⋆

Mechanistic Theories of Causality

⋆

Jon Williamson

Draft of February 16, 2011

Abstract

After introducing a range of mechanistic theories of causality and some of the problems they face, I argue that while there is a decisive case against a purely mechanistic analysis, a viable theory of causality must incorporate mechanisms as an ingredient. I describe one way of providing an analysis of causality which reaps the rewards of the mechanistic approach without succumbing to its pitfalls.

Contents

§1 Mechanistic and Causal Talk

Mechanistic talk and causal talk appear at face value to be quite different. Mechanistic explanation proceeds in a *downwards* direction: given a phenomenon or capacity to be explained, its mechanism is the constitution of reality that is responsible for it. Indeed this sense of the word 'mechanism' has been quite standard throughout its usage in the English language: thus in 1662 Stillingfleet talks of the mechanism of nature; in 1665 Hooke discusses that of the foot of a fly; in 1715 Desaguliers that of a chimney; in 1770 Percy that of Icelandic

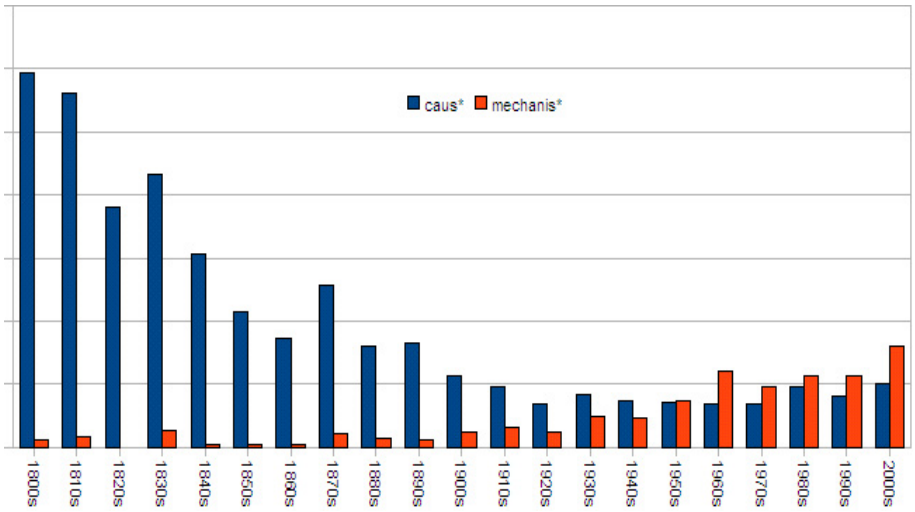

Figure 1: Proportion of English-language books with 'caus-', 'mechanis-' in the title (1800-present). Data source: British Library.

metre; in 1862 Darwin that of the flower.[1] Causal explanation, on the other hand, proceeds in a *backwards* direction: given an event to be explained, its causes are the events that helped bring it about. Thus in 1751 Samuel Johnson writes, 'the greatest events may be often traced back to slender causes'.[2]

The usage of these two words has also followed independent fashions. In the long term—over the last two centuries—'cause' has been falling out of fashion while 'mechanism' has been coming into vogue. Thus we have a steady decrease in the proportion of books with a word in the title beginning with 'caus-', accompanied by a steady increase in the proportion of books with a word in the title beginning with 'mechanis-' (Fig. 1). However, in the short term—over the last four decades—the use of mechanistic language has stabilised while causal talk is on the increase, at least in academic research papers (Fig. 2).

While there are these differences between mechanistic talk and causal talk, there are of course close connections between the concepts of mechanism and causality. On the one hand our knowledge of underlying mechanisms guides our causal ascriptions, while on the other, evidence of causal relationships helps us discover mechanisms. That mechanisms are evidence for causal relations and vice versa suggests some metaphysical connection between the two.

A *mechanistic theory of causality* holds that this metaphysical connection is very close: two events are causally connected if and only if they are connected by an underlying physical mechanism of the appropriate sort (see, e.g., Ney, 2009). Broadly speaking there are two main kinds of mechanistic theory: a process theory (§2) and a complex-systems theory (§3).[3] After giving an introduction

[1]Source: Oxford English Dictionary.

[2]The Rambler, no 141. English usage here has been more varied. Since at least the 13th century 'cause' has also been used more generally for reason or explanation; only relatively recently has it lost a connotation of necessity and become attached primarily to events.

[3]Cartwright's dispositional account of causality might also be classified as a mechanistic theory—this account is discussed separately in Williamson (2006b).

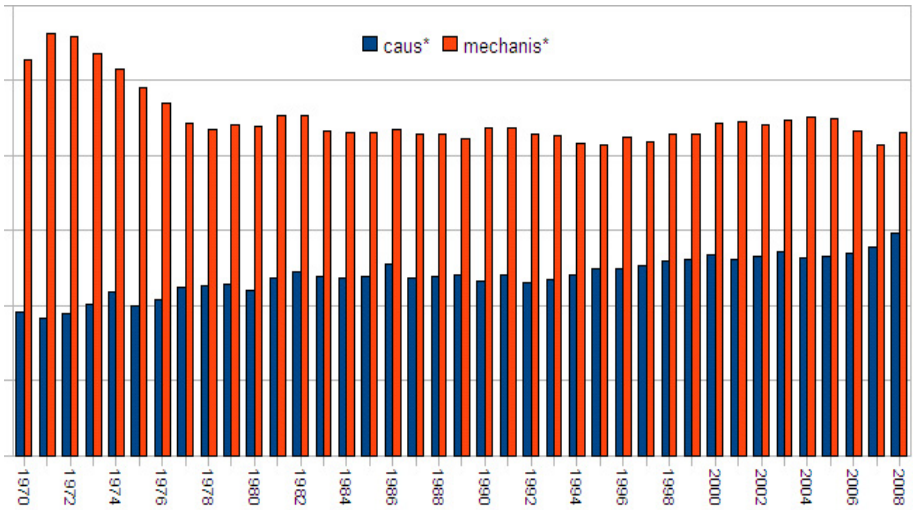

Figure 2: Proportion of English-language research papers with 'caus-', 'mechanis-' in the title (1970-present). Data source: Web of Science.

to these two kinds of account and taking a look at some specific problems that they face, in §4 we shall examine some general problems for a mechanistic theory of causality. I will suggest that no purely mechanistic theory of causality can overcome these problems.

Mechanistic theories of causality are normally contrasted with *difference-making theories of causality*. This heading includes probabilistic, counterfactual, regularity, agency and some dispositional theories of causality. According to a difference-making account, two events are causally connected if and only if a change to one makes a difference (of an appropriate sort) to the other. While the categorisation of theories of causality according to the distinction between mechanistic and difference-making accounts is quite useful to bear in mind, it can be flouted: e.g., one might give a mechanistic account which is essentially difference-making because the mechanisms in question are given a counterfactual analysis, or one might give a difference-making account which is at root mechanistic if the differences in question have mechanisms as truth-makers. We will neither consider these anomalies, nor indeed difference-making accounts in general, in any detail here.

One might expect that if mechanistic theories of causality fail, then a difference-making account will succeed. But there are independent reasons for thinking that no purely difference-making theory will succeed, discussed in §4 and in Williamson (2009). Problems with purely mechanistic and purely difference-making theories are well recognised and have led to a recent upsurge of interest in pluralistic theories to causality (see, e.g., Hall, 2004). *Conceptual pluralism* holds that we have two concepts of cause—typically, one mechanistic and one difference-making. On the other hand, according to *metaphysical pluralism* there are two causal relations—again, usually one mechanistic and one difference-making. While these kinds of pluralism can go hand-in-hand, it is also possible to have one without the other. For example, a conceptual pluralist

might say that although we use different concepts of cause in different contexts, each kind of causal claim is made true by the same relation. Or a metaphysical pluralist might say that while we in fact use a single, unified concept of cause, our usage is ambiguous, sometimes picking out a mechanistic relation and sometimes a difference-making relation. (There is also an *epistemological* or *methodological* pluralism which holds that we discover causal relationships by invoking different kinds of evidence or different kinds of methods; this kind of pluralism is relatively uncontroversial.) I have argued in Williamson (2006a) that conceptual or metaphysical pluralism is a last resort. Conceptual pluralism appears to be factually incorrect. For good or for bad we have a homogeneity of causal usage: while we have lots of specific causal words—e.g., push, raise, shove—we have no problem recognising them to be causal in the same general sense. On the other hand metaphysical pluralism is unattractive because it is not parsimonious. Moreover, we do not seem to need to clarify any apparent ambiguity when we make a claim like 'smoking causes cancer'. Finally, both kinds of pluralism fall to the same problems that beset mechanistic and probabilistic theories (Williamson, 2006a, §2; Russo and Williamson, 2007, §6; Russo and Williamson, 2011).

