling from a rooftop, and is then smashed by someone with a baseball bat on its way down. There is a 'result' of the vase getting smashed, but it seems importantly different how it comes about.

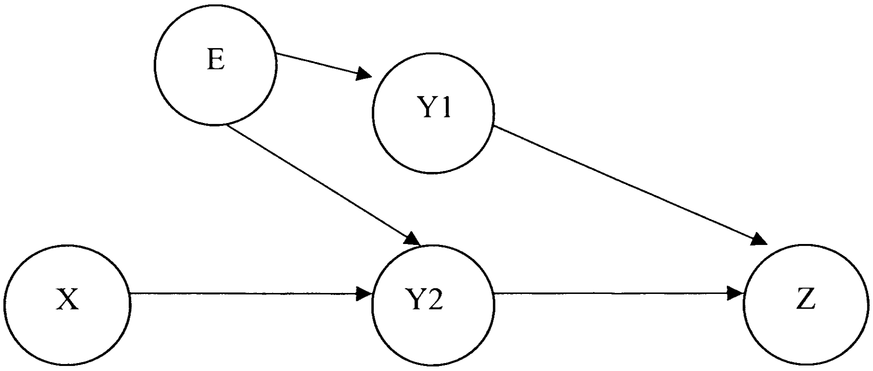
I think the correct response to this is just to point out that Strevens seems again to want it both ways. We briefly mentioned in chapter 2 the thesis, endorsed by Woodward and others, of *modularity*. Modularity is a constraint on how to model a system, and it stipulates that for each mechanism that connects two variables/nodes, a separate arc must be drawn. When considering the birth control pills and its effect on thrombosis, we were obliged, on pain of misrepresenting reality, to draw separate routes on the graph representing the direct chemical route and the indirect pregnancy-preventing route. We shall soon have a chance to expound further on this condition, but for now it is clear that the different ways a cause can act in order to bring about some effect need to be represented with separate arcs and separate structural equations. This is all applicable to Strevens' more sophisticated counter-example. Strevens must choose – if it is true that the *way* that $Z = z$ is brought about differs depending on whether Y is 1 or 2, then these different ways must be represented by different arcs on the graph. If there is only one way that the effect is brought about, then X will not be a cause after all.

But this is problematic, since we have not seen any development so far in the causal modelling literature which allows one to draw more than one arrow between the same two nodes. And even if we could do that, it would not solve the present problem, since we would end up with a graph looking like this:



The problem remains – since we have a bottleneck through *Y*, holding fixed *Y* renders any change in variables upstream from *Y* causally impotent.

The solution requires that we separate *Y* into two further variables, as follows:



Here, Y1 is a (binary) variable which can only represent whether *Y* takes value 1 or not, similarly Y2 represents whether or not *Y* takes value 2. We now have the two required routes through the DAG, and as such we can hold fixed Y1 at its actual value (False) and then intervene to change X to see what happens at Z. And indeed, we now have the required relationship holding between X and Z.

One complication arises, however. Y1 and Y2 are not independent of each other. In fact there is a rather strong *logical* relation which holds between them. *Y* can only be in one state at a time. We do not draw arrows between Y1 and Y2 because there is no *causal* relationship which holds between them. But as long as we are clear on the nature of this relationship, I think it is possible to have both nodes on the graph. One must always bear in mind that, when setting the value of one of these variables, one constrains what values the other variable may take. One could, of course, draw different arcs between logically related nodes, demonstrating their logical dependence.[3] Extending interventionist modelling in this way would be useful, providing another tool for the modeller to represent reality in more ways.

Another possibility[4] is that, given that there is a fundamental difference between type and token level DAGs, we can accept that two *seemingly* logically related variables can be represented at the level of types, as follows. Modularity, we

[3] An additional point to bear in mind is that a graph whose nodes are connected other than by causal links may not comply with the causal Markov condition (CMC). As discussed in chapter 2, the CMC dictates that any node is rendered independent of its non-descendants, when we condition on its parents. But this cannot be guaranteed if there are *logically* related nodes on the graph. The debate about the CMC mostly concerns whether and how we should use it when we are trying to find causal links from some statistical data – that is, for causal inference. If we have some algorithm that assumes the CMC, and where two nodes are logically related, the algorithm will insist that the two are causally linked, since conditioning on either of their parents will not render them independent of one another. But when we are not doing causal inference, but are rather concerned with the meaning of causal statements, we might be able to accept logically related variables in a DAG. This is an area of research which requires a great deal more work and is beyond the scope of this thesis.

[4] Suggested to me by Rory Madden

said, requires us to separate the ways that some effect can be brought about. In this example, Strevens stipulates that there are two different ways that the effect we are interested in can be brought about. Lets call Y being in state 1 a 'unitronizating' state, and it being in state 2 a 'duovalecy' state. Unitronization and duovalency both bring about the effect $Z = z$. We can model each of these states as binary variables at the level of *types*,[5] since there is no restriction on how many entities exist to cause the effect $Z = z$. Of course by stipulation, in the *actual* case, there is only one entity, Y, which can only take one value, but this does not prevent us from implementing essentially the same solution to regular pre-emption, as discussed in the previous chapter. This solution requires a little more work to make it work at the token level – indeed perhaps the force of the counter-example is precisely that, in the actual scenario, since Y can only take one value at a time, and since it is the only candidate for bringing about the effect of interest, there is indeed a back-up cause which we cannot hold fixed. This counter-example does need a little more scrutiny than I can give it here.

4.2 Relativism and Objectivity

In both Strevens' original and forthcoming critiques, the spectre of relativism is raised. The particular form of the charge is framed so that contradictory results

[5] We can ask of a given scenario, "is there any unitronization?" and "is there any duovalency?"

are arrived at, depending on the variable set that is used. Since interventionist definitions of causation are relative to a variable set, this can occur when we move between different levels of analysis.

