The Time-Asymmetry of Causation

Huw Price and Brad Weslake

1 Introduction

One of the most striking features of causation is that causes typically *precede* their effects - the causal arrow seems strongly aligned with the temporal arrow, as it were. *Why* should this be so? This is the puzzle of the time-asymmetry of causation. In this chapter we offer an opinionated guide to this problem, and to the solutions currently on offer.

1.1 Hume's semantic conventionalism

A good place to start is with the parsimonious patriarch of philosophy of causation in the modern era, David Hume. Early in the *Treatise,* Hume offers the following 'definition' of 'the relation of cause and effect':

> We may define a CAUSE to be 'An object precedent and contiguous to another, and where all the objects resembling the former are

> plac'd in like relations of precedency and contiguity to those objects that resemble the latter.' (*Treatise,* Bk. 1, Part III, sec. XIV)

This proposal makes it a matter (literally) of definition that causes precede their effects. Hume takes the core of the causal relation to be the symmetric notions of contiguity and regularity, and proposes that we impose an asymmetry upon these symmetric relations, by labelling as 'cause' and 'effect' the earlier and later of a pair of appropriately related events. If Hume is right, then the relation between the causal arrow and the temporal arrow is merely a matter of *semantic convention*.

Hume's proposal has some evident attractions. It implies that there is no separate problem about the causal asymmetry, which is just an oblique way of referring to the temporal asymmetry. But despite its economical advantages, Hume's view has not been popular. There are two main objections. The first is that Hume's view makes the connection between causal asymmetry and temporal asymmetry *too tight*. Many philosophers have felt that there is an interesting issue as to whether there are, or could be, instances of *simultaneous causation,* in which the cause happens at the same moment as the effect; or even *backward* (or *retro-*) *causation,* in which the cause happens later than the effect. Hume's view turns these issues into conceptual confusions.[1] If we share the intuition that backward and simultaneous causation are not obviously absurd, we must reject Hume's view, at least in its simple form.

The second difficulty with Hume's view is that it is *too weak,* in the following sense. Causation seems connected to *deliberation*. In particular, the temporal asymmetry of causation seems to have something to do with the fact that it doesn't make sense to deliberate with *past* ends in view. Hume's proposal does not begin to explain this fact. To see this, imagine that we have a ticket in a lottery drawn yesterday. The results have not yet been announced, and we are hoping that we have won. Why does it seem so absurd to try to do something *now* to ensure, or make it more likely, that our ticket was drawn from the barrel some hours ago? If Hume is right, it is no answer to be told that because the draw took place in the past, its outcome cannot be an *effect* of a present action. For on Hume's view, this just amounts to repeating the claim we were trying to *explain,* viz., that we act for later ends (i.e., for Hume, 'effects'), but not earlier ends ('causes'). If there were a present action which would *guarantee* our success in yesterday's draw, why should we care whether it could properly be said to cause it?[2]

The limitations of Hume's view thus bring into focus two general *desiderata* for an adequate account of the time-asymmetry of causation. It should explain the fact that the causal arrow is typically - though perhaps not *necessarily* - aligned with temporal arrow. And it should help us to make sense of a matter of great practical importance in our lives, the fact that we can act for future ends but not past ends (at least in normal circumstances).

We will be stressing the latter point, in particular, at various stages in this chapter - we will call it the *Practical Relevance Constraint* (PRC). It turns

on the intuition that an account of the time-asymmetry of causation should be able to explain the time-asymmetry of deliberation, or at least emerge as part of the same package. We shall argue that its ramifications are wider than usually appreciated; it creates difficulties for some popular attempts to explain the asymmetry of causation.

1.2 The physicalist constraint

Another constraint stems from physicalism - e.g., from the intuition that the abilities the world grants us, and restrictions it imposes on us, are determined ultimately by physics. Hence, apparently, we should look to physics for the origins and nature of the causal asymmetry. Yet this raises a new puzzle. Fundamental physics seems to be time-symmetric, in the sense that if it permits a process to occur in one temporal direction, it also allows it to occur in the opposite temporal direction. How could time-symmetric physics yield something as time-asymmetric as the cause–effect distinction?

One tempting response is to appeal to those parts of physics that are *not* time-symmetric, such as thermodynamics. We shall return to this approach below. First, it should be noted that some writers conclude at this point that there is no time-asymmetric causal arrow. A common view among physicists is that the only physically respectable notion of causation is time-symmetric: viz., the notion of what may be deduced from what in accordance with deterministic laws. For example, Stephen Hawking (1994: 346) describes his encounter with Reichenbach's (1956) work on the direction of time:

> It laid great stress on causation, in distinguishing the forward direction of time from the backward direction. But in physics we believe that there are laws that determine the evolution of the universe uniquely. So if state A evolved into state B, one could say that A caused B. But one could equally well look at it in the other direction of time, and say that B caused A. So causality does not define a direction of time.

Clearly, this symmetric attitude does not explain the asymmetry of practical reasoning. Nor, apparently, is it consistently applied in science. Physicists use ordinary asymmetric causal reasoning as much as anyone else does, e.g., in thinking about the consequences of possible experimental interventions.

A simple example: imagine a photon passing through two polarizers, on its journey from a distant light source. Consider the photon in the region between the two polarizers. Physicists, as much as anyone else, find it natural that the state of the photon at that point depends on the orientation of the first polarizer - the one through which it passed *in the past*. They find it highly counterintuitive that it might similarly depend on the orientation of the second polarizer - the one through which it passes *in the future*. This asymmetry is reflected in the description of such a case in textbook quantum mechanics, according to which the state of the photon reflects the fact that it *has passed* the earlier polarizer, but not the fact that it *will pass* the local future polarizer.

It is not only physicists who have taken the time-symmetry of fundamental physics to provide a reason for denying that there is any such thing as time-directed causality. This was also a motivation for the twentieth century's most famous philosophical critic of causality, Bertrand Russell (1912–1913).[3] Again, however, Russell's view leaves us with a puzzle. What are we to make of the fact that we seem unable to influence the past? If Russell were right that 'physics has ceased to look for causes,' would we be free to make money on yesterday's horse race? On the contrary, obviously, our puzzle would be intact and unsolved, as the issue as to why our practical abilites are so strongly aligned with the temporal arrow.