Granting all this, one might reason that since monistic accounts of causality all fail, we are forced to pluralism whether it be attractive or not. While this inference is valid, it is not sound. While a *purely* mechanistic theory or a *purely* difference-making theory is bound to fail, a monistic account that combines both aspects may yet succeed. In §5 I sketch such an account, the *epistemic theory of causality*, that, I argue, does succeed.

§2
The Process Theory

The idea behind the process theory is that A causes B if and only if there is a physical process of the appropriate sort that links A and B. There are two views as to which kind of physical process is appropriate for underpinning causal relations. One view has it that the process should be one capable of transmitting a mark from A to B (Reichenbach, 1956, §23; Salmon, 1980a, §2). According to the other view, a causal process transmits (Salmon, 1997, §2) or possesses (Dowe, 2000b, §V.1) a conserved physical quantity, such as energy-mass (Fair, 1979), linear momentum, angular momentum or charge, from A to B. We shall now take a closer look at both these views.

Hans Reichenbach's Process Theory

Reichenbach, in his quest to shed light on relativity theory and on the direction of time, developed the idea of a physical process that propagates marks:

> By "signal" we understand a physical process that travels from a real point P to another point P' and has the following property: if this event is marked at P, the mark can also be observed at P'. The word "signal" pinpoints this very property because it means a transmission of signs. The expression "causal chain" is also frequently used in such instances. (Reichenbach, 1924, p. 27)

On the one hand, being embedded in such a physical process is necessary for causation:

> *If E_1 is the cause of E_2, then a small variation* (a mark) *in E_1 is associated with a small variation in E_2, whereas small variations in E_2 are not associated with variations in E_1. ...*
>
> An example: We send a light ray from A to B. If we hold a red glass in the path of the light at A, the light will also be red at B. If we hold the red glass in the path of the light at B, it will *not* be colored at A. (Reichenbach, 1928, pp. 136–138)

On the other hand, such a process is also sufficient for causation:

> If a mark made in an event A_i shows in an event A_k, then A_i is *causally relevant* to A_k. ...
>
> The marking process is sufficient to make possible the construction of the causal net (Reichenbach, 1956, p. 200)

Here the causal net is the complete set of causal relations as determined by the account in question. This set of relations is usually represented by a directed acyclic graph whose nodes are the A_i, A_k, etc. and whose arrows represent direct causal connections.

It is interesting to consider the relationship between this mechanistic theory of causality and the difference-making approach. In fact Reichenbach (1956) developed two theories of causality: the mechanistic theory outlined above and also a probabilistic theory—a probabilistic theory which is at the core of the currently popular *Bayesian net* approach to causality (Williamson, 2005). According to this probabilistic theory, the causal relation can be reduced to patterns of probabilistic dependencies and independencies.

As did Hume before him,[4] Reichenbach viewed his two accounts of causality as co-extensive. He suggested that processes and statistical relations can both provide evidence for causal relationships, and can be used interchangeably:

> it is also possible to combine the marking process with other statistical methods. ...
>
> A mixed procedure of this kind, however, presupposes a certain correspondence between the order established by marking processes and the order derived from the [statistical] relations. This correspondence must now be studied. It is based on certain assumptions ... (Reichenbach, 1956, pp. 200–201)

These assumptions are as follows:

Assumption α. If a mark made in A_i shows in A_k then $P(A_k|A_i) > P(A_k)$.

Assumption β. If a mark is made in A_i, then either $P(A'_k|A'_i) = P(A_k|A_i)$ or $P(A_k|A'_i) = P(A_k|A_i)$, where A'_i and A'_k are the marked versions of A_i and A_k respectively.

[4] 'We may define a cause to be *an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second.* Or in other words *where, if the first object had not been, the second never had existed.*' (Hume, 1748, p. 76).

Assumption γ. If A_2 screens off A_1 from A_3, and if a mark made in A_1 shows in A_3, then it also shows in A_2.[5]

Assumption δ. If $A_2^1 \cdots A_2^n$ screen off A_1 from A_3, and if a mark made in A_1 shows in A_3, then it also shows in at least one of the events $A_2^1 \cdots A_2^n$.

It is not hard to see, however, that these assumptions all admit counterexamples. Consider Assumption α. Suppose a golf-ball has been hit towards the hole, bounces off a squirrel's foot (A_i) but nevertheless proceeds into the hole (A_k). Now a mark made in A_i (say by an inky foot) may show in A_k, since the ink-stain may still be on the ball when it reaches the hole. But this is a classic example of a case in which there is probability-lowering rather than probability-raising: bouncing the ball off a squirrel's foot lowers the probability of getting it into the hole (see, e.g., Salmon, 1980b, pp. 215). Let us turn to Assumption β, which says that adding a mark does not change the relevant conditional probabilities. Consider another golfing example in which A_i is the striking of the ball and A_k is its arrival in the hole. Now marking the ball with ink (A_i') may indeed not change the underlying probability: $P(A_k'|A_i') = P(A_k|A_i)$ if the ink stays on or $P(A_k|A_i') = P(A_k|A_i)$ if it comes off by the time of arrival into the hole. But marking the ball with lead surely could, in which case the assumption fails. For a counterexample to Assumption γ, consider a situation in which A_2 happens if and only if A_1 happens. For instance, consider a perfectly reliable firing squad involving two shooters: each shooter fires (A_1, A_2 respectively) if and only if the command to fire is given, and each shooter is accurate in the sense that the target is hit (A_3) if either shooter fires. Then A_2 screens off A_1 from A_3, but marking A_1, say by scoring that shooter's bullet, will not result in that mark showing in A_2. This counterexample extends straightforwardly to Assumption δ—just consider n shooters in the firing squad.

Since these assumptions all admit counterexamples, they can at best be reinterpreted as *conditions* that need to be satisfied in order for Reichenbach's process theory of causality to agree with his probabilistic theory. But while there may be agreement between the two theories in some cases, there can be no general agreement because there exist counterexamples: the (unrestricted) causal net determined by the process theory will disagree with the (unrestricted) causal net determined by the probabilistic theory. Hence Reichenbach's process theory should be regarded as a rival to his probabilistic theory. Indeed, these days it is widely acknowledged that a mechanistic theory of causality will disagree with a difference-making account over some causal claims (see, e.g., Hall, 2004).

Wesley Salmon's Process Theory

Wesley Salmon developed Reichenbach's process account in Salmon (1980a, 1984, 1997, 1998).[6]

Salmon began by adopting Reichenbach's account of causal processes as transmitters of marks. Thus if one places a rotating white beacon in the middle

[5] Here A_2 *screens off* A_1 from A_3 iff A_2 renders A_1 and A_3 probabilistically independent: $P(\pm A_1 | \pm A_2 \wedge \pm A_3) = P(\pm A_1 | \pm A_2)$, where $+A_i$ represents the occurrence of A_i and $-A_i$ represents its absence. This relation is often written $A_1 \perp\!\!\!\perp A_3 \mid A_2$.

[6] An accessible introduction to Salmon's views can be found in his 'A new look at causality' in Salmon (1998).

of a round building like the Colosseum, the transmission of light from the beacon to the wall is a causal process (since placing a red filter on the beacon will impose a mark on—i.e., modification of—the process, which will endure without further interventions). On the other hand, the spot of light that travels round the wall is not a causal process, but rather a *pseudo-process*, because placing a red filter over the spot will not impose an enduring mark (Salmon, 1998, p. 16). According to Salmon, a mark is introduced by a causal interaction: when two processes meet this intersection is a *causal interaction* if both processes are modified, would not have been modified otherwise, and the modifications persist. It is a *noncausal intersection* otherwise.

These notions provide the key ingredients of Salmon's process account of causality:

> The causal connection Hume sought is simply a causal process. For example, when I arrive at home in the evening, I press a button on my electronic door opener (cause) to open the garage door (effect). First, there is an interaction between my finger and the control device, then an electromagnetic signal transmits a causal influence from the control device to the mechanism that raises the garage door, and finally there is an interaction between the signal and that mechanism. (Salmon, 1998, pp. 17–18).

Salmon later revised his account to more closely resemble Dowe's views, to which we shall turn next.

Phil Dowe's Process Theory

The key idea behind this process theory is that a process is causal if it manifests a conserved quantity (Dowe, 1992, 1993, 1996, 1999, 2000a,b). Conserved quantities include linear momentum, angular momentum, energy, electric charge, colour charge and weak isospin, but there are others. An intersection of two processes is a causal interaction if there is an exchange of a conserved quantity between them.