In fact, we have already seen one way in which interventionism has to tread quite carefully with regard to an aspect of relativism. In chapter 3 we saw that the problem of omissions requires any theorist about causation to set out criteria for ruling in the 'right' and ruling out the 'wrong' sorts of omissions as causes.[6] We saw that Woodward uses a plausible strategy whereby only serious possibilities get recognized. In the case of the hiker and the falling boulder, for example, we do not consider the lack of the boulder from the hiker's head as a serious possibility. We do not imagine the boulder just appearing, ex-nihilo, as it were, and claim that the lack of such an event was the cause of hiker's continued survival. We may begin to put pressure on this strategy when considering more prosaic cases. If an aircraft engineer is trying to model what will happen in various emergency scenarios involving technical malfunctions of the plane, it would just be annoying to explicitly model the launching (or not) of surface to air missiles by a terrorist group. But it is not as if terror attacks havn't been the

[6] Or at least we saw the demand for an explanation of why we *judge* this to be so – remember that Lewis ultimately rules any and all omissions in, so long as there is counterfactual dependence between the omission and the effect. He then explains away the absurd cases as being merely inappropriate things to *say*.

cause of a significant number of plane losses,[7] and they are certainly scenarios to be taken seriously[8]. Rather, the engineers are just not interested in modelling this kind of emergency. But if the non-existence of surface to air missiles really is a cause of the continued flight of aircraft, don't the engineers construct a 'wrong' model if they leave them out? So it is not only a matter of which things we take seriously, but we will also need to justify a kind of 'interest relativity' with regard to how we model. This will be the focus of some discussion below, but first we must consider Strevens' attack as a prelude to these elucidations.

Strevens' charge of relativism is intermingled with that of circularity, which we looked at in chapter 2. It will therefore be necessary to consider what we said there, namely that although Woodward's definitions are circular, this is only in the analytic sense. Inferential circularity, a more vicious form, would indeed pull the rug from Woodward's project, because it would not even give application conditions for when something was a cause of something else. Strevens begins by noting that contributing causation is a relativistic notion, since the definition of contributing causation is made with respect to a variable set. As a reminder, this said that

[7] And that, therefore, the lack of such attacks can be said to be a cause of the continued flight of a plane.

[8] Not just because of their devastating effects, but also because they are not uncommon, relatively speaking.

> "If *X* is a contributing type-level cause of *Y* with respect to the variable set **V**, then there is a directed path from *X* to *Y* such that each link in this path is a direct causal relationship; [...]." (Woodward, p57)

Strevens would be happy to accept this if the notion of an intervention were not also relativized in this way. But he claims, legitimately, that it is, since a condition for one variable being an intervention with respect to another variable was that it had to satisfy the following condition:

> "I3. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y*, and is not a cause of any causes of *Y* which do not pass through *X*." (ibid. p98)

References to cause in this definition mean contributing cause, and so we have an implicit relativization of an intervention to, presumably, the same variable set that contributing causes are themselves relativized to. Strevens does not put it in these terms, but it looks as if we might end up in inferential circularity. In order to even know when to apply the concept of (contributing) causation, one has to be able to say what an intervention is. But in order to do *that*, one must know what it is for something to *be* an intervention. If one of the conditions for being an intervention is that it is not a contributing cause via some other route than the one we are interested in, then this looks viciously circular. In an apparent backtrack, then, Woodward de-relativizes the notion of contributing cause. Strevens quotes Woodward as saying that this can be done with a new, relaxed, definition. *X* is a contributing cause of *Y* as long as

> "there exists *some* variable set relative to which X is a relativized cause of Y" (Strevens forthcoming p5, emphasis added).

We retain, merely as a means to an end, the notion of relativized causation. This new definition will de-relativize causation, but at a cost. In order to find a counter-example, one needs to find two different variable sets, both of which represent the same causal system, but one for which *X* counts as a cause of *Y*, and another where it does not. The problem will be particularly acute if we can find a rich variable set for which *X* no longer counts as a cause, where previously, using an impoverished set, it was a cause. Woodward's response is then to insist that causation is monotonic – if *X* is a cause of *Y* with respect to some variable set, then it will remain a cause under any variable set which is richer, given that the new variable set is a *better* representer of reality. This is surely what we want to keep hold of if causation is objective. We can tolerate, in our approach to modeling some real world system, the elision of certain causal information. We can, for example, ignore certain factors which are not particularly relevant – as with certain types of omissions. So, we could tolerate the loss of certain causal links on a graph in moving from a richer set of variables to a more impoverished set. What appears intolerable is the loss of causal links in moving in the other direction. Strevens' counterexamples exploit this, and he tries to show that causation is thus non-monotonic.

Suppose we are interested in investigating the links between bottled water consumption and heart disease. It might be the case that the only possible way of intervening on the level of bottled water consumption is by increasing the level of salty food consumption. Salty food, though, causes heart disease, and so if we do not explicitly model salty food consumption, and only use it as a means to manipulate the level of bottled water consumed, it will turn out that bottled water consumption counts as a cause of heart disease. Strevens immediately acknowledges that intervening on bottled water consumption by adjusting salty food intake is an inept method, since there is an independent causal pathway leading from salty food consumption to heard disease. His point is just that, when we move to a richer variable set – one that includes salty food consumption explicitly (and where we do not make inept interventions) – the apparent causal link between bottled water consumption and heard disease disappears. This proves, he claims, that causation is non-monotonic because even though the intervention used in the impoverished variable set is an inept one, and thus not something which can qualify as an intervention at all, this ineptitude only shows up when we have the larger set, since we can then *see* that the variable *actually* intervened on is one which has its own causal ramifications on heart disease, and further that there is no link between water consumption and heart disease.