1.3 Hyperrealism

We might be tempted to respond to the tension between the time-symmetry of physics and asymmetry of causal dependence by denying physicalism - by regarding causation as something "over and above" physics. Physics itself may be time-symmetric, but perhaps there is a further, causal, aspect of reality which is asymmetric. Call this the *hyperrealist* view of causation. It takes causation to be as real as the aspects of the world with which physics is immediately concerned, but not reducible to or supervenient on those aspects.[4]

The main difficulty with hyperrealism is that in putting causation beyond physics, it threatens to make it both *epistemologically inaccessible* and *practically irrelevant*. After all, if the causal direction is detached from physics, then presumably the world could have had the same physics, with an

oppositely-directed causal arrow - in which case, apparently, we have no way of knowing whether our ordinary ascriptions of the terms cause and effect are correct, or back to front. Perhaps the past actually depends on the future. How could we tell? And either way, what practical difference does it make to the choices we face as agents?[5] Hyperrealism thus seems an unpromising solution to the puzzle of the time asymmetry of causation.

1.4 Grounding the causal arrow

Let's review the problem. It may seem that any explanation of the time-asymmetry of causation will need to rest on some account of the nature of causal asymmetry itself - that is, of the intrinsic *difference* between cause and effect (with the issue of time orientation set aside). But Hume shows us another possibility. Perhaps there is no causal asymmetry, as such - no asymmetric causal relation in the world - but only a semantic convention to label symmetric relations with an image of the past–future asymmetry.

By way of analogy, imagine someone puzzled by the difference between royalty and the rest of us. What (he wonders) are the distinctive qualities of *royal* individuals, and why are those qualities correlated with constitutional role - why are they found in particular among the families of hereditary rulers? The analogue of Hume's view - uncontroversial, presumably, in this case! - is that there are no such distinctive qualities. 'Royal' is simply a label applied by convention to the families of rulers of this sort, and the only asymmetry is the constitutional one.

At the opposite extreme from Hume lies hyperrealism. This view not only postulates a real causal asymmetry in the world, but takes it to be a primitive feature, not reducible to physics. We have seen that both extremes seem unsatisfactory. Among other failings, neither meets PRC - on both views, the practical asymmetry of deliberation remains mysterious.

At this point, there are two main options. The first agrees with the hyperrealist that there is a real causal asymmetry, but seeks to make it physical rather than 'extra-physical'. In other words, it seeks a physical asymmetry with the right relation to the temporal arrow - usually but perhaps not necessarily aligned past-to-future - and the right kind of relevance to our deliberative lives. Following Price (1996: Ch. 6), let us call such an asymmetry a *third arrow.* It would provide a link between the causal arrow, on one side, and the temporal arrow, on the other.

If we could find a suitable third arrow, the following kind of account would be on offer:

1. The cause–effect distinction turns on the fact causes are 'upstream' and effects 'downstream', with respect to the third arrow.

2. The link between the causal asymmetry and the temporal arrow turns on the fact that the third arrow has a prevailing temporal orientation; usually, though perhaps contingently, it points 'past-to-future'.

3. The relevance of the third arrow to deliberation ensures that this, too, picks up the usual temporal orientation of the third arrow itself.

Where might we find such a useful piece of philosophical weaponry? Not in (time-symmetric) fundamental physics, presumably, but this leaves the possibility mentioned earlier. The third arrow might be linked to some striking respects in which physics is *not* time-symmetric, such as the time-asymmetry of thermodynamic phenomena. We turn to this proposal in a moment.

The second option is to side with Hume rather than the hyperrealist on the issue as to whether there is an objective causal asymmetry in nature. Perhaps Hume was right to deny this, though wrong in his alternative suggestion concerning the meaning of 'cause' and 'effect'. Hume proposed that these terms indicate the time-ordering of pairs of events in the appropriate (symmetric) relationship, but perhaps this misses the crucial point. In some cases, an asymmetry is a product of an asymmetric viewpoint on a symmetric state of affairs. Think of the distinction between the *left* side of the street and the *right* side; between *nearby* places and *remote* places; or between *locals* and *foreigners*. All these distinctions are drawn 'from a perspective' (and reverse their directions, in the obvious ways, if the perspective changes). As we shall explain, the main alternative to the third arrow strategy proposes that the direction of causation is a case of this kind; and that it is our perspective as *deliberators* that underpins the distinction between cause and effect.[6] We shall return to this proposal in due course.

2 The search for the third arrow

The most prominent example of the third arrow strategy is that of David Lewis. Though not originally proposed as an account of the causal asymmetry in terms of the thermodynamic asymmetry, Lewis's view turns out to be best defended along these lines. We shall explain why this is so, and then turn to a recent proposal in which the link is explicit.

2.1 The asymmetry of counterfactual dependence

Famously, Lewis (1973) defends a counterfactual analysis of causation.[7] The central idea behind such an analysis is that it is typically the case for causally related events that had the cause not occurred, the effect would not have occurred. Of course it is also typically the case that had the effect not occurred, it would have been because the cause did not occur. So what the analysis requires in order to distinguish causes from effects is an analysis of a variety of counterfactual dependence according to which effects counterfactually depend on their causes but not *vice versa*. For our purposes such an analysis is also required in order to address the puzzle of the connection between the asymmetry of causation and the time-symmetry of physics. That is, the analysis should make it clear not only how it is that effects depend on their causes but not *vice versa*, but how this asymmetry is grounded in some asymmetric fact about our world consistent with the time symmetry of fundamental physics.

Lewis (1979) provided just such an attempt within the framework of his possible worlds analysis of counterfactuals.[8] According to this analysis, a counterfactual is true just in case, among worlds in which the antecedent is true, the consequent is true in at least one world closer to the actual world than any in which it is false. The analysis therefore requires an account of *closeness*, or *similarity,* between possible worlds. Lewis rejected the option of making the similarity relation one according to which by definition, for any possible world, worlds preserving the past are always more similar overall than worlds not preserving the past. This would have turned the counterfactual account of causal asymmetry into a variant of Hume's conventionalism, which Lewis rejected for the first reason we discussed in §1.[9] Instead, the similarity relation Lewis opted for was designed to make it a contingent matter that at least generally, with respect to the *actual* world, worlds preserving the past are more similar overall than worlds not preserving the past.