Wesley Salmon came to adopt this conserved-quantity approach, in order to eradicate his previous account's reliance on counterfactual conditionals. Saying that a process is causal if it transmits a conserved quantity avoids saying that a causal process has a capacity to transmit marks. Since a capacity can be thought of as a disposition to display some behaviour *should* its conditions be triggered, positing a capacity requires the corresponding subjunctive conditionals to be true. But subjunctive—especially counterfactual—conditionals are often treated with suspicion; in particular, it is easy to doubt whether many counterfactual conditionals admit objective truth conditions. Hence it is tempting to replace talk of capacities by talk of conserved quantities, where possible. Similarly, saying that an intersection of two processes is causal if it involves the exchange of a conserved quantity avoids saying that the processes *would* not have changed but for the interaction, i.e., avoids a counterfactual conditional.

Now one might suspect that *transmission* is itself a causal concept. But Salmon adopts an *at-at* theory of transmission: something is transmitted from A to B iff it is present in the process at every point between A and B in the absence of further interactions (Salmon, 1998, p. 21). The resulting theory is able to handle a wider class of causal interactions than previously. Not only can

it handle Salmon's X-interactions (2 processes before, 2 after) but it can account for Y-interactions (1 before, 2 after)—e.g., a nucleus emitting a particle—and also λ-interactions (2 before, 1 after)—e.g., an atom absorbing a photon.

Problems with the Process Theory

We see, then, that the process theory has evolved from the mark-transmission theory of Reichenbach and the early Salmon to the conserved-quantity theory of Dowe and the later Salmon. Naturally this view is not without its detractors, and there are two main kinds of worry.

One concern is that the theory is not low-level enough, failing to account for causation in quantum mechanics. In an Einsten-Podolsky-Rosen (EPR) thought experiment (Einstein et al., 1935; Bell, 1964) an electron and a positron are emitted from a source ensuring that their spins are opposite, so that when Alice measures the spin of the positron in the z-axis and finds it to be positive or upwards, $A = up$, and Bob measures the spin of the electron, he finds it to be negative or downwards, $B = down$. If the emission from the source, s, is the common cause of the values that measurements A and B take, then the common cause does not screen off one measurement from the other, $P(B = down|s) = 0.5$ but $P(B = down|s, A = up) = 1$. This seems rather counter-intuitive and violates Reichenbach's *Principle of the Common Cause*, which says that when two variables are probablistically dependent and neither is a cause of the other, then there is some common cause (or set of common causes) of the two variables which screens off this dependence, i.e., such that the two variables are probabilistically independent conditional on the common cause(s). On the other hand, suppose that this view of s as the common cause of the values of A and B is incorrect. If, instead, the measurement A is a cause of simultaneous measurement B then we get action at a distance: A causes B instantaneously, without time for a signal to be transmitted from the former to the latter. This equally goes against the grain of physics and is a problem for a mark-transmission theory with an at-at theory of transmission. Consequently Dowe (2000a, Chapter 8) goes for a third causal story: the measurement A is a cause of s which in turn is the cause of the value measurement B takes. However, this is perhaps the most counter-intuitive claim of the lot, since the causal connection from A to s is backwards in time. So it appears that none of the three causal accounts of the experiment are particularly palatable: we can choose between a failure of the Principle of the Common cause, action at a distance, or backwards causation. I would suggest that the proponent of a process theory should take the trilemma by the first of these three horns: while current probabilistic analyses of causality are wedded to the Principle of the Common Cause (Williamson, 2009), a mechanistic analysis can take or leave this principle—causal inference is easier with the principle intact but by no means impossible without it (Williamson, 2005)—while action at a distance and backwards causation, on the other hand, stand more in conflict with our mechanistic intuitions. Hence, in my view, there is a way out of the problem posed by the EPR experiment available to the proponent of the process theory: just accept that there are cases in which common causes can fail to screen off the correlation between their effects.

A second concern with the process theory is the opposite concern that conserved quantities are *too* low-level, far removed from most of our causal claims.

Consider a claim like *eliminating the 10% tax band caused an increase in inequality*: not only does such a claim not explicitly mention conserved quantities, it is very hard to see how it could be about conserved quantities at all. Most of our causal claims are high-level claims like this, and it appears that the process theory has some work to do before it can provide a convincing interpretation of such claims.

This second concern is perhaps the more problematic. There are, broadly speaking, three main options for the proponent of a mechanistic account of causality: to bite the bullet, to move to a pluralist theory of causality, or to move to a more general notion of mechanism.

The first option is to maintain that, despite appearances, high-level causal claims in, e.g., economics are in fact reducible to claims about processes. This option is unattractive inasmuch as it is hard to test and establish this claim. It is hard, even, to see how the reduction *could* work. Accordingly, there is a corresponding epistemological problem: it is hard to see how, when we are learning high-level causal relationships, we are actually learning about the progress of conserved quantities; under this view, the epistemology of causality seems totally unrelated to its metaphysics. Furthermore, this first option seems to preclude causal relations that are due to high-level organisation: any causal relation must be underwritten by low-level causal processes, and two scenarios in which the processes differ only by relations other than interaction relations must have the same causal structure; this seems to undermine the view that high-level sciences are causally autonomous from physics.

The second option is to turn pluralist: to advocate a process theory for causality in physics but to accept that causality in economics differs from causality in physics. Several questions immediately arise. What is causality in economics? Do other sciences require yet more notions of cause? Why do we use a single concept of cause to apply to very different ontological relations? This last question is particularly worrying: there seems to be a homogeneity to our causal claims—we make no distinction between uses of the word 'cause' as we move from science to science, and our methods for learning causal relations seem to transcend disciplinary boundaries. Arguably, therefore, pluralism is a last resort, to be appealed to if no coherent monistic account of causality can be found (§1).

While the first two options are by no means refuted, they are unappealing because they seem to raise more questions than they settle. This makes the third option—that of moving to a more general notion of mechanism that is not tied to a particular science—worthy of sustained investigation. We shall turn to this third option now.

§3
The Complex-Systems Theory

The idea behind the complex-systems theory is that A and B are causally related if and only if they both feature in the same complex-systems mechanism. A complex-systems mechanism is a complex arrangement of parts that is responsible for some phenomenon partly in virtue of the organisation of those parts (Machamer et al., 2000; Glennan, 2002; Bechtel and Abrahamsen, 2005).

Machamer, Darden and Craver's Complex-Systems Theory

Machamer, Darden and Craver characterise a mechanism thus:

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions (Machamer et al., 2000, p. 3).

Intuitively the start may be thought of as a cause of the finish, and it is the job of a complex-systems theory of causality to flesh this thought out.

Machamer et al. (2000, §3.1) put forward one view:

> Activities are types of causes. Terms like "cause" and "interact" are abstract terms that need to be specified with a type of activity and are often so specified in typical scientific discourse. Anscombe ... noted that the word "cause" itself is highly general and only becomes meaningful when filled out by other, more specific, causal verbs, e.g., scrape, push, dry, carry, eat, burn, knock over. An entity acts as a cause when it engages in a productive activity. This means that objects *simpliciter*, or even natural kinds, may be said to be causes only in a derivative sense. It is not the penicillin that causes the pneumonia to disappear, but what the penicillin does. (Machamer et al., 2000, p. 6)

Hence A causes B if and only if A engages in an activity to produce B in a mechanism.

According to this approach, the complex-systems theory is to be viewed as a generalisation of the process theory:

> Our emphasis on mechanisms is compatible, in some ways, with Salmon's mechanical philosophy, since mechanisms lie at the heart of the mechanical philosophy. Mechanisms, for Salmon, are composed of processes (things exhibiting consistency of characteristics over time) and interactions (spatiotemporal intersections involving persistent changes in those processes). It is appropriate to compare our talk of activities with Salmon's talk of interactions. Salmon identifies interactions in terms of transmitted marks and statistical relevance relations ... and, more recently, in terms of exchanges of conserved quantities .... Although we acknowledge the possibility that Salmon's analysis may be all there is to certain fundamental types of interactions in physics, his analysis is silent as to the character of the productivity in the activities investigated by many other sciences. Mere talk of transmission of a mark or exchange of a conserved quantity does not exhaust what these scientists know about productive activities and about how activities effect regular changes in mechanisms. As our examples will show, much of what neurobiologists and molecular biologists do should be seen as an effort to understand these diverse kinds of production and the ways that they work. (Machamer et al., 2000, p. 7)

Stuart Glennan's Complex-Systems Theory

Stuart Glennan characterises a mechanism as follows:

> A mechanism underlying a behavior is a complex system which produces that behavior by the interaction of a number of parts according to direct causal laws. (Glennan, 1996, p. 52)

For example, the cardiovascular and respiratory systems have mechanisms for pumping blood, inhaling oxygen and oxygenating blood. Glennan emphasises that 'the choice of decomposition into parts depends on the capacity or behavior to be explained' (Glennan, 1996, p. 52).