I think a number of points are in order here. First of all, if I understand Strevens correctly, his assumption about what counts as an intervention does not depend only on the causal system under investigation – does not depend, in other words, on the objective status of the variable to be intervened on. Rather, he thinks that an intervention is proper so long as the variable set we have drawn for ourselves, in modeling a given scenario, says that it is. But I think we can quarrel with this idea, even if an intervention is only defined in relation to a given variable set. We must not lose sight of the idea that interventions are *idealized*. It may be true that in a given case we may not have any way of making such an idealized intervention, and must intervene on the variable of interest only indirectly. We can only hope that such a non-ideal intervention will not prove to be inept. But if it is inept, then this ineptitude disqualifies it from counting as a proper intervention. Furthermore, we can always know when we are making a *potentially* inept intervention, since we will always know whether we are changing the variable we are interested in directly, or via some other, potentially misleading, indirect change. And because we will always know this, we should by rights incorporate in the DAG a variable to represent the changes being made on this variable.

Secondly, we need to refocus on what interventionism is really a theory of. We said already that the definitions of causation, because of their circularity, only leave us with 'applicability conditions' for when the concept applies. That is, we have a way of finding out what the causal links connecting a set of variables is, given that we can make idealized interventions (singly and in combination) to see what happens. Though it is true that we cannot always be sure that we have hit upon the correct causal graph, an assumption of the theory is that there must be at least one graph which represents properly the causal system under investigation. Perhaps, then, the notion of causation, in the interventionist school, can best be captured by considering the conceptual role which it plays. This takes us back to an idea we discussed in chapter 2 – also in relation to the circularity worry. Though we did not spell the idea out in very great detail, we said that even if a definition is non-reductive, if it ties together two quite different notions, we can achieve mutual illumination of one concept by the other. The concepts of cause and manipulation via intervention, though they are both fundamentally causal concepts, can nevertheless illuminate each other. As a theory of meaning, this comes out in a bi-directional way. The meaning[9] of the

[9] At least, the *sense* of the term.

term 'cause' can only be understood, on this view, within a close circle of other nomic concepts, none of which have priority or can be grounded. All we have is that, for a given causal system, the causal relationships will reveal themselves when ideal interventions are made.

4.3 Harmless subjectivity

I think we have said enough, in the above paragraphs, to ground the objectivity of causation. Causation is not something mind-dependent or subjective, even if the definitions are relative to a variable set. Using a faulty graph to represent a real world system provides you with faulty results, just as we would expect them to.

However, there is a mild form of subjectivity concerning interventionism and its definitions. Here we come back to the cases for which the modeler is interested in analyzing a system at a particular *level*. If an engineer is modeling the way in which a plane will be caused to move due to the changes in position of its ailerons, stabilizers and rudder, she will not want to model the tiny gravitational effect that these movements will have. This is because they will be dwarfed by the far greater aero-dynamic implications of the movement of these parts. She is then well within her rights to elide these tiny forces, even though that means that part of causal reality is not being represented. So the causal graph that we use to represent the plane will imply that there is no causal effect due to gravity – even though we know, in reality, that such effects are there. One way we can answer this is by utilizing context sensitivity, perhaps similar to the kind of theory employed by Lewis and others in giving an account of knowledge. What counts as knowledge, on this view, is dependent on the context against which the knowledge claim is being made. If Bill and Ben start a conversation, and Ben claims that he knows who the prime minister is, this claim is true, since the background context is one in which the level of evidence required is nominal. But Bill can change the context by 'raising the bar', for example, he could ask Ben, "how do you know that the PM hasn't just been murdered?", or he could raise it even further by asking how Ben even knows that an external world exists. Ben's claim to knowledge might then fall away, unless he can give further justification for his claim to know who the prime minister is. Similarly, we can tolerate context sensitivity in relation to causation. We can challenge the engineer, saying to her, "what about the small gravitational causes? Are you saying that they are not causes?" Against this more detailed background, the engineer will be forced to agree that they are. But just as with the context sensitivity of knowledge, the context sensitivity must always be monotonic. In moving to a causal graph which represents the system more richly, we always conserve the causal links from the more coarse-grained model. The same kind of reasoning can also be applied to causation by omission.

Chapter 5 - Probabilistic Causation and Interventionism

We have now seen the ways in which interventionism can cope with some of the more pressing problems faced by Lewis and his followers. We saw, for example, how interventionism, in its ability to represent causal systems in a richer way than just whether an event happened or not, allows it the flexibility to deal with issues of transitivity. Also important is the way in which interventionist graphs can represent causal systems using nodes and arcs, whereby we need not worry about the potentially dubious metaphysical status of what is being represented. This was particularly useful for causation by omission, for which we did not worry either about the status of the event itself (node), or the mechanism connecting the causal relata (arc). However, maintaining this level of metaphysical neutrality might become problematic when we want to represent real world systems whose causal interactions are probabilistic rather than deterministic. Particularly pressing, as we shall shortly see, is the move from the level of types to the level of tokens. As already stressed, interventionism considers causation at the level of types to be primary; it is here that the causal and counterfactual information is encoded. When we consider the truth of causal statements, we see that cases of token causation derive their truth from their type-level parents. But there is notorious difficulty in linking single case chance with probability of events which we have observed many times (long run frequency).