The contingent feature of the world supposed to secure this outcome can be understood by considering how the nearest possible world where the antecedent is true is to be identified, according to Lewis. Call the actual world w_0, the nearest world w_1, and the time of the antecedent t. Under the assumption of determinism, according to which two possible worlds are qualitatively identical either always or never, we know that if the past of w_1 is identical to w_0 and yet w_0 different from w_1 at t, some violation of the laws of nature of w_0 must occur in w_1. This difference between the laws of nature in the two worlds Lewis refers to as a miracle. Intuitively, in w_1 the past is identical to w_0 up until just

before t, at which point things go just slightly differently enough to have the antecedent occur at t. What happens later is left to the laws of w_1 to settle.

Consider now w_2, a competitor to w_1 for similarity. We attempt to construct w_2 by following the temporally reversed strategy - in w_2 the future is identical to w_0 except for just after t, at which point things go just slightly differently enough to have the antecedent occur at t. What happens earlier is left to the laws of w_2 to settle. To put it figuratively, in w_1 we run the tape forwards and diverge just in time to secure an alternative future in which the antecedent occurs, while in w_2 we run the tape backwards and diverge just in time to secure an alternative past in which the antecedent occurs.

What Lewis required here was a reason for thinking there an asymmetry between w_1 and w_2 with respect to the actual world. His strategy was essentially to *deny* that there are worlds such as w_2, in which the antecedent world differs from the actual future only by a *small* miracle.[10] Lewis did not take this alleged *asymmetry of miracles* to be primitive; rather, he took it to reflect a contingent empirical asymmetry that he called the *asymmetry of overdetermination*. A *determinant* is defined by Lewis (1979: 474) as "a minimal set of conditions jointly sufficient, given the laws of nature, for the fact in question", and what Lewis claims is that there are in our world many more future determinants than past determinants for events. Since we are assuming determinism, this is in addition to whole states of the world determining earlier times - as Lewis (1986*b*: 57–58) puts it, there are "plenty of very incomplete cross sections that postdetermine incomplete cross sections at earlier times". And so, if we believe

that had some cause had not occurred, the effect would not have, Lewis claims that even under the assumption of determinism we can not conclude that if the effect had not occurred the cause would not have - since there generally exists some other (future) effect (or set of effects) that are sufficient given the laws to determine the cause. Figuratively, when we run the tape backwards and try to diverge just in time to secure the antecedent, we find that we cannot, since the antecedent is determined by many widespread facts about the future.

2.2 Overdetermination and thermodynamics

Lewis himself professed to uncertainty about the relationship between the asymmetry of overdetermination and that of thermodynamics. His paper ends with the remark: "I regret that I do not know how to connect the several asymmetries I have discussed and the famous asymmetry of entropy" (Lewis, 1986*a*: 51). However, he believed originally that the asymmetry of overdetermination is not a statistical asymmetry; and therefore, by implication, that it is distinct from the thermodynamic asymmetry, to the extent that the latter does rest on a statistical asymmetry. Field (2003: 458) reports that Lewis changed his mind about this, and came to regard the asymmetry of overdetermination as a statistical asymmetry. And an argument due to Elga (2000) makes it very clear that the asymmetry of overdetermination is defensible, if at all, only in this form. Unless we restrict the options in the way that the second law of thermodynamics does, miraculous convergence is ridiculously easy.

Elga's argument exploits a very fundamental feature of a widely accepted statistical explanation of the second law of thermodynamics, the essential elements of which are due to Ludwig Boltzmann (1844–1906). Boltzmann's explanation combines two main ingredients. The first is a statistical consideration. For any macrostate of a physical system which is not already in thermodyamic equilibrium, there are many more microstates compatible with that macrostate whose evolution would be *towards* equilibrium, than microstates which would evolve away from equilibrium. This might seem sufficient to explain the fact in our experience, isolated systems do evolve towards equilibrium.

The flaw in this reasoning was first pointed out by Boltzmann's teacher and colleague, Josef Loschmidt (1821–1895). The statistical considerations are time-symmetric. If they alone imply that entropy increases towards the future, then they alone would also imply that entropy increases towards the past: time-symmetric statistics cannot break the symmetry, to explain the monotonic increase of entropy we actually observe. To explain what we observe, we need to supplement Boltzmann's statistics with a second assumption, a time-asymmetric 'boundary condition'. We need to assume that the observed universe begins in an extremely low entropy condition, at some point in the distant past. Borrowing a term from Feynman (1965: 110), Albert (2001) calls this assumption the *Past Hypothesis* (PH).[11]

Loschmidt's point implies that the *actual* microstate of our familiar universe is always remarkably 'special', in the following sense. The vast

majority of microstates compatible with the actual macrostate are associated with histories very *unlike* that of the actual world (as we believe it to be) - histories in which entropy *increases* towards the past, rather than *decreasing* towards the past. As Elga points out, this means that there is actually a huge *superabundance* of microscopic miracles, providing exactly the cases Lewis's asymmetry of overdetermination is meant to exclude: worlds that converge from very different histories, to differ from the actual world by a tiny local miracle. Without the restriction imposed by PH, in other words, the asymmetry of overdetermination would fail on an absolutely massive scale.

Elga's argument suggests that to the extent that there is an objective physical asymmetry of the kind that Lewis took to ground the asymmetry of counterfactual dependence, it involves macroscopic, statistical phenomena, of the same kind as ordinary manifestations of the thermodynamic asymmetry; dependent, in particular, on the same initial conditions.[12] Indeed, it is tempting to characterise these phenomena, generically, as examples of the dispersal of precisely the kind of macroscopic concentrations of energy that are produced by PH. To the extent that Lewis's intuitions lead us in the direction of a genuine physical asymmetry - a possible candidate for a third arrow - it seems to be this one.

In a moment we turn to an explicit proposal for linking the asymmetry of causal and counterfactual reasoning to PH, from recent work by Albert, Kutach and Loewer. Before that, let's distinguish two questions that need to be raised about Lewis's proposal. First, has Lewis successfully identified an objective

temporal asymmetry with the right distribution to provide a third arrow - has he found a physical asymmetry in more or less the right place? Second, can the resulting account meet PRC - can it account for the asymmetry of deliberation?