Glennan then puts forward a mechanistic account of causality:

> a relation between two events (other than fundamental physical events) is causal when and only when these events are connected in the appropriate way by a mechanism. (Glennan, 1996, p. 56)

This account appears circular, since Glennan's characterisation of mechanism appeals to the notion of a causal law. However, Glennan holds that the causal laws are themselves reducible to lower-level mechanisms (Glennan, 2009, p. 317). Hence the worry might be more one of regress than of circularity. This regress might be halted by an appeal to certain fundamental mechanisms. However, Glennan—in contrast to Machamer, Darden and Craver—does not hold that his mechanistic account encompasses causality in fundamental physics. In fact he suggests that the EPR experiment shows that there is no causality at the fundamental level. (As discussed in §2, more plausibly there is causality but no screening off in the EPR experiment, in which case one might, contra Glennan, appeal to a process theory to handle fundamental mechanisms.) Glennan's own position is to bite the bullet and accept that, since his mechanistic account has no fundamental mechanisms, it must appeal to fundamental causal laws and is circular after all (Glennan, 2009, p. 318).

Glennan (2010) adopts a kind of pluralism, distinguishing two concepts: one event is *causally relevant* to another if the latter counterfactually depends on some property of the former (or on background conditions), while an event *causally produced* another if they are connected by a continuous chain of causal processes. This turns out to be conceptual, rather than metaphysical, pluralism, because mechanisms are the truthmakers for both kinds of causal claim.

Problems with the Complex-Systems Theory

This mechanistic account clearly goes some way towards overcoming one problem that besets the process theory—namely the problem that the process theory is too low-level. On the other hand one might complain that the complex-systems theory is too high-level: if, as in Glennan's account, there is no analysis of causality at the level of fundamental physics then the theory is at best an incomplete analysis of causality, and at worst no analysis at all, on account of the ensuing worries about circularity discussed above. But the approach of Machamer, Darden and Craver suggests one way out: perhaps one can take process-theory mechanisms as the fundamental mechanisms. This course is not altogether plain sailing though. In particular, it is not clear that a Salmon-Dowe process *is* a mechanism in the complex-systems sense, since it need not be a complex arrangement of parts. So the resulting account may become a hybrid of the complex-systems and process theories, rather than a complex-systems theory that admits processes as a special case.

One might also worry that there are mechanisms unaccompanied by causal relationships. In particular, the components of a complex-systems mechanism might not be the sort of things that can stand in the cause-effect relation. Here's an example: the government is seeking extra tax revenue; bachelors have plenty of disposable income, but simply imposing a tax on bachelors would fall foul of sexual discrimination legislation, so the government instead imposes a tax on all unmarried people. A mechanism for taxing bachelors is created—one that proceeds by taxing the unmarried. But taxing the unmarried is not a cause of taxing bachelors, because the putative relata are not disjoint events and the causal relation is normally only taken to relate disjoint events. Hence this is a case of a mechanism without causation. Similarly, mathematical mechanisms are apparently not causal: forcing is a mechanism for deriving the independence of the continuum hypothesis, but it cannot be said to be a *cause* of the existence of models which do not satisfy the continuum hypothesis; **?**, p. 55 talk of 'the mechanism typically used to show presaturation'—another example of a mechanism with no corresponding causal relations. Poetry offers further examples: Thomas Percy talks of 'the construction and mechanism of the Gothic or Icelandic Metre' (**?**, Volume 2, p. 191).

In response to this problem, one can of course simply accept that there are mechanisms without causality but argue that this warrants revising rather than rejecting the complex-systems account. Perhaps one can say that A and B are causally related if they are linked by a mechanism *and are disjoint physical events*. But this may not be good enough. While this move may well handle cases in which A and B do not qualify as causal relata, there are arguably other cases in which it is the relation that is at fault, rather than the relata. In the human body, for example, mechanisms are everywhere: there are any number of mechanisms linking the heart to the kidney, but that is not on its own sufficient for a particular kidney event to be a cause of a particular heart event. In general, that two events are connected by a mechanism does not imply that one event makes a difference to the other, and difference-making is often required to warrant a causal claim. In this respect the complex-systems theory is arguably worse off than the process theory, for in the latter theory it is much more obvious and explicit as to which mechanisms are causal and which are not. One might suggest augmenting the complex-systems theory with an anologue of mark transmission to distinguish the causal from the non-causal mechanisms. But this would reinstate the worry of Salmon's that the resulting theory would essentially be a counterfactual—rather than mechanistic—theory. Indeed, while Bogen (2008, §4) acknowledges the need to distinguish genuine from spurious causally productive mechanistic activities, he despairs of finding some criterion for so doing.

§4
General Problems for Mechanistic Causality

In this section we shall survey some general problems for the mechanistic approach to causality—problems that affect both the process theory and the complex-systems theory.

¶ Causality without Mechanisms. We have already seen cases in which there are mechanisms without causality. Cases of causality without mechanisms are, if anything, even more worrisome for a mechanistic account of causality. The standard example here is the case in which the cause or effect or both are absences. For instance, *my missing my flight in London is a cause of my talk being cancelled in Australia.* The problem is that it is hard to see how the missing of the flight and the lack of a talk can be connected by a physical process or activity, or could be components of a physical mechanism.

The standard response to this concern is to say that cases of causation involving absences should be given a counterfactual, rather than mechanistic, interpretation (Dowe, 2000a, Chapter 6; Dowe, 2001, 2009; Machamer, 2004, §5; Hall, 2004; Glennan, 2010). *If I were to have caught my flight, my talk would have taken place.* As Machamer puts it, absences are '*causally relevant* rather than *causally efficacious*' (Machamer, 2004, p. 36). This leads to a pluralistic view of causality. As discussed earlier, pluralism is a last resort—such a view is by no means attractive. Why should one give radically different analyses to *a lack of oxygen caused the fire to go out*, and *abundant oxygen caused the fire to rage*? The lack of oxygen is but one particular value of a variable here: if the proportion of oxygen in the air is between 0% and x% then the fire will go out, while if it is more than x%, the fire will continue to burn. What is so special about 0% that, in that case and in that case only, the causal claim should have a radically different analysis?

One might try to cash out the relevant counterfactuals in terms of mechanisms (Glennan, 2010). *If I were to have caught my flight my talk would have taken place* because there would be a mechanism linking the catching of the flight and the talk. If this strategy works we have conceptual rather than metaphysical pluralism: at base the analysis is in terms of mechanisms. Alas this strategy is rather dubious because the relevant counterfactuals may themselves involve absences. Suppose coffee was served in the slot vacated by my talk. *My missing my flight in London is a cause of the coffee being served in Australia.* Here the cause, but not the effect, is an absence. The relevant counterfactual would be: *if I were to have caught my flight the coffee would not have been served.* In this case the consequent of the counterfactual conditional is an absence and there is no apparent mechanism linking the catching of the flight and the non-existent coffee. Lewis (2004, §7) has a strategy for dealing with this case: *my missing my flight in London is a cause of the coffee being served in Australia* iff, had I caught my flight, an event would have taken place (my talk, in this case) that is incompatible with the coffee being served, and some mechanism links the catching of the flight and the incompatible event. The trouble with this move is that my talk is not incompatible with coffee being served—while the serving of the coffee during my talk may be less probable than in its absence, the two events are quite compatible. This suggests a way of fixing Lewis' strategy: *my missing my flight in London is a cause of the coffee being served in Australia* iff, had I caught my flight, an event would have taken place (my talk, in this case) that renders the coffee being served less probable, with some mechanism linking the catching of the flight and the talk. But alas this too is prone to counterexamples. Coffee might have been *more probable* during the talk than in its absence, yet the absence of the talk still be the cause of the coffee being served. (Indeed, examples of probability-lowering causes abound in the literature—see, e.g., Salmon (1980b, §2).) So this strategy seems to offer

little hope to the mechanist.

An alternative strategy involves substituting what is actually present for the absence (see, e.g., Thomson, 2003, §4) and invoking mechanisms to explain the causal relation. Accordingly, one might substitute the coffee break for my missing talk but also the actual boarding of the plane by the remaining passengers for my missing my flight. But this strategy is not helpful because the causal relation relates the absences not the presences: it is simply not true to say that *the actual boarding of the flight in London caused the coffee break in Australia*. One might object that what actually is present (the boarding of the other passengers) entails that I missed my flight, so must make a difference to the coffee break in Australia. But, as Armstrong (2004, §5.2) points out, this is not so: what is present only entails the absence when conjoined with the fact that what is present is *all* that is present. Such totality-facts are no more palatable to the mechanist than absences, since it is just as hard to see how a totality-fact can be a component of a physical mechanism as it is to have absences as components. In sum, this strategy is another dead end.