5.1 Indeterminism and Pseudo-Indeterminism, Tokens and Types

At the outset, we need to make two important distinctions. Firstly, there are two kinds of probabilistic causation. We have cases of genuine indeterminism, whereby the events in question are not governed by laws which lay down how and when those events come about – the laws merely specify that there is some probability that they will. The laws governing radioactive decay, for example, do not determine that such and such conditions must be in place in order for some unstable atom to decay. Rather, the atom only has some probability that it will decay within some time interval. We might wonder whether the concept of causation can even survive such a discovery. (Certainly, so-called 'immanent' causation will begin to look shaky – it just is not true that the state of the atom at time *t* caused it to be in some other state at time *t2*.)

By contrast, what I will call pseudo-intederministic causation involves causal claims where we do not know (or perhaps do not care) the exact mechanism linking cause and effect, and where we therefore can only say, based on *statistical* data (from a population, say) that *A* causes *B*, but where *ex hypothesi* the laws in play are deterministic. A good example is the claim that 'smoking causes lung cancer'. Plausibly, this claim does not say that everyone who smokes will get lung cancer, or even that most people who smoke will get lung cancer (See Hitchcock (1995) p265 and ch1 of Eells (1991)). It can be read as saying merely that smoking is a positive causal factor for lung cancer, and where this is a claim made about a group or population. Such a probabilistic claim is fully compatible with determinism, since it might be the case that for a given level of smoking, some members get lung cancer while others do not. We can imagine, for now, that the probabilities are given an epistemic reading, since *ex hypothesi* the relationship between smoking and lung cancer is deterministically governed. If this is the case, there cannot be any objective chance concerning one person's susceptibility to lung cancer from smoking – they either will or will not get cancer, within a certain timeframe, given a certain level of smoking (holding fixed any other causes of lung cancer which might be acting simultaneously).

Our second distinction is the familiar one we have already seen, that between type and token causation. The distinction is especially important here, and some philosophers have sought to give separate accounts of probabilistic causality for each (Sober (1985), Eells (1991 ch6)). This is so because of the special problems posed by single case chance, and also because of the way that our different notions of causation can interact. Some examples will help to bring this out below, but for now it will suffice to say that it is often clear that, though *in general* (at the level of types) something may be a positive causal factor for something else, in an individual circumstance the presence of that factor may not be a cause. Hitchcock (2004) gives the example of two riflemen who shoot at a vase; both have a probability of .5 of hitting the vase.[1] One would think, then, that the overall probability of the vase smashing is .75. But what actually happens is that one rifleman hits and the other misses, so although the second rifleman's shot raised the probability of the vase smashing, it appears not to be a cause of that event. Conversely, events which *generally* (type) lower the probability of certain effects *can* be causes. These kinds of cases are more frequently discussed in the literature; the paradigm example is of a golfer whose ball is kicked seemingly off course by a squirrel but through an unlikely sequence of ricochets, ends up in the hole. Squirrel kicks are the kinds of event which generally speaking lower the probability of a ball landing in the hole, and so should be seen as preventers rather than causes. But here, so the argument goes, the kick should be counted as a cause. Anyone wanting to give an integrated account of token and type causation will need to address these two kinds of cases.

[1] It is stipulated that these probabilities are genuine chances.

This chapter is organised as follows. Firstly I shall discuss the distinction between genuine and pseudo-indeterminism, with some commentary on the various interpretations we can give to each theory. I shall then proceed to consider how interventionism might seek to capture the various kinds of cases under discussion. It will become apparent that interventionism, because it is designed as a theory mainly of type causation, may not have quite the resources to account for token indeterminism.

5.2 Genuine indeterminism

The concept of causation did survive the discovery that the world may be fundamentally indeterministic, but in a weakened form. If we consider not just a single atom and its decay, but rather a situation where we can influence the decay in some way, then it looks as if we can cause the decay even if it is not determined *exactly* how things will pan out. Thus the probability-raising theory of causation was born. Very roughly, this says that for C to be a cause of *E*, the presence of C must raise the probability of *E* happening. This can be couched in terms of conditional probability:

PR: C causes E if and only if $P(E \mid C) > P(E \mid {\sim}C)$

However, as Hitchcock (2002) points out, there are two major problems which *PR* faces. The first is that, whilst causation is an asymmetric relation, conditional probability is not. Causes are generally taken to precede their effects, but what probability some proposition or event has, conditional on another is not sensitive to when either of the two happen. Secondly, there can be what are known as spurious correlations, whereby two types of event, though they are not directly causally related,[2] are nevertheless correlated due to them both being the effects of a common cause. For example, low barometer readings and storms are correlated, so that the probability of a storm conditional on some (set of) low barometer readings is higher than it would have been with a different set of (higher) readings. But the low barometer readings do not cause storms. Rather, low barometer readings and storms are both the effects of a common cause, namely, low atmospheric pressure. There could also be cases in which the events in question have their probabilities correlated but where this is just accidental, although this is the subject of some dispute.

As Hitchcock (2002) notes, both Hume and Mill provide ready answers to the asymmetry problem - just stipulate that causes must precede their effects. This might seem like a rather ad-hoc solution, for, as Hitchcock further points out, this is to rule out backwards causation *a priori,* and it also prevents one, on pain of

[2] That is, neither of them cause the other

vicious circularity, of developing a causal theory of temporal order. As we said in chapter one, however, it is not an uncommon move to make, even within contemporary philosophy.