We shall return to the latter question in §3.3 and §4 below. Concerning the former, there are some evident difficulties. As Price (1996: Ch. 6) notes, grounding causal asymmetry on a macroscopic statistical asymmetry seems likely to imply that there is no causal asymmetry at a microscopic or substatistical level. True, it is easy to impose an asymmetry at that level by fiat, by using the macroscopic asymmetry as a kind of 'signpost'. But this is much the same as Hume's view, with the reference to earlier and later replaced by reference to the direction in which entropy increases, or something similar. As a result, the same objections apply. Don't we exclude microscopic retrocausality by fiat, for example?

3 Appealing to the Past Hypothesis?

The most explicit attempt to link the asymmetry of causation and counterfactual dependence to that of thermodynamics lies in recent work by Albert (2001), Kutach (2001, 2002, 2007) and Loewer (2007). For present purposes we ignore various differences between these authors, referring to the proposal collectively as the *AKL* view.

The AKL proposal tries to use PH to explain the asymmetry of counterfactual dependence. The basic idea is to argue that in virtue of PH, small,

local changes - the kind of things we could use as 'causal handles', as Albert (2001: 128) puts it - produce much bigger and more diverse changes in the future than they do in the past. Intuitively, PH is supposed to do the job of ensuring that if we wiggle a causal handle in the present, we produce corresponding wiggles in the future but not in the past - or at least, not in the macroscopic, noticeable past. Loewer explains this idea using the figure of a tree, branching to the future but confined to one trunk in the past. PH is supposed to do the job of excluding (macroscopic) branching to the past. The initial plausibility of this idea is easily seen by recalling Elga's objection to Lewis's asymmetry of overdetermination. Elga's demonstration that convergence to the actual world is, *pace* Lewis, actually very easy, relies precisely on Loschmidt's anti-thermodynamic worlds - worlds *without* PH, in other words.

3.1 A web not a tree?

The AKL proposal has been sharply criticised in a series of papers by Mathias Frisch (2005*a*, 2007, forthcoming). In particular, Frisch challenges the claim that PH supplies the required tree structure. In many cases, he argues, the actual structure seems more like a web than a tree. In other words, it contains divergence to the past, as well as the future - which would imply, by AKL's lights, that small, local changes could produce macroscopic changes in the past, as well as the future. For example, Frisch considers a gas in a two-chamber container, which was initially in one of two low entropy conditions: all the gas was in the left chamber, or all the gas was in the right chamber. After the gas has

dispersed between the two chambers, it may well be the case that only tiny local changes separate microstates evolved from the two distinct initial conditions. In this case, the AKL approach seems to imply that a tiny present change could cause the gas to *have been* in one chamber rather than the other. (As Frisch points out, thermodynamics itself implies that this kind of case is likely to be very common, for it is simply a consequence of equilibration.) Frisch also notes that even setting aside this kind of gross counterexample, the AKL approach seems unsatisfactory. The intuitive asymmetries of causation and counterfactual reasoning seem sharper, more general, and far less sensitive to the micro–macro distinction, than the AKL proposal can possibly account for.

3.2 Would a Future Hypothesis prevent us affecting the future?

Another class of objections to the AKL approach rests on the observation that if it were true that PH (in conjunction with the time-symmetric resources noted above) were sufficient to explain our inability to affect the past, then - by symmetry of reasoning - an analogous low-entropy boundary condition in the future would prevent us from affecting the future. But would a 'Future Hypothesis' (FH) have this consequence? We think not.

The first question is whether such a future constraint would imply that our deliberative phenomenology would be a future-directed analogue of what we are trying to explain with respect to the past: the sheer apparent absurdity, at least in ordinary cases, of acting so as influence the past. It is hard to see why this should be so. Restrictions in the distant future - even extreme restrictions, much

tighter than PH itself - seem to have virtually no bearing on our present sense that we can affect the future. Suppose God tells us tells us that as a matter of law, the final state, some fifteen billion years from now, will be constrained within some tiny region of phase space (comparable in size to that required by PH). Better still, suppose he offers to tell us the *actual* final microstate, to as many decimal places as we wish. Either way, the AKL tree of possible trajectories suffers the kind of pruning towards the future that PH requires towards the past. Do we lapse into fatalism, coming to think it absurd that we might seek to influence our immediate future? It is hard to see why we would, or should.[13] Hence, by symmetry, it is hard to see why a remote past hypothesis should be incompatible with taking ourselves to be able to affect the near past.

It might be objected that this argument trades too much on the fact that it considers only a *distant* future constraint. Setting aside the obvious reply that PH is rather distant too, let us turn to consequences of much closer future constraints. Would these necessarily be perceived as making deliberation absurd? On the contrary, we think, they might provide a new degree of control, an influence over matters previously thought to be independent of our actions.

To adapt an old example from the decision theory literature (Gibbard and Harper, 1978: 136), suppose we believe that we are destined to meet Death at noon on a certain day. We regard this as a lawlike future boundary condition.[14] It is now 09:05 on the fateful morning, and we sit in Aleppo airport, with a boarding pass for the flight to Damascus. We know that Death will meet us in one place or other; and moreover (since he refuses to fly) that he is already on the

road to whichever place it is to be. Is it *absurd* to think that we are still free to choose whether to board the plane? On the contrary, apparently. While the boundary condition certainly deprives us of many options - the option to be anywhere other than Damascus or Aleppo at noon, for example, or to be anywhere at all, later in the day - it also yields some new abilities: in particular, the ability to influence Death's movements, even somewhat *earlier* on the day in question.[15]

The example suggests that while a lawlike future constraint can limit the options, it does not make it absurd to think that we exercise control within those limits. Within those limits, its effect seems to be not to prevent us from achieving ends, but to ensure that the world conspires to bring about those ends. Far from preventing us from achieving the ends, in other words, it gives us a new kind of control over *other* events - the ones that need to be appropriately arranged, in the light of the new constraint, for our ends to come to pass. This means, in particular, that we may be able to affect the remote present, and the past, via a kind of zig-zag. We choose the future in some respects, and the future constraint ensures that the remote present and past keep in sync, in order to achieve the required final state. If this is how things would go in a world with lawlike future boundary conditions, shouldn't PH have the same kind of effect? Shouldn't it merely *limit* our capacities to influence the past, and compensate by giving us new powers - powers, say, to affect the remote present, by affecting bits of the past with which the antecedents of the remote present are necessarily correlated?