While cases involving absences constitute one way in which there can be causality without a mechanism, there are other ways: double-prevention, for instance (see, e.g., Schaffer, 2000). Arguably, if *his oversleeping prevented him from carrying out the safety check which would have prevented the fire*, then *his oversleeping was a cause of the fire*, despite the fact that there is no mechanism linking his oversleeping and the fire.

¶ Non-Mechanistic Evidence. There is another general problem for mechanistic theories of causality. This is an epistemological problem: mechanistic theories cannot explain the importance of non-mechanistic evidence for causal claims. If a mechanistic theory of causality is correct, then, once the relevant mechanism is known, there should not be any need for further evidence for a causal claim. But typically there is need for further evidence, namely evidence that the putative cause *makes a difference* to the putative effect. The requirement for difference-making as well as mechanistic evidence has long been explicit in the health sciences (Hill, 1965; Russo and Williamson, 2007, 2011), and is becoming increasingly acknowledged in the social sciences too. It is also borne out by psychological studies: when attempting to substantiate a causal claim, subjects ask questions to elicit the underlying mechanism but also ask questions about difference-making (Ahn et al., 1995).[7]

The need for these two sorts of evidence is attributable to the Janus-faced nature of causality. On the one hand, causal relationships are used to *explain*, and mechanisms are paradigm explainers (Machamer et al., 2000). Thus if one wants to explain an event B by invoking a cause—i.e., in an explanation of the form *B occurred because A occurred and A is a cause of B*—it had better be the case that the causal relation is accompanied by a mechanistic relation: that A is a cause of B should signify that there is some mechanism or chain of mechanisms linking A and B that explain the occurrence of B and invoke A. On the other hand, causal relationships are also used for *prediction and control*, and difference-making is required for these tasks. Thus if one wants to predict B on

[7] This is not to say that mechanistic and difference-making evidence are the *only* sorts of evidence for a causal claim—clearly temporal cues are important for assessing the direction of causality, and evidence that A and B correspond to disjoint events is important for determining whether the causal relation can relate A and B at all.

the basis of cause A, or to control B by manipulating A, then it had better be the case that A makes a difference to B: that A is a cause of B should signify that there is some difference-making relationship between A and B that supports the use of the causal relation for prediction and control. Accordingly, a causal relation typically signifies the existence of both a mechanistic and a difference-making relation, and evidence of the existence of both the mechanistic relation and the difference-making relation is typically required to establish the causal claim.

I say 'typically' here because there are important qualifications. We have already seen that absences and double prevention offer cases in which a causal relation need not be accompanied by a mechanistic relation. (This poses a problem for any purely mechanistic theory of causality.) But there are also cases in which a causal relation need not be accompanied by a difference-making relation. If effect B is already bound to occur, independently of cause A, then A can hardly be expected to make a further difference to B. This is known as a problem of *overdetermination* and is a problem for all purely difference-making accounts of causality, including counterfactual accounts (Hall, 2004, §3) and probabilistic accounts (a cause cannot raise the probability of its effect if that probability is already 1). A second problem of overdetermination occurs when there is a partition $\{A_1, \ldots, A_n\}$ of possible causes of B, no one of which is more efficacious than any other (Williamson, 2005, §7.3). If you can get to B only via precisely one of $A_1, \ldots, A_n$, and you are bound to visit some A_i, and you are equally likely to get to B whichever A_i you visit, then while, for any i, visiting A_i is a cause of getting to B, visiting A_i does not make a difference to your prospects of getting to B. Again this is a problem for both counterfactual and probabilistic notions of difference-making. In sum, there are exceptions—due to, e.g., absences, double prevention and overdetermination—that make the requirement of both mechanistic and difference-making evidence less than universal.

That there is such a requirement at all puts paid to a purely mechanistic theory of causality, while the fact that this requirement is not universal puts paid to a conjunctive analysis of causality that deems A to cause B if and only if A and B are related both by a suitable mechanistic relation and by a suitable difference-making relation.

§5
Inferentialism and the Epistemic Theory

Having surveyed the more prominent mechanistic theories of causality, encountered some of the key problems they face, and found these problems to be substantial, the question naturally arises as to whether one can say something more positive about causality. Is there a theory of causality that captures what is right about the mechanistic accounts without succumbing to their pitfalls?

I would argue that there is such a theory. However, the envisaged theory moves away from a realist view of causal relationships as physical mechanisms towards a Humean view of causal relationships as inferential habits. According to Hume, the cause-effect relationship should be understood in terms of 'foretelling one upon the appearance of the other', 'the mind is carried by habit, upon the appearance of one event, to expect its usual attendant', a 'customary tran-

sition of the imagination from one object to its usual attendant' (Hume, 1748, paragraph 59). The Humean inferentialist view was adopted by Mach (1883, §4.4.3) and Ramsey (1929), but seems to have fallen rather out of fashion now.

The *epistemic theory of causality* is one such inferentialist account (Williamson, 2005, Chapter 9). According to this view, an agent's *causal map*—her posited network of causal relationships—allows her to draw certain inferences: in particular, it allows her to construct explanations, to make predictions and to decide how to control her environment. Her causal map is correct to the extent that it allows her to draw successful inferences. If the arguments of §4 are right, a causal map will license successful inferences just when it latches onto mechanistic and difference-making relations in the right way. But it is not simply that each causal relationship in the map should trace both a mechanistic and a difference-making relationship—we saw that this is not plausible due to considerations arising from absences, double prevention and overdetermination. Rather, a causal relationship should be accompanied by a mechanistic relationship just in those cases in which one might expect a mechanistic relationship (*not* in the cases of absences and double prevention) and it should be accompanied by a difference-making relationship only where one might expect such a relationship (*not* in the overdetermination cases). So the map cannot be said to chart mechanistic and/or difference-making relationships; instead it maps out appropriate inferences, and it is a subtle mixture of mechanistic and difference-making relationships that explain the success of the corresponding inferences.

This is left at a schematic level: it remains to say more precisely what the mechanistic and difference-making relations are that explain the success of a causal map, and how evidence of such relations helps to determine a causal map. There are various options open here. In the light of the preceeding discussion, one might appeal to a process theory to understand low-level mechanisms, while high-level mechanisms might be taken care of by a complex-systems account, perhaps with the use of mark transmission to distinguish those mechanisms that are relevant to the causal relation from those that are not (here there is no need to worry about any appeal to counterfactual conditionals in the explication of mark transmission because we are not attempting a purely mechanistic analysis of causality). Turning to the characterisation of difference-making, a standard probabilistic account says that A makes a difference to B iff A and B are probabilistically dependent conditional on B's other direct causes. As to the question of how evidence of mechanistic and probabilistic relations helps determine a causal map, a candidate formal epistemology is given in the appendix, §A.

The causal epistemology of §A is given in some detail, in order to show how a causal epistemology might be specified in a non-circular way. A causal epistemology can be thought of as a function which, when applied to an agent's evidence, yields a causal map that is an appropriate basis for inference for an agent with that evidence. Some causal epistemologies are better than others, and we can imagine an ideal epistemology—one that yields a map that generates the most successful inferences. Now imagine an agent with total evidence—i.e., whose evidence comprises all fundamental matters of fact—who can apply an ideal causal epistemology to that evidence. There is a sense in which a causal map determined by applying this ideal epistemology to total evidence is itself a complete picture of causality. Indeed, one can hypothesise that the facts about causality are determined by the application of an ideal causal epistemology to total evidence. Causality itself just *is* a special causal map. And one can be said

to have causal *knowledge* if one's causal map correctly latches onto features of this ultimate causal map, and not accidentally so, but through the application of a sensible causal epistemology.[8] Thus we see that the epistemic theory (i) understands the nature of causality by appealing to causal epistemology, and (ii) can account for such objectivity of causality as there is.

The schematic view of the epistemic theory is enough to see how this theory might succeed where mechanistic theories fail. Although it retains a link, posited by mechanistic theories of causality, between mechanisms and causality, the epistemic theory does not require that *every* causal relationship be accompanied by a mechanism linking cause and effect. This renders it immune to counterexamples stemming from, e.g., absences and double prevention. That was the first general problem for mechanistic theories (§4). The second problem was an epistemological problem: mechanistic theories cannot account for the need for evidence of difference-making over and above evidence of the required mechanisms, when establishing a causal claim. It is plain to see that the epistemic theory *can* account for this need for the two sorts of evidence: any decent causal epistemology must incorporate the need for both kinds of evidence, so establishing an ultimate causal map (i.e., establishing causality itself) must require both kinds of evidence; this is so because a causal map grounds predictions and inferences about control, as well as explanations.