A by now standard response to the problem of spurious correlations is to require that *PR* be augmented with a further condition. According to Eells (1991) and Cartwright (1979), causes must raise the probability of their effects *ceteris paribus*. What this means is that, given a *range* of different backgrounds, the effect must still have a higher probability, given the cause. This can be formalised as:

PR_b: C causes E if and only if $P(E \mid C\&B) > P(E \mid {\sim}C\&B)$

So how does this work? If we take our barometer and storm example, for it to be the case that low barometer readings cause storms, the conditional probability of a storm given the proposed causally influential barometer reading(s) must be higher than that of there being a storm given some other set of readings, both in a background where the pressure is low and high. This is just another way of saying what interventionism says. In epistemic terms, we can say that in order to *test* whether barometer readings really do cause storms, we need to ensure that any potential common causes are tested for. Interventionism does a clean job of spelling out the way in which these 'test scenarios' are to be implemented; we need to hold fixed any off path variables and also hold fixed the direct causes (parents) of the cause itself. Then, by surgical intervention we see if there is any change at the 'storm' variable for some change in the 'barometer' variable.[3]

What is interesting is that the probability raising analysis can serve for both indeterministic and pseudo-indeterministic cases, at least at the level of types. For genuine indeterminism, this is simple. The probabilities measure event types like "neutron bombardment" and "radioactive decay". Does neutron bombardment cause decay? To answer this question, we need to consider the probability of decay (within some time interval) given bombardment, as compared with non-bombardment. It is clear that even when we condition on background factors, the probability of decay given bombardment increases.[4]

[3] I have couched this in epistemic terms, but as already stressed the definition of causation is based on a model which is *already known* to be 'correct', where this correctness is nevertheless perspectival in some sense.

[4] But what makes this true? There is some conceptual difficulty here. On the one hand, it could be that neutron bombardment has the effect it does because some of the neutrons 'whack' the atom directly, causing it to decay (or weaken, perhaps) while other neutrons miss (and that is why it is a probabilistic law). On the other hand, we could say that each firing raises the probability of decay by *x* amount and where this is just the way that nature is – it deals inherently with probabilities. On the first view, the distinction between indeterminism and pseudo-indeterminism just collapses – everything would be pseudo-indeterministic on this view. I take it that, at least on the dominant interpretation of quantum mechanics, this is not the way to read genuine indeterminism – some things are genuinely, inherently indeterministic. A minority think that we can hold onto classical deterministic physics, for example by positing so called 'hidden variables' which would ultimately account for what *appear* to be irreducibly indeterministic systems.

For genuine indeterminism, then, there seems to be a much tighter interplay between token and type causation. Any token case in which some neutron is fired at an unstable atom, will raise the probability of the decay of the atom. Hitchcock (2003 p18) considers three possibilities for interpreting token indeterminism of this kind. The example is where we have a bombardment and decay, and where the probability of decay given the bombardment is much higher than without:

(1) The bombardment was a cause the decay. This is the position held by David Lewis (1986) and Paul Humphreys (1989). The indeterministic nature of the relationship between bombardment and decay does not prevent it from being a cause. It is important to note that, on this view, causation can still be seen as a relation between events.

(2) The bombardment causes the raising of the chance, but that is all it does – whether there in fact is a decay is still a matter of sheer chance. Dan Hausman (1998) takes this line. There is great difficulty for this

view in understanding causation as a relation between events. Instead, causation might have to link facts.[5,6]

(3) The matter is underdetermined. In any given case, it might be that the bombardment caused the decay, or it might be that it merely raised the probability of decay. Interestingly, this is Woodward's (1990) view. Does this have any implications for what the relata of causation are? There is some ambiguity here stemming from what 'underdetermination' means. Let us explore this in some more detail.

Woodward's (1994) – quoted in Hitchcock (2004)) discussion is actually in relation to a subtly different example. The case involves two carcinogens, both of which act indeterministically:

> "Suppose we know...that each of the two different carcinogenic materials C1 and C2 [...] can cause [stomach tumours] in mice (E). Suppose also that there is no evidence for any interaction effect between C1 and C2 when both are present. Now suppose that a particular mouse is exposed to C1 and C2 and develops cancer (E). It follows on Humphreys' account that since both C1 and C2 increase

[5] For example, the fact that there was a neutron bombardment caused the chance of decay to be raised to such and such a level.

[6] It is not entirely clear what the difference is between (1) and (2). See footnote 8.

> the probability of cancer, both cause or have causally contributed to cancer. But why should we believe this? How do we know that the cancer was not instead caused by C1 alone or C2 alone? We know that when C1 occurs in isolation, it is perfectly possible for it to increase the probability of E and yet fail to cause E. Here probability increase is not sufficient for actual causation. How do we know that the envisioned case is not also one of these cases, in which C1 fails to cause E, and E is instead caused by C2 alone?"

This echoes the arguments found in both Foster (1985) and Armstrong (2004) in their discussions of singularism. They both use thought experiments to show that there must be a 'further fact' to causation beyond the global regularities concerning what the laws of nature are. The common feature of all these cases is that they involve indeterministic causation, and where two instances of the law overlap in some way. The problem essentially involves that of moving from a case where only one law is instantiated, and where even though there is some indeterminism we can say *for sure* that the effect was caused by some event *x*, to a situation where, because of the overlap, we can no longer say this. There are, instead, two candidates in play. As an example, Foster imagines a world in which it is a law that a sphere of some metal, when heated to some temperature *t*, produces a flash within two radii of the centre of the sphere. If two spheres were situated close to one another, so that they had intersecting 'flash fields' and were both heated to the requisite temperature, it could happen that both flashes occur in the intersecting region. Two possibilities would then be open as to what happened. Either Sphere S1 caused Flash F1 and Sphere S2 caused Flash F2, or vice versa. But given that there may be no further law governing where the flash occurs *within* the defined region, there is nothing we can appeal to in order to determine which of these possibilities was actual. So if we want to hold onto the idea that there is a definite fact of the matter as to which sphere caused which flash, there must be fully local, irreducible causal facts which will ground this for us. If there is no fully local further fact concerning what causes what, then we are in a bind, and must say that both are causes. Armstrong says:

> "The obvious thing to say [...] is that *e* may or may not causally depend on *c* [...] Suppose that besides *c*, a potential cause *c1* is also present and that it is the latter which gives the smaller chance of *e* occurring. It seems to be a perfectly objective question, when *e* occurs, whether it is *c* or *c1* that is the cause, although it is more likely to have been *c*."