This possibility has been missed, apparently, because AKL have failed

to notice an ambiguity in the requirement that we consider the consequences of small, *local* changes - causal handles, to use Albert's term. The requirement that the handles be local is needed to avoid a trivial falsification of the theory, because in the assumed context of a deterministic theory, it is immediate that large-scale differences will make a difference at earlier times, as well at later times. But this restriction to small, local handles should not be taken to imply that the *consequences* of wiggling the handles cannot be simultaneous - otherwise we exclude simultaneous causation by fiat.

These considerations play out in two ways. First, and more directly, they suggest that the consequences of lawlike future constraints would be nothing like a future-directed analogue of what we are trying to explain with respect to the past: the sheer apparent absurdity, at least in ordinary cases, of acting so as influence the past. As we have seen, remote constraints provide little inclination to fatalism, and while immediate constraints would certainly restrict our choices, they would also give us new options.

Second, the argument suggests that *microscopic* effects on the distant past - which AKL allow to be a consequence of their view - cannot be prevented from being magnified into less microscopic effects on the less distant past, and the remote present, by means of a zig-zag.[16] The engine of the second stage of this process - the 'zag' by means of which the influence of a present action returns from the distant past - will be the very process of amplification of small differences which is central to the account's own proposal concerning macroscopic branching. Suppose it is true (as the AKL account allows) that, had

I lifted my little finger a moment ago, there would have been differences in the positions a number of atoms, billions of years in the past. What changes might the movement of those ancient atoms have wrought, over such a vast period of time? Not changes enough to dispose of me and my little finger, certainly, for I am here, now, by stipulation, in the history in question. But there is no such protection for the rest of my familiar universe, anywhere within the future light cone of those ancient microscopic changes.

3.3 A general objection to the third arrow strategy?

We conclude that the AKL approach does not yield a satisfactory explanation of the asymmetry of deliberation. Moreover, the argument just outlined suggests a powerful objection to *any* attempt to ground the time-asymmetry of causation on the kind of macroscopic statistical asymmetries we find in our world. As already noted, in seems highly plausible that these asymmetries have their origin in PH. But we have just argued that since FH would not make it absurd to deliberate for future ends, PH cannot explain why we do not deliberate for past ends. So *any* account of the causal arrow which seeks to reduce the time-asymmetry of causation to the kind of asymmetries that derive from PH seems destined to be similarly powerless to explain the time-asymmetry of deliberation - destined, in other words, to share the failings of Hume's proposal in this respect.

This brings us back to a question we deferred in §2. In §1, generalising from this objection to Hume's view, we formulated the Practical Relevance Constraint: an account of the time-asymmetry of causation should be expected to

explain the time-asymmetry of deliberation. In §2, we observed that it is not *obvious* why we should care about counterfactuals in deliberation in the first place, and hence how Lewis's account might deal with PRC (even if succeeds in accounting for the time-asymmetry of counterfactual dependence). We now return to that issue.

4 Why care about counterfactuals?

Can Lewis's account meet PRC? Alternatively, can it maintain that PRC is an optional matter for a satisfactory account of causation? Interestingly, these issues have been on the table for many years, in a different guise. There is a long-standing debate between two rival accounts of rational decision, *causal* decision theory (CDT) and *evidential* decision theory (EDT); and a much-discussed class of cases, known generically as *Newcomb problems,* in which the two theories seem to give different recommendations.

The original Newcomb problem (see Nozick, 1969) goes like this. We are presented with two boxes, one transparent and one opaque. The transparent box contains $1,000, and we are told that the opaque box may contain either $1,000,000 or nothing. We are offered the choice of taking only the opaque box, or taking both boxes. It seems obvious that we should take both boxes, for that way we are $1,000 better off, *whatever* the opaque box contains. However, we are also informed that the choice of what to put in the opaque box is made by an infallible (or almost infallible) predictor, who places $1,000,000 in the opaque

box if and only if he predicts that we will take *only* that box. This information seems to imply that if we take just the opaque box it is very likely to contain $1,000,000; whereas if we take two boxes, the opaque box probably contains nothing. Doesn't it now make sense to take just one box? Isn't a high probability of $1,000,000 much better than a high probability of $1,000? No, says the rival decision principle, for our choice won't *affect* what is in the opaque box - and whatever it is, we're $1,000 ahead if we take both.

Thus 'one-boxers' argue that we should be guided solely by *evidential* considerations (i.e., by EDT), while 'two-boxers' claim that rationality dictates that we consider *causal* or *counterfactual* considerations (as required by CDT). (Lewis himself was a prominent two-boxer.) The connection with our present concerns is that the issue raised by PRC is a more general form of the issue that divides one-boxers and two-boxers. After all, Newcomb problems are precisely problems in which, according to one-boxers, it is appropriate to act for the sake of an end that one does not *cause* - e.g., to raise the evidential probability that the predictor has placed $1,000,000 in the opaque box. The two-boxer's task is to explain why such a decision policy is irrational. And the danger, from the two-boxer's point of view, is that whatever he says about the meaning of cause and effect, the one-boxer is going to respond: "But if that's what these terms mean, then what's wrong with acting for a end which is *not* an effect of one's action?" This is exactly the challenge that PRC raised to Hume's view.

Thus for a view such as Lewis's, a successful response to PRC and a successful defense of two-boxing would amount to much the same thing. What

does the history of these debates tell us about the prospects for such a defense? It reveals a widespread acceptance, even on the part of two-boxers themselves, that there is no such argument to be found. Lewis himself remarks that the debate "is hopelessly deadlocked" (Lewis, 1981*a*: 5). Elsewhere, he puts it like this (Lewis, 1981*b*: 378):

> [I]t's a standoff. We [two-boxers] may consistently go on thinking that it proves nothing that the one-boxers are richly pre-rewarded and we are not. But [one-boxers] may consistently go on thinking otherwise.[17]

These remarks support the following assessment of the status of PRC for Lewis's view of causation (and, apparently, for any other view with a similar investment in the issue between CDT and EDT). On the one hand, such views cannot set aside PRC, for they are heavily committed to the relevance of causation to rational deliberation. On the other hand, they have nothing better to offer than a blunt appeal to intuition, in response to the challenge posed by PRC (or, what comes to the same thing, by the one-boxer's challenge to CDT).