In sum, if mechanistic theories of causality turn out to be unsustainable, all is not lost. One can retain their attractive aspects by turning to a Humean inferentialist account of causality, along the lines of the epistemic theory.

§6 Conclusion

There is clearly something right about mechanistic theories of causality: if causal relationships are to be explanatory then, by and large, they ought to be accompanied by underlying physical mechanisms, for to explain an event is typically just to point to the constitution of reality that is responsible for it. But there is also clearly something wrong about mechanistic theories: there is no simple isomorphism between causal and mechanistic structure since A can be a cause of B without there being a chain of mechanisms linking A and B (e.g., in cases involving absences and double prevention) and the existence of a mechanism between A and B is insufficient for A to cause B since A may make no difference to B. This last point leads to an epistemological problem: a mechanistic theory cannot explain why evidence of difference-making is required to establish a causal claim in cases where the existence of an appropriate mechanism is already established.

I have suggested that the way out of this quandary is not to abandon mechanisms in favour of a difference-making account—which is itself subject to overdetermination problems and an analogous epistemological problem—nor to adopt a pluralist account, which should be thought of as a last resort. Rather, what is right about mechanistic theories can be retained in an inferentialist account of causality such as the epistemic theory.

[8] If there is more than one ultimate causal map, then arguably causal knowledge must latch onto features common to *all* ultimate causal maps.

Acknowledgements

I am very grateful to David Corfield, Phyllis McKay Illari and Jan Lemeire for helpful comments and discussion, and to the Leverhulme Trust and the British Academy for supporting this research.

§A
A Formal Epistemology for Epistemic Causality

This section provides the details of a formal epistemology that fits with the epistemic theory of causality.[9] The idea is not to develop a psychologically plausible account of causal learning—though the approach developed here fits well with one stream of current psychological research in causal learning (see, e.g., Gopnik and Schulz, 2007)—but to develop a normative framework which permits both mechanistic and difference-making evidence to constrain an agent's causal map, thereby overcoming the objection concerning non-mechanistic evidence discussed in §4. The exact way in which mechanistic and difference-making evidence induces constraints will not be explored in any detail here; rather the focus is on how to integrate those constraints in a general framework compatible with the epistemic theory of causality.

Our starting point is an agent with language $\mathcal{L}$, evidence $\mathcal{E}$, and belief state $(P, \mathcal{C})$, where P is a probability function on $\mathcal{L}$ representing the degrees of belief that are appropriate in the light of $\mathcal{E}$ and $\mathcal{C}$ is a directed acyclic graph on $\mathcal{L}$ representing appropriate causal beliefs, i.e., $\mathcal{C}$ is the agent's causal map. For simplicity we shall take $\mathcal{L}$ to be a propositional language on a set of variables. (A propositional language is normally taken to be defined on a set of propositional variables, each of which can take one of two possible values, *true* or *false*. We will denote the assignments $A = \mathit{true}$ and $A = \mathit{false}$ by a and $\bar{a}$ respectively, and use $\pm a$ to denote an arbitrary assignment to A.) We take $\mathcal{E}$ to include everything the agent takes for granted in her current operating context, including background knowledge, observations, theoretical assumptions and so on. For simplicity we shall suppose that $\mathcal{E}$ can be represented by two components: π, a set of constraints on P determined by the agent's evidence of chances, and κ, a set of constraints on $\mathcal{C}$ determined by the agent's other evidence, including evidence of physical mechanisms. We shall take constraints in π to generated by inequality constraints on probabilities, e.g., $P(b|a) \geqslant 0.3, A \perp\!\!\!\perp B \mid C$ (i.e., variable A is probabilistically independent of B conditional on C), $C \rightleftharpoons D \mid E$ (i.e., C and D are probabilistically dependent, conditional on E). We shall take constraints in κ to be generated by constraints of the form $A \longrightarrow B$ (A is a cause, either positive or negative, of B).[10] Mechanistic knowledge typically imposes negative constraints—the lack of a mechanism between A and B might imply that $A \not\longrightarrow B$—but can also impose more complex constraints such as $(A \longrightarrow B) \Rightarrow (A \longrightarrow C \longrightarrow B)$ (if A causes B then it causes it via C).

The question then arises as to which belief states $(P, \mathcal{C})$ are rational for the agent to adopt, given her evidence $\mathcal{E}$.

[9] See Williamson (2005, Chapter 9) for further discussion of the motivation behind this framework. This appendix supersedes Williamson (2006a, Appendix A), with improvements made in the light of comments from Jan Lemeire, to whom I am very grateful.

[10] A more fine-grained approach could distinguish positive causation, $A \longrightarrow^{+} B$, from prevention, $A \longrightarrow^{-} B$.

Arguably P should satisfy all the constraints in π and κ if those constraints are consistent. π contains constraints that explicitly mention P, so it is straightforward to see how P can satisfy those constraints. Williamson (2005, §5.8) provides an account of how κ constrains P: if A is not a cause of any of the other variables in the language then when P is restricted to those other variables, it should match the probability function that would have been obtained had A not been included in the language (as long as any probabilistic evidence concerning A does not imply otherwise). Arguably P should also be sufficiently equivocal—the agent should only believe a proposition to an extreme degree if forced to by her evidence—i.e., P should have sufficiently high entropy $H(P) = -\sum_{\omega\in\Omega} P(\omega)\log P(\omega)$ (Williamson, 2005, Chapter 5; Williamson, 2010).

We are principally interested in determining the agent's causal map: the graph $\mathcal{C}$ on the variables in $\mathcal{L}$ that represents causal beliefs that the agent should adopt on the basis of evidence κ and π. (Arrows in the graph $\mathcal{C}$ correspond to direct causal connections.) Any causal graph $\mathcal{C}$ will be assumed to be a directed acyclic graph (dag). With respect to such a graph, D_A is the set of direct causes of variable A and NE_A is the set of A's non-effects.

In this case it is clear how constraints in κ constrain $\mathcal{C}$, but the way in which constraints in π constrain $\mathcal{C}$ needs to be clarified. Causal connections are typically (but not always) accompanied by a certain sort of difference-making relation: if one intervenes to change a cause A while controlling for an effect B's other causes, then B also changes. This kind of difference-making, commonly ascertained using controlled experiments, can be explicated in terms of a probabilistic dependence between A and B when holding fixed B's other direct causes (controlling for those causes) and holding fixed A's direct causes (which captures the idea of intervening to change A):

Definition A.1 (Strategic Dependence) *There is a* strategic dependence *from variable A to variable B with respect to probability function P and causal graph $\mathcal{C}$, written $A \Rightarrow B$, iff A and B are probabilistically dependent conditional on B's other direct causes and A's direct causes, $A \rightleftharpoons B \mid D_B \backslash A, D_A$.*

One might expect an agent's belief state $(P, \mathcal{C})$ to satisfy all constraints imposed by her evidence, to causally account for each strategic dependency, and to probabilistically account for each causal connection:

Definition A.2 (Full Fit) *A probability function P and a causal graph $\mathcal{C}$* fully fit *π and κ iff:*

1. *the constraints in π and κ are all satisfied,*
2. *for each pair A, B of variables, $A \Rightarrow B$ if and only if $A \longrightarrow B$.*

But such an expectation may fail to be met, and for good reason: the evidence may be inconsistent, or the agent may have evidence that a strategic dependence is anomalous in that it does not correspond to a causal connection, or the agent may have evidence that a causal connection is anomalous in the sense that it does not correspond to a strategic dependence. So one should not always expect full fit and we need to consider the following residuals:

Definition A.3 (Evidential Residue) *Given $P, \mathcal{C}, \pi, \kappa$, the* evidential residue *of P and $\mathcal{C}$ is the set of constraints in π and κ that are not satisfied by P and $\mathcal{C}$.*

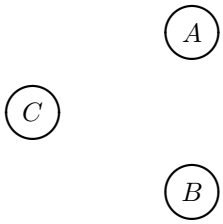

Figure 3: An empty graph.

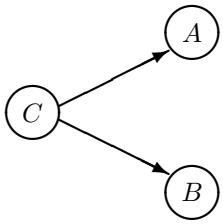

Figure 4: A common cause.