This seems to be, basically, Woodward's position, putting him firmly in the singularist camp.

Hitchcock (2004) considers the Lewisian response by invoking the idea of a 'probability pool'. Lewis and Humphreys consider any probability raiser to be a cause because they fundamentally reject Woodward's picture of how probabilistic causality works. Humphreys says:

> "...the situation with both carcinogens *is* different from the situation with only one – the chance is higher than with either alone because both chemicals have contributed to the value of that chance. And that is all there is. To think otherwise is to conceive of the example in terms of a *deterministic image* where the tumour was 'entirely caused' by the first chemical and the second chemical was thereby irrelevant. [...] The second chemical is not irrelevant on this (or any other) occasion for it contributes to the chance on this occasion, as does the first chemical, and after they have done this, *nothing else causal happens*. It is...a matter of sheer chance whether the tumour occurs or not."[7]

As Hitchcock further explains, these two theories are fundamentally at odds concerning what probabilistic causality is. Lewis and Humphreys articulate a position for genuine indeterminism which says that probabilistic causes "bring about their effects only via their contributions to a probability pool" (ibid. p408).

[7] Actually, this last sentence makes Humphreys sound like he's closer to the Hausman view. What is involved in indeterministic causation? Only the raising of chances, - this is what Hausman holds. But I think the key phrase is that 'nothing else causal happens' – in other words, there is nothing else to probabilistic causation than this. Perhaps the difference is this. While Lewis and Humphreys think that, although there is nothing more to probabilistic causality than the raising of the chance of the effect conditional on the cause, this somehow does not prevent the causal relation holding between events. Hausman thinks that this is implausible; probabilistic causality involves the deterministic raising of chances, and so whatever kind of thing chances are, they must play the role of being (at least one of) the causal relata.

Before considering the implications of this debate for interventionism, it will be instructive to consider some issues thrown up by pseudo-indeterministic causality.

5.3 Pseudo-Indeterminism

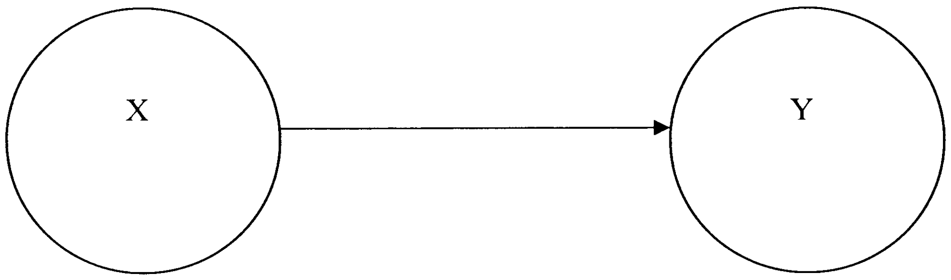
The three options we have laid out concerning the interpretation of genuine indeterminism do not seem readily applicable to pseudo-indeterministic cases. The only reason we had (say, 20 years ago, before anything was known about any kind of mechanism linking tar deposits with cancer) for saying that smoking increases the probability of developing lung-cancer was that, in a population, the two were correlated. Assuming that a common cause could be ruled out,[8] we could say that smoking was a positive causal factor for lung-cancer. We might even be able to quantify the probabilities. Lets say that for a given population, there is a baseline lung-cancer rate of 2%. This means that 2% of the population will develop lung cancer at some point in their lives, no matter what else happens. Further assume that, for every pack of cigarettes smoked per day, the rate goes up by 10%. So, if everyone in the population smoked two packs a day, the rate of lung-cancer would be 22%. Intuitively speaking, there is a sense in

[8] This was no mean feat. In fact tobacco companies argued for a long time that there was a common cause lying behind the correlation between smoking and lung cancer, and that therefore one would not prevent lung cancer by cutting down on smoking.

which these probabilities can be interpreted so that they are genuine properties of the population. Of this population, it can be said that it has the property of its members being susceptible to lung-cancer from smoking to such and such a degree (on average). Indeed this kind of information is vital for guiding and implementing government policy, say. We might be interested in the effect, broadly speaking, of banning cigarette advertising. The effects we are interested in here are not those at the level of the individual, but only at the population level. So even though we have said that this kind of probabilistic causality is pseudo-indeterministic, there is a good sense in which we can abstract away from the individuals who make up the population, and treat the population *as if* it had these probabilities genuinely. The importance of this will be made apparent directly.

5.4 Interventionism, types and tokens

What implications do genuine and pseudo-indeterministic causality have for interventionism? Need it take a stance on any of the metaphysical positions we have spelled out, or can it glide over the metaphysical messiness as with, say, omissions? Let us begin with the type and token level analysis of pseudo-indeterminism. For smoking and cancer, we can draw the following, very simple, DAG:



$$Y = aX + b$$

This DAG and its associated structural equation say that Y, the probability of a member of the population getting lung cancer within a lifetime, is increased for a given increase in the level of smoking (*X*). Because there is baseline probability of lung cancer, we need an additional factor 'b' to represent this. In the example I gave above, the equation would be $Y = 0.1X + 0.02$.