In the present context, our interest is in the asymmetry of causation and deliberation. Our reason for mentioning Newcomb problems was that they illustrate so strikingly the gap between proposing an explanation of the causal asymmetry and providing an explanation of the asymmetry of deliberation. One-boxers personify the challenge of PRC, by defending a conception of

deliberation which doesn't keep step with causation, at least as ordinarily construed.[18]

But Newcomb problems hold a second message for our present concerns. Why are real-life Newcomb problems comparatively rare, and arcane? Largely, apparently, because even *evidential* deliberation displays a marked temporal asymmetry. If this were not so, after all, then the many cases there would then be of evidential deliberation about past ends would themselves be Newcomb problems. The realisation that it is so raises an interesting puzzle, and an inviting prospect. The puzzle is how to characterise and explain this purely evidential asymmetry of deliberation - an asymmetry of an epistemic and 'pre-causal' kind, presumably. The prospect is that once we have succeeded in doing so, we might have the basis for an understanding of causation itself - an understanding which, by incorporating some of the structure of the epistemic perspective, would gain the means to explain the two things that have so far proved illusive: the temporal orientation of causation, and its relevance to deliberation.

5 The time-asymmetry of material deliberation

Consider a typical case in which we believe that *if* we perform an action *A* (which we take to be within our power to perform or not to perform), an outcome *O* will occur; and in which we don't have reason to think that *O* will occur, independently of whether we perform *A*. Interpreted in material terms, what we

believe is simply that the disjunction $\neg A \vee O$ is true. Moreover, we believe it *inferentially,* as we might say - i.e., not simply in virtue of already believing one or other disjunct to be true.[19]

Let's call disjunctions of this form - disjunctions held true on inferential grounds, such that the truth of one disjunct is held to be a matter of future choice - *action-linked inferential disjunctions* (ALIDs, for short). Here's a striking fact about ALIDs. They are common in cases in which the outcome disjunct (O, in our example) concerns a time *after* that of the action disjunct; rare, or perhaps even unknown, in cases in which it concerns a time *before* that of the action disjunct. Call this the *temporal asymmetry of disjunctive deliberation* (TADD).

In the present context, the relevance of TADD is that it reveals a temporal asymmetry which on the one hand is closely linked to deliberation, and on the other seems entirely epistemic in nature - a temporal asymmetry in our typical pattern of disjunctive beliefs about the *actual* world, in cases in which one disjunct concerns one of our own future actions. As we noted, this implies that an account of the causal asymmetry in terms of the counterfactual asymmetry will be blind to at least one significant aspect of the deliberative asymmetry. More intriguingly, it also holds out the prospect that if we could explain TADD then we could also explain everything that needs to be explained about the asymmetries of counterfactuals and causation, if these could be grounded on epistemic or disjunctive deliberation.[20]

Against the latter proposal, it may be objected that there are familiar reasons for distinguishing epistemic from counterfactual deliberation, and for

preferring the latter when the two come apart. After all, the former corresponds to one-box reasoning, the latter to two-box reasoning. The epistemicist argues that he knows that he'll have $1,000,000 if and only if he takes one box; the counterfactualist that if he were to take both boxes, he would be $1,000 richer than if he were to take just one box.[21] But our point is that the present context suggests a novel argument on behalf of one-boxing in these debates. In the present context, even a two-boxer needs to explain TADD - and the two-boxer, of all people, must insist that this is a different matter from explaining the analogous asymmetry of counterfactual reasoning. So TADD is a two-boxer's problem, too. Two-boxers have two temporal asymmetries to explain, in effect. Whereas a one-boxer has the prospect of an argument that TADD is the *only* asymmetry we need, to account for the asymmetry of deliberation.

5.1 What about Cartwright?

It may seem that this prospect is a poor one, in that it collides head-on with the message of a famous paper by Nancy Cartwright (1979). Cartwright argues that causal notions are needed to ground an important distinction between effective and ineffective decision strategies. She describes cases in which evidential and causal deliberation (i.e., EDT and CDT) seem to come apart, and in which it is simply *obvious* that rationality goes with the latter. How, then, could the former kind of deliberation possibly ground the latter?

Our answer is in two parts. First, we note that as subsequent discussion of the kind of decision problems introduced by Cartwright's paper has shown,

clear cases are hard to find. Cartwright's examples include so-called 'medical' Newcomb problems, such as one based on the hypothesis that there is a 'smoking gene' that predisposes both to smoking and to cancer. In this case, Cartwright's argument is that a decision to smoke would be evidence that one has the gene, and hence that one has a higher chance of cancer; but that it would clearly be irrational to refrain from smoking on those grounds, if it is what one would otherwise prefer to do.

In such cases, however, it turns out to be far from clear that a rational agent who believes the smoking gene hypothesis should take her own decision to smoke to be evidence that she herself has the gene. Arguably, her knowledge of distinctive features of her own case renders invalid an application of the relevant statistical generalisations (e.g., that most smokers have the gene) to her own decision.[22] If so, then the obvious irrationality of not smoking this situation rests on faulty evidential reasoning, not on any difference between the recommendations of EDT and CDT. Give EDT the right probabilities, and it, too, recommends that one should smoke.

The remaining cases are both more extreme and far less realistic. For example, they ask us to imagine an agent who has statistical data even about the choices of agents 'just like herself', who have faced exactly her present choice. These cases are much more like the classic Newcomb problem. As well as being highly unrealistic, they share with the classic case the ability to confront us with a deep conflict between seemingly rational intuitions. Hence they are far from clear counterexamples to the approach we are now exploring.

Second, we want to stress that Cartwright's examples *cannot* be clear cases, at least on reflection, if the notion of an effective strategy is to be tied to that of causal or counterfactual reasoning. For in that case, as we have urged, PRC demands an answer. If 'effectiveness' means ... - here plug in your favourite causal or counterfactual story - then why should we care about it? Why not be satisfied with an 'ineffective' but probability-raising strategy?

Cartwright is thus in much the same boat as Lewis. On the one hand, she is heavily invested in the link between causation and rational deliberation, and so cannot afford to set PRC aside, as irrelevant to an account of causal asymmetry. On the other hand, as is revealed both by the inconclusiveness of arguments for two-boxing in the classic Newcomb problem and by the inability of appeals to PH and the thermodynamic asymmetry to account for the stark asymmetry of deliberation, she has very little prospect of a satisfactory response to PRC.