Definition A.4 (Explanatory Residue) *Given $P, \mathcal{C}$, the* explanatory residue *of P and $\mathcal{C}$ is the set of constraints of the form '$A \Rightarrow B$ implies $A \longrightarrow B$' or '$A \longrightarrow B$ implies $A \Rightarrow B$' that are not satisfied by P and $\mathcal{C}$.*

On the other hand, one would not want any evidential residue in cases where the evidential constraints are consistent. Nor would one want any explanatory residue in cases where there is no evidence of anomalous disassociation between strategic dependencies and causal connections. This motivates conditions 1–3 in the following definition:[11]

Definition A.5 (Rational Belief States) *Given evidential constraints π and κ, the set $\mathbb{R}_{\pi,\kappa}$ of* rational belief states *is formed by:*

1. *taking the set $\{(P, \mathcal{C}) : P$ is a probability function on $\mathcal{L}$ and $\mathcal{C}$ is a dag on $\mathcal{L}\}$;*
2. *eliminating those belief states that do not have mimimum evidential residue;*
3. *eliminating those belief states that do not have mimimum explanatory residue;*
4. *eliminating those belief states whose probability function does not have sufficiently high entropy.*

The role of step 3 can be understood in the light of the following example. Suppose the language $\mathcal{L}$ is defined on three binary variables, $\mathcal{L} = \{A, B, C\}$. Suppose that $\pi = \{P(b|a) \geqslant P(b) + 0.3\}$, and that $\kappa = \{A \not\longrightarrow B, B \not\longrightarrow A\}$. Let P be the probability function satisfying π that has maximum entropy: this can

[11] Note that Definition A.5 is a point of departure from the approach of Williamson (2005, §9.5). There it was suggested that $\mathcal{C}$ be determined simply by choosing a minimal graph from all those that satisfy the constraints.

be specified by $P(a) = 0.5, P(b|a) = 0.8, P(b|\bar{a}) = 0.2, P(c|\pm a \pm b) = 0.5$. Then the belief state consisting of P together with the empty graph, Fig. 3, satisfies the constraints and so has no evidential residue, but it does have explanatory residue $\{(A \Rightarrow B) \Rightarrow (A \longrightarrow B), (B \Rightarrow A) \Rightarrow (B \longrightarrow A)\}$: it doesn't explain the probabilistic dependence of A and B. On the other hand, there exists P for which the graph Fig. 4 satisfies the constraints with no residue at all: such a P renders A and C dependent, B and C dependent, and A and B unconditionally dependent but independent conditional on C. Intuitively the latter graph is to be preferred, even though it has more arrows, because it includes an explanation of the dependence between A and B—it attributes the dependence to a common cause. Thus the explanatory residue should be taken into account.[12]

Step 4 is motivated by the consideration that one should adopt equivocal beliefs unless more committal beliefs are forced by one's evidence (Williamson, 2005). Entropy is the standard measure of the degree to which a probability function on $\mathcal{L}$ equivocates between the basic alternatives expressible in $\mathcal{L}$. In particular, if the evidence leaves open three options, (i) A raises the probability of B, (ii) A lowers the probability of B, and (iii) A and B are probabilistically independent, the third option would normally stand out as the maximum entropy option (Williamson, 2005, Chapter 5). The qualification that the entropy should be 'sufficiently' equivocal is needed to deal with the special case in which for each $(P, \mathcal{C}) \in \mathbb{R}_{\pi,\kappa}$ there is always some $(P', \mathcal{C}') \in \mathbb{R}_{\pi,\kappa}$ with higher entropy—this happens in the presence of strict inequality constraints, e.g., $P(a) > 1/2$. In that case one might consider as rational any belief state that is sufficiently high up the entropy ordering—as to what counts as *sufficiently* high is a pragmatic question (Williamson, 2010).

Step 4 rules out overly complex belief states. Suppose again that $\mathcal{L} = \{A, B, C\}$ and $\pi = \{P(b|a) \geqslant P(b) + 0.3\}$, but that now there are no causal constraints, $\kappa = \varnothing$. In this case there are null-residue belief states involving any of the following graphs: $\mathcal{C}_1$ which has only an arrow from A to B; $\mathcal{C}_2$ which has only an arrow from B to A; $\mathcal{C}_3$ which is as in Fig. 4; $\mathcal{C}_4$ which is $\mathcal{C}_3$ but with an extra arrow from A to B; and $\mathcal{C}_5$ which is $\mathcal{C}_3$ but with an extra arrow from B to A. Of these belief states, only those on $\mathcal{C}_1$ and $\mathcal{C}_2$ can have maximum entropy. Hence $\mathcal{C}_1$ and $\mathcal{C}_2$ are picked out as the optimal causal maps.

One can refine this approach in various ways.

First one can extend the approach to take other kinds of belief into account (Williamson, 2005). It may be that a correlation between A and B is induced by a semantic relationship rather than a causal relationship: if A stands for *unmarried* and B for *bachelor* then $P(a|b) = 1 > P(a)$ so $A \rightleftharpoons B$, but this dependence is explained by the meanings of the two variables rather than by

[12] One might worry that taking the explanatory residue into account could lead to the causal model *overfitting* the data. Data is likely to contain spurious correlations and invoking a causal connection to account for each of those correlations will produce a causal map that is a poor basis for inferences about new data. But avoiding explanatory residue requires that causal connections be invoked to explain dependencies of *rational degrees of belief*, not to explain every dependency of the *data*. If an agent is rational to use A as a basis for predicting B then some explanation is required for this fact. However, it is not the case that any correlation between A and B in the data automatically translates into a constraint on rational degrees of belief such as $P(b|a) \geqslant P(b) + 0.3$. Only those correlations that are unlikely to be artefacts of the data (i.e., that are likely to reflect population dependencies) should be translated into constraints on rational degrees of belief.

any causal connection between the two. In which case an agent's *semantic map*, rather than her causal map, might be expected to account for the dependence. If we consider semantic, ontological, logical, mathematical and other kinds of connection as well as causal connection, then the notion of explanatory residue should be generalised. Plausibly, each dependency should be explained by *some* kind of connection between the variables, and if the variables refer to spacio-temporally distinct events then a *causal* connection would be the default explanatory connection.

Second, one can also refine the approach to take richer evidence into account. For instance, if one had evidence of overdetermination then that might be good reason not to expect a causal connection to correspond to a strategic dependence, and to condone some explanatory residue accordingly.

Third, the approach can be extended to take 'latent' variables into account. The above approach operates on a fixed language $\mathcal{L}$, and any common causes are found from within that set of variables (see Fig. 4). But sometimes it is appropriate to postulate a new variable as a common cause that explains a strategic dependence—it may simply not be mechanistically plausible that one of the variables currently under consideration is the common cause. Consequently it is natural to explore ways in which language, as well as the belief state, might change in the light of new evidence (Williamson, 2005, Chapter 12).

Fourth, one might object to Definition A.5 that step 2 is constructed in such a way that the set of rational belief states can depend on the way in which the constraints π, κ are formulated: in certain cases one can rewrite the constraints in a way that is logically equivalent but which ensures that different belief states have minimum evidential residue. In fact step 2 implements a very simple form of consistency maintenance procedure, namely that of considering only states that satisfy maximal consistent subsets of the constraints, and this consistency maintenance procedure is well known to depend on the syntactic form in which the constraints are expressed. But this is not the only possible consistency maintenance procedure and one can easily reconstrue step 2 by appealing to whichever such procedure is most appropriate. The resulting version of Definition A.5 is likely to look more complicated but the protocol for isolating rational belief states is essentially the same.

Causal Markov Methods

This formal epistemology generalises methods based on the Causal Markov Condition, as we shall now see.

Definition A.6 (Causal Markov Condition) *The* Causal Markov Condition *(CMC) is said to hold for $(P, \mathcal{C})$ if each variable A in $\mathcal{L}$ is probabilistically independent of its non-effects, conditional on its direct causes, $A \perp\!\!\!\perp NE_A \mid D_A$.*

The following properties follow from the definition of probabilistic independence (see, e.g., Pearl, 1988, Theorem 1):

Proposition A.7 (Properties of Independence) *For $R, S, T, U \subseteq \mathcal{L}$,*

Symmetry. *$R \perp\!\!\!\perp S|T$ if and only if $S \perp\!\!\!\perp R|T$.*

Decomposition. *$R \perp\!\!\!\perp S, U|T$ implies $R \perp\!\!\!\perp S|T$ and $R \perp\!\!\!\perp U|T$.*

WEAK UNION. *$R \perp\!\!\!\perp S, U | T$ implies $R \perp\!\!\!\perp S | T, U$.*

CONTRACTION. *$R \perp\!\!\!\perp S | T$ and $R \perp\!\!\!\perp U | S, T$ imply $R \perp\!\!\!\perp S, U | T$.*

INTERSECTION. *If P is strictly positive then $R \perp\!\!\!\perp S | U, T$ and $R \perp\!\!\!\perp U | S, T$ imply $R \perp\!\!\!\perp S, U | T$.* □

LEMMA A.8 *A belief state $(P, \mathcal{C})$ satisfies CMC if and only if it satisfies all constraints of the form $(A \Rightarrow B) \Rightarrow (A \longrightarrow B)$.*

PROOF: [⇐] Suppose $\mathcal{C}$ satisfies all such constraints. Suppose for contradiction that $\mathcal{C}$ does not satisfy CMC. Then there is some variable A and non-effect B such that $A \rightleftharpoons B \mid D_A$. This implies $A \rightleftharpoons B, D_B \mid D_A$ by the contrapositive of the Decomposition property of conditional independence, which in turn implies $A \rightleftharpoons B \mid D_A, D_B$ by the contrapositive of the Contraction property.