The problem seems to be that, though this DAG represents very well what happens at the population level, we seem unable to use it at the level of tokens, unless the probability of getting lung cancer is given the special interpretation of being *epistemic*. This is because, as we said, since there is a definite fact of the matter concerning each person's susceptibility to lung-cancer from smoking, the average susceptibility captured by the population level DAG may not actually represent any individual's probability of getting cancer for a given level of smoking.[9] But given that we have continually stressed the priority of type level graphs over token levels, because it is at the level of types that the causal information is encoded, will this schema not now fall apart?

The answer is obvious. It depends on what you mean by 'type'. We can, if we so wish, represent what goes on at the population level, where we aggregate the different effects that smoking has on each individual. This will leave us with a probability measure for the population's cancer rate, given a certain level of smoking per capita, and will provide useful policy guidelines concerning what to do at the level of populations.[10] What it will not do is tell us for sure that those policies will be effective for everyone in the population. For example, there may be members who are not susceptible to cancer no matter how much they smoke. For these members, the level of advertising that the government allows will make no difference at all. There is only one sense in which the type level DAG will 'translate down' to the token level. Given that it is not *known* where on the spectrum each member of the population lies in respect of their susceptibility to cancer from smoking, the probability of getting cancer for a given level of smoking can be interpreted as epistemic. How sure should each member be that they will get cancer for a given level of smoking? In their state of ignorance, the

[9] Indeed, if the law governing smoking and cancer is deterministic, there is not any sense to be made of talk of probability. One either will or will not get cancer for a given level of smoking, *ceteris paribus*.

[10] The government will want to formulate policies which will cut down smoking across the board, via 'broad brush' actions like banning advertising.

best they can do is consult what happens to the average person. So the DAG can still be used to represent the 'type' of person that exists in this population, in that sense.

If we really want to capture the causal relation that holds or does not hold at the token level, however, we cannot use a simple aggregation like the one above. Instead, we must consider each 'susceptibility group' separately. As a simplification, lets say that there are only two types of person. Type *A* people will definitely get cancer if they smoke more than one pack a day. Type *B* people will only get cancer if they smoke two packs a day, again this will happen deterministically. The two types must be represented on two different DAGs if we are to capture exactly the causal relations which exist. For each type of DAG, we can then represent every individual within the population, and the causal relationship which holds between smoking and cancer.

5.5 Interventionism and genuine indeterminism

Genuine indeterminism perhaps provides a greater challenge to the interventionist. If we take Hitchcock's three options as being exhaustive of the interpretation we can give to genuine indeterministic causality, then interventionism might have to take a stand; if it does then this will threaten the metaphysical neutrality of the theory. This would not be a disastrous result, so long as each of the three interpretations is equally plausible. I shall not here have space to extensively discuss the relative merits of the three interpretations, but in any case showing that interventionism has some metaphysical commitments would be surprising in and of itself.

We have already seen that Woodward, in the example involving two indeterministic carcinogens, embraced the singularist doctrine, which is committed to the view that there is always a matter of fact as to which of the carcinogens caused the cancer. Perhaps because of specific features of the case we can feel some sympathy with this view, even if for other cases our intuitions would be pulled in other directions. It might have been presumed that since cancer production is a bio-chemical process, there must be a spatio-temporal chemical pathway leading from only one carcinogen to the tumour. We can probably do better by considering the bombardment – decay examples. There is some chance of a particle decaying within the next 5 seconds, lets say it is close to zero. If we fire a neutron, the chance of decay within this time frame goes up to 0.5. So, firing 2 neutrons increases the chance of decay to 0.75 (assuming that the two firings do not interfere with one another). If one buys the singularist line here, then there is a determinate matter of fact as to which of the two neutrons caused the decay.[11] So post-hoc, one neutron caused the decay with 100% certainty, whilst the other merely drifted on past the atom. But I don't think that this result sits easily within the interventionist paradigm. If we were to model the situation at the level of types, we don't seem to be able to discriminate among those atoms which are going to be the causes and those which are not. This is because the system is not deterministic, in the way that smoking might be for cancer. Every neutron bombardment must count as a cause because every bombardment raises the chance, genuinely, that the decay will occur. Singularism says that there is a way that the world is on every occasion in which a cause-effect pair is instantiated, even though there is nothing we can point to, even *in principle*, to ground this. It is just that, as things happen, there is a determinate matter of fact about what was the cause. But we cannot represent this at the level of types, since the type level graph will have to count anything and everything which raises the chance of the effect as a cause.

Singularism thus commits one to breaking with the idea that there is a strong link between type and token causation. Another way of making the same point is that interventionism uses the idea of reproducibility, relying on the stability of the causal system in play in order to ground the counterfactuals concerning what

[11] It is important to distinguish between determinate-ness and determinism. Determinism says that if we were to reset the conditions back to the way they were before the neutrons were fired, then it is guaranteed that the same neutron will be the cause. Determinateness is compatible with indeterminism, since it is only committed to the thesis that one or the other, at the end of the day, was the cause.

will happen under interventions on the system. But determinateness of cause (singularism) coupled with indeterministic process disallows this reproducibility. For whilst the indeterminism remains when we intervene to test what is the cause, singularism is committed to only one thing *being* the cause, but this could change in moving from one test to another. So we will end up getting different results on different occasions, destroying the reproducibility which grounds the interventionist counterfactuals.

Everything begins to look rosier on either of the other two conceptions of what the nature of probabilistic causality is. This is because reproducibility is restored. As Humphreys noted in his response to Woodward, his view asserts that on each and every occasion on which a probability-raiser acts it is a cause, because that is all a probabilistic cause is. So this is perhaps one way in which interventionism is not as metaphysically neutral as it would at first seem. At least in the indeterministic realm, it is constrained in its approach, forced to choose a perhaps controversial interpretation of what causality ultimately is.