Far from providing a major obstacle to the suggestion that epistemic deliberation be made the basis of everything else, Cartwright's argument thus provides another illustration of how much is to be gained, if the epistemic approach can be made to work. To do so, however, it needs to find an explanation of the temporal asymmetry of material deliberation (without appealing to a primitive causal asymmetry, of course). We now turn to this project.

5.2 Explaining TADD

How are we to explain the asymmetry of disjunctive deliberation? A good first question is whether the deliberative aspect - i.e., the fact that concerns

disjunctions one disjunct of which we take to under our control - is likely to play any crucial role. Or does the asymmetry persist if we move to a slightly larger class of disjunctions, without this restriction?

It is easy to see that the asymmetry does not hold if we impose no restriction at all on the form of the disjuncts. Trivially, any disjunction of the form $X \vee Y$ in which one disjunct concerns matters later in time than the other disjunct is equally a disjunction of which the temporal inverse holds. Following the lead of the AKL approach, however, we might suspect that the asymmetry re-emerges when one disjunct concerns a small, local matter, and the other something larger. In this case, too, a material version of AKL might suggest, disjunctions held true on inferential grounds are always such that the 'small local' disjunct concerns a matter earlier in time than the other disjunct.

This simply isn't true, however. After all, consider disjunctions relating forensic evidence (say) to the past states of affairs for which it is evidence. Small differences in the evidence may be indicative of very different histories at earlier times - that's *why* we pay such close attention to forensic details, of course. Thus it may be true, for example, either (S) that a silver medallion just found in the sand does *not* bear the tiny inscription "CG 1753", or (T) that this beach is the long-lost last resting-place of Captain Greybeard (the oldest sea-dog of his day) - and we may believe $S \vee T$ on inferential grounds - despite the fact that T concerns a matter much earlier than S.

What isn't normally the case, of course, is that we hold true such a disjunction on inferential grounds, *and* believe that the truth of the later disjunct

is under our control. (We might believe that whether the medallion bears the inscription "CG 1753" is under our control, in the sense that we could easily have the inscription added or removed, but in this case we don't hold the disjunct itself true, at least not on inferential grounds.) So the restriction to deliberative cases is crucial to TADD - which raises the question: Is there something temporally asymmetric about agency, about our own deliberative standpoint, that might account for the fact that it seems to introduce an asymmetry in these disjunctive cases, which wasn't present without it?

5.3 The asymmetry of agency

We have just observed that we can't use evidence as a 'causal handle' to influence the earlier states of affairs for which it provides evidence. This suggests that the distinguishing feature of causal handles isn't a temporal-direction-neutral fact about the correlation of small local differences with big remote differences. On the contrary, it seems to lie in the simple fact that we can only wiggle handles which lie in the immediate *future,* with respect to our own deliberations on the matter. If this is right, then the source of the temporal asymmetry of TADD is our own asymmetric perspective as agents - the fact that we are always contemplating actions in the near *future,* with respect to the time of deliberation - not some independent fact about the structure of reality.

Looking at this from the point of view of the matters we contemplate bringing about in deliberation, this asymmetry plays out in a marked temporal asymmetry in associated states of affairs, in the immediate temporal vicinity of

the matters in question. To think of the matters we bring about *as* products of deliberation is to think of them as *having a particular history* - as being immediately *preceded* by our own deliberation, in effect. This makes a huge difference to their evidential significance in that direction, of course, as our last example illustrates: the evidential bearing on past states of affairs of the presence of an inscription on an old medallion is highly sensitive to whether we have just chosen to put it there.

In other words, the very presence of deliberation ensures that the events contemplated in deliberation are *not* typical as regards their associations in the past - for in the past lies the deliberation itself. And yet there is no such restriction in the future. No wonder, then, that that inference *from* the fact of the occurrence of such an event should work so differently in the two temporal directions.

The crucial difference here, compared to the AKL approach, is that we have shifted from considering the evidential consequences of small, local changes *in general* - wiggles of 'causal handles', or local changes produced by agents with *arbitrary* temporal orientation - to thinking of those such changes that are the products of deliberation by agents with *our* temporal orientation: agents for whom actions *follow* deliberation, in the usual time sense. In an account of this kind, then, the asymmetry is being supplied by the asymmetry of our own particular deliberative standpoint, rather than by an objective asymmetry such as PH.[23] It is thus analogous to cases such as those we mentioned at the end of §1: the distinctions between near and far, or local and foreigner, or left and

right.

So far, we are talking about TADD, and hence about cases in which changes are thought of as possible actions. For the moment, the claim is simply that the temporal asymmetry of the deliberative standpoint itself does a good job of accounting for TADD. If we are to make the further claim that the asymmetry of the deliberative standpoint underlies that of counterfactuals and causation in general, it needs to be explained how we are to make the step from this restricted case to the general case - if the asymmetry of the deliberative standpoint is to do the work in the general case, it will need to be argued that when we assess counterfactuals, we think of the antecedents *as* potential actions, with the asymmetry intact. We'll return to this issue in a moment. First, before we leave the relative simplicity of the disjunctive case, it is worth asking whether TADD itself is a strict temporal asymmetry, or whether the account allows for backward-directed - 'retroactive' - disjunctive deliberation.

5.4 Retroactive disjunctive deliberation?

Retroactive disjunctive deliberation ('RetroDD') seems to exist in two varieties. The first is illustrated by our modified Death in Damascus example, from §3.2. In this case, we believe a disjunction of the form:

(We will stay in Aleppo) $\vee$ (Death is already on his way to Damascus) (1)

We believe it on inferential grounds, and we take the first disjunct to be one that we can decide to make true or false, as we wish. So the case meets the criteria for disjunctive deliberation, despite the fact that the second disjunct concerns a time in the past, relative to that of the deliberation. Let's call the pattern exemplified here *zig-zag* RetroDD - it turns on the fact something we can choose to make the case in the future is suitably correlated with a state of affairs in the past (even in the circumstances in which we take ourselves to have the choice).