Since $D_B = D_B \backslash A$, there is a strategic dependence from A to B, $A \Rightarrow B$. But this contradicts the assumption that $\mathcal{C}$ satisfies the given constraints, since $A \not\longrightarrow B$ in $\mathcal{C}$. Thus $\mathcal{C}$ does satisfy CMC after all.

[⇒] Suppose $\mathcal{C}$ satisfies CMC. Suppose for contradiction that $A \Rightarrow B$ but that $A \not\longrightarrow B$ in $\mathcal{C}$. There are four cases:

(i) If B is an (indirect) effect of A then CMC $\Rightarrow B \perp\!\!\!\perp A, D_A \mid D_B \Rightarrow B \perp\!\!\!\perp A \mid D_A, D_B$ (by the Weak Union property) which contradicts $A \Rightarrow B$.

(ii) If A is an indirect effect of B then CMC $\Rightarrow B \perp\!\!\!\perp A, D_B \mid D_A \Rightarrow B \perp\!\!\!\perp A \mid D_A, D_B$ contradicting $A \Rightarrow B$.

(iii) If A is a direct effect of B then $A \Rightarrow B$ implies $A \rightleftharpoons B \mid D_B, D_A$ which is impossible since $B \in D_A$.

(iv) If neither is a cause of the other then CMC $\Rightarrow B \perp\!\!\!\perp A, D_B \mid D_A \Rightarrow B \perp\!\!\!\perp A \mid D_A, D_B$ contradicting $A \Rightarrow B$.

Thus in each case we have the required contradiction. □

THEOREM A.9 *A belief state $(P, \mathcal{C})$ has no explanatory residue if and only if it satisfies CMC and no subgraph of $\mathcal{C}$ satisfies CMC with respect to P.*

PROOF: [⇒] Suppose $(P, \mathcal{C})$ has no explanatory residue. Then it satisfies CMC by Lemma A.8. Suppose for contradiction that some subgraph $\mathcal{C}'$ of $\mathcal{C}$ satisfies CMC with respect to P. Let $A \longrightarrow B$ be an arrow that is in $\mathcal{C}$ but not in $\mathcal{C}'$. Let D_B, D'_B be the set of direct causes of B in $\mathcal{C}, \mathcal{C}'$ respectively. Since $(P, \mathcal{C})$ has no explanatory residue, $A \Rightarrow B$, i.e., $A \rightleftharpoons B \mid D_B \backslash A, D_A$. Now neither A nor indeed any of the variables in D_B or D_A are effects of B in $\mathcal{C}'$, so by CMC, $B \perp\!\!\!\perp A, D_B \backslash A, D_A \mid D'_B$. By the Decomposition and Symmetry properties of conditional independence, $A \perp\!\!\!\perp B \mid D_B \backslash A, D_A$. But this gives the required contradiction.

[⇐] If $(P, \mathcal{C})$ satisfies CMC then it satisfies all constraints of the form $(A \Rightarrow B) \Rightarrow (A \longrightarrow B)$ by Lemma A.8. Consider an arrow $A \longrightarrow B$ in $\mathcal{C}$. Now $A \Rightarrow B$ for otherwise deleting $A \longrightarrow B$ would yield a graph that would also satisfy CMC. Hence $(P, \mathcal{C})$ also satisfies all constraints of the form $(A \longrightarrow B) \Rightarrow (A \Rightarrow B)$ and there is no explanatory residue. □

Consequently, in cases where π, κ admit rational belief states with null residues, the plethora of algorithms that exist for finding minimal graphs that satisfy CMC might be applied to the task of identifying such belief states.

References

Ahn, W., Kalish, C. W., Medin, D. L., and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3):299–352.

Armstrong, D. (2004). *Truth and Truthmakers*. Cambridge University Press, Cambridge.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36:421–441.

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, 1:195–200.

Bogen, J. (2008). Causally productive activities. *Studies in History and Philosophy of Science*, 39:112–123.

Dowe, P. (1992). Wesley Salmon's process theory of causality and the conserved quantity theory. *Philosophy of Science*, 59(2):195–216.

Dowe, P. (1993). On the reduction of process causality to statistical relations. *British Journal for the Philosophy of Science*, 44:325–327.

Dowe, P. (1996). Backwards causation and the direction of causal processes. *Mind*, 105:227–248.

Dowe, P. (1999). The conserved quantity theory of causation and chance raising. *Philosophy of Science (Proceedings)*, 66:S486–S501.

Dowe, P. (2000a). Causality and explanation: review of Salmon. *British Journal for the Philosophy of Science*, 51:165–174.

Dowe, P. (2000b). *Physical causation*. Cambridge University Press, Cambridge.

Dowe, P. (2001). A counterfactual theory of prevention and 'causation' by omission. *Australasian Journal of Philosophy*, 79(2):216–226.

Dowe, P. (2009). Absences, possible causation, and the problem of non-locality. *The Monist*, forthcoming.

Einstein, A., Podolski, and Rosen (1935). Can quantum-mechanical description of reality be considered complete? *Physical Review*, 47:777–780.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14:219–250.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69:S342–S353.

Glennan, S. (2009). Mechanisms. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*, pages 315–325. Oxford University Press, Oxford.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44:49–71.

Glennan, S. S. (2010). Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research*, 81(2):362–381.

Gopnik, A. and Schulz, L., editors (2007). *Causal learning: psychology, philosophy, and computation*. Oxford University Press, New York.

Hall, N. (2004). Two concepts of causation. In Collins, J., Hall, N., and Paul, L., editors, *Causation and counterfactuals*, pages 225–276. MIT Press, Cambridge MA and London.

Hill, B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Hume, D. (1748). Enquiry into the human understanding. In *Enquiries concerning human understanding and concerning the principles of morals*. Clarendon Press, Oxford, 1777 edition.

Lewis, D. K. (2004). Void and object. In Collins, J., Hall, N., and Paul, L. A., editors, *Causation and Counterfactuals*. MIT Press, Cambidge, Mass.

Mach, E. (1883). *The science of mechanics*. Open Court, fourth (1919) edition.

Machamer, P. (2004). Activities and causation: the metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, 18(1):27–39.

Machamer, P., Darden, L., and Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67:1–25.

Ney, A. (2009). Physical causation and difference-making. *British Journal for the Philosophy of Science*, 60:737–764.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo CA.

Ramsey, F. P. (1929). General propositions and causality. In Mellor, D. H., editor, *F. P. Ramsey: philosophical papers*, pages 145–163. Cambridge University Press (1990), Cambridge.

Reichenbach, H. (1924). *Axiomatization of the theory of relativity*. University of California Press, Berkeley and Los Angeles, 1969 edition. Trans. Maria Reichenbach.

Reichenbach, H. (1928). *The philosophy of space and time*. Dover, New York, 1958 edition. Trans. Maria Reichenbach and John Freund.

Reichenbach, H. (1956). *The direction of time*. University of California Press, Berkeley and Los Angeles, 1971 edition.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Russo, F. and Williamson, J. (2011). Generic versus single-case causality: the case of autopsy. *European Journal for Philosophy of Science*, forthcoming.

Salmon, W. C. (1980a). Causality: production and propagation. In Sosa, E. and Tooley, M., editors, *Causation*, pages 154–171. Oxford University Press, Oxford.

Salmon, W. C. (1980b). Probabilistic causality. In *Causality and explanation*, pages 208–232. Oxford University Press (1988), Oxford.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton NJ.

Salmon, W. C. (1997). Causality and explanation: a reply to two critiques. *Philosophy of Science*, 64(3):461–477.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press, Oxford.

Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, 67(2):285–300.

Thomson, J. J. (2003). Causation: omissions. *Philosophy and Phenomenological Research*, 66(1):81–103.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford.

Williamson, J. (2006a). Causal pluralism versus epistemic causality. *Philosophica*, 77:69–96.

Williamson, J. (2006b). Dispositional versus epistemic causality. *Minds and Machines*, 16:259–276.

Williamson, J. (2009). Probabilistic theories. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*, pages 185–212. Oxford University Press, Oxford.

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press, Oxford.