Bibliography

Anscombe, G. E. M. (1975) "Causality and Determination," in E. Sosa, ed., *Causation and Conditionals*. Oxford: Oxford University Press, pp. 63-81

Armstrong, D.,(2004) "Going Through the Open Door Again" in J. Collins, N. Hall, and L. A. Paul, eds., *Causation and Counterfactuals*. Massachusetts: The M. I. T. Press, pp445-457

Beebee, Helen (2004), "Causing and Nothingness," in J. Collins, N. Hall, and L. A. Paul, eds., *Causation and Counterfactuals*. Massachusetts: The M. I. T. Press, pp. 291-308

Carroll, John (1994) *Laws of Nature*. Cambridge: Cambridge University Press

Cartwright, Nancy. (1979) "Causal Laws and Effective Strategies," *Noûs* **13**: 419-437

Collins, J., Hall, E., and Paul, L. (2004): *Causation and Counterfactuals*. Cambridge, Mass: MIT Press

Eells, Ellery. (1991) *Probabilistic Causality*. Cambridge, U.K.: Cambridge University Press
— (1988) "Probabilistic Causal Interaction and Disjunctive Causal Factors." In: Fetzer (1988), pp 189-209.

Fetzer, J. (ed) (1988) *Probability and Causality: Essays in Honour of Wesley C. Salmon.* Dordrecht: Reidel.

Foster, J. (1985): *Ayer,* London: Routledge & Kegan Paul.

Gopnik, A. *et al.,* (2004), "A theory of causal learning in children: Causal maps and Bayes nets", *Psychol. Rev.* **111** pp. 1–31

Hall, N., (2004): "Two Concepts of Causation", in Collins, Hall, and Paul (2004), pp.225-76.

Hausman, Daniel. (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press

Hitchcock, Christopher (1995), "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies* **78**: 257 – 291
—(2002) "Probabilistic Causation", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), forthcoming URL = http://plato.stanford.edu/archives/fall2008/entries/causation-probabilistic/
—(2003) "Of Humean Bondage", *British Journal For the Philosophy of Science* 54 pp1-25
— (2004): "Do All and Only Causes Raise the Probabilities of Effects?", in Collins, Hall and Paul (2004), pp.403-418

Humberstone I. L., (1997) 'Two Types of Circularity', *Philosophy and Phenomenological Research* 57, 249-280.

Humphreys, Paul. (1989) *The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences,* Princeton: Princeton University Press.

Kvart, Igal (1986) *A Theory of Counterfactuals*. Indianapolis: Hackett Publishing

Lewis, D. (1973a): *Counterfactuals*. Oxford: Blackwell.
— (1973b): "Causation", *Journal of Philosophy,* 70, pp.556-67. Reprinted in his (1986a).
— (1979): "Counterfactual Dependence and Time's Arrow", *Nous,* 13, pp.455-76. Reprinted in his (1986a).
— (1986a): *Philosophical Papers: Volume II*. Oxford: Oxford University Press.
— (1986b): "Events", in his (1986a).
— (1986c): "Postscripts to 'Causation'", in his (1986a).
— (2004a): "Causation as Influence", in Collins, Hall, and Paul (2004),pp,75-106.
— (2004b): "Void and Object", in Collins, Hall, and Paul (2004), pp.277-90

McDermott, Michael (1995) "Redundant Causation," *British Journal for the Philosophy of Science* **46**, pp. 423-44

Mellor, D. H. (1995): *The Facts of Causation*. London: Routledge

Menzies, Peter and Price, Huw (1993) "Causation as a Secondary Quality," *British Journal for the Philosophy of Science* **44**, pp. 187-203

Nolan, D., (2005) *David Lewis,* Acumen Publishing

Paul, L. (2004): "Aspect Causation", in Collins, Hall, and Paul (2004), pp.205-224

Pearl, J. (2000): *Causality*. Cambridge: Cambridge University Press.

Price, H and Corry, R. (2007): *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press

Reichenbach, Hans (1956) *The Direction of Time*. Berkeley: University of California Press

Russell, Bertrand (1912) "On the Notion of Cause," in J. Slater, ed., *The Collected Papers of Bertrand Russell v6: Logical and Philosophical Papers 1909-1913*. London: Routledge Press, pp. 193-210

Sober, Elliott. (1985) "Two Concepts of Cause" in Peter Asquith and Philip Kitcher, eds., *PSA 1984, Vol. II* (East Lansing: Philosophy of Science Association), pp. 405-424

Sosa, E. and Tooley, M. (eds.)(1993): *Causation*. Oxford: Oxford University Press

Spirtes,P., Glymour, C, and Scheines, R. (1993): *Causation, Prediction, and Search.* New York: Springer.

Strevens, M., (2007) "Review of Woodward, Making Things Happen" *Philosophy and Phenomenological Research* 74 (1) , 233–249
—-(forthcoming)"Comments on Woodward, Making Things Happen" available at http://www.strevens.org/research/expln/WoodRiposte.pdf

Tooley, Michael (2004) "Probability and causation," in P. Dowe and P. Noordhof, eds., *Cause and Chance: Causation in an Indeterministic World*. London: Routledge, pp. 77-119

Woodward, James (1990), "Supervenience and Singular Causal Claims", in D. Knowles (*ed.*), *Explanation and its limits,* Cambridge: Cambridge University Press pp. 211-46
—- (2003): *Making Things Happen: A Theory of Causal Explanation,*Oxford: Oxford University Press

Wright, C., (1992) *Truth and Objectivity,* Harvard University Press

Yablo, Stephen (2002) "De Facto Dependence," *Journal of Philosophy* **99**, pp. 130-48