The second kind of RetroDD - in some sense, a limiting case of the first - is where we take *our choice itself* to be correlated with an earlier event. This is the case associated with medical Newcomb problems, such as the smoking gene example from §5.1. Consider the extreme version, in which the hypothesis is that all and only those who have the gene become smokers. For someone who believed both that this correlation holds, and that he nevertheless had a choice as to whether to smoke, the following disjunction, too, would meet the relevant criteria:

$$(\text{I will not smoke}) \vee (\text{I have the cancer gene}) \qquad (2)$$

Of course, it is hard to imagine why someone should combine both the required beliefs. Prima facie, they seem to be in tension. (Perhaps the original Newcomb problem does as good a job as can be done of presenting a case in which it seems reasonable that we might believe both.) But for the moment, what matters is simply that for someone who did combine them, the result would be an example of RetroDD - we might call it *direct* RetroDD.

We emphasise again that this discussion has been confined to the epistemic case. At this stage, counterfactuals and causal reasoning are simply not in the picture. But the fact that disjunctive deliberation allows, at the margins, for these retroactive cases, implies that if epistemic deliberation can be made the foundation for counterfactual deliberation, then it, too, stands to inherit the same temporal character: overwhelmingly 'past-to-future', though with loopholes for exceptional cases. And as we noted at the beginning, this seems to be precisely what we want of an account of the temporal asymmetry of causation.

6 The attractions of subjectivism

At the end of §1 we observed that if we reject two extreme views - Hume's conventionalism and hyperrealism - we seem to be left with two options for explaining the nature and temporal orientation of the causal arrow. The first, the third arrow approach, looks for some objective physical asymmetry to ground the causal asymmetry. We argued that the only apparent candidate, some sort of *de facto* statistical asymmetry linked to the thermodynamic asymmetry, seems unpromising. For one thing, it reduces to something very much like Hume's view in the case of microscopic and substatistical systems, where the causal asymmetry becomes nothing more than a conventional label, applied to mark alignment with a macroscopic statistical asymmetry. For another, its link to deliberation is at best obscure. In particular, the statistical asymmetry does a poor job of explaining why we don't (typically) deliberate with respect to past ends.

The second option, we noted, is to agree with Hume that there is no intrinsic asymmetry of causation, but to look for some better story than Hume's own about why our causal notions show such a strong and temporally-asymmetric asymmetry. In §§4–5 we have been investigating the credentials of one obvious candidate for the beginnings of such a story, viz., our own perspective as agents and deliberators. We have discovered that if we think of deliberation, initially, in epistemic, evidential or 'pre-causal' terms, it nevertheless exhibits a strong temporal asymmetry: an asymmetry explicable, apparently, in terms of our own asymmetric temporal orientation, as 'players' in the dynamical environments in which we live; and an asymmetry that allows, at the margins, for the epistemic analogue of retrocausality.

This is a very striking result. If it could be elaborated into a plausible explanation (or better, *genealogy*) of our ordinary causal concepts, and of associated matters, such as counterfactual reasoning, it would tick all the hard boxes, apparently. It wears its link with deliberation on its face, so there are no problems with PRC. It has good physicalist credentials so long as the notion of agency itself does: in other words, so long as biology and physics can account for the existence of creatures like us; and it links to the thermodynamic asymmetry so long as that explanation does so. It gets the character of the temporal asymmetry just about right: predominantly though perhaps not universally past-to-future, in our time sense (and plausibly linked to *de facto* physical asymmetries, for the reason just mentioned). It gets the scope of the causal asymmetry just about right, too, in the sense that so long as our

deliberative perspective is blind to the micro–macro distinction, then so is the causal asymmetry. And it makes it immediately obvious, in a way that Hume's own conventionalism does not, why we have an interest in marking (what we come to call) the cause–effect distinction: we thereby mark something of first importance, from an agent's perspective.

Despite these advantages, many philosophers feel that this approach to the causal asymmetry gives away too much: it renders the causal asymmetry insufficiently objective. It is worth noting, however, that there is one sense in which this battle has already been lost. The main rival, the statistical view, has already conceded that there is no instrinsic asymmetry at a fundamental level. Critics thus do better to focus their attention on the challenges of the project of turning the subjectivist's proposed raw materials into a plausible genealogy for our causal concepts and cognitive machinery.

We cannot explore the prospects for that project here, but we close with a suggestion about how to think of the 'subjectivism' of this view, and with two notes about how it might tie in in interesting ways with aspects of the theory of causation normally thought of in other ways.

6.1 A subjectivist's guide to objective causation?

The project is to ground the asymmetry and practical relevance of causation on that of deliberation, epistemically construed. This idea seems strikingly analogous to a viewpoint long familiar in the case of probability. In that case, probabilistic 'subjectivists' are united by the thought that a proper account of

probability needs to begin on the practical and epistemic side - i.e., with *credence,* defined in terms of its role in decision under epistemic uncertainty. Not all subjectivists think of this as incompatible with recognising more objective notions of probability as well, but their common motto is that if an account of probability doesn't build the link with decision in at the beginning, it will never be able to recover it later - never be able to justify the link between objective probability and credence which Lewis calls the Principal Principle.[24]

We suggest that the lesson of PRC be viewed in the same light, and be called 'subjectivist' for the same reason. Indeed, PRC itself seems to play a role analogous to that of the Principal Principle. And subjectivism here consists in reading its implications in a similar way: unless an account of causation starts with deliberation, epistemically construed, it is not going to be able to explain why causation matters to deliberation, in the way that it does. As in the case of probability, this starting point leaves room for a range of possible views, at the more objectivist end of which might be causal analogue of Lewis's view of chance. But what these views will have in common will be a recognition that for causation, as for probability, the practical, epistemic perspective is importantly prior to the metaphysical perspective.

6.2 Folk physics and the fixity of the past

We noted earlier that Lewis observes that one might treat the asymmetry of counterfactual dependence as the product of a convention - a stipulation that when we assess counterfactuals, we 'hold the past fixed.' He rejects this option

for much the same reasons that many philosophers reject Hume's conventionalism, e.g., that it puts the asymmetry in by hand, and rules out backward dependency by fiat. But the subjectivist view gives new interest to the idea that counterfactual reasoning might be governed by such a convention. If the relevant species of counterfactual reasoning develops from the kind of hypothetical reasoning needed in epistemic deliberation, the principle that one should hold the past fixed provides a simple codification of the asymmetry of the deliberator's perspective - a codification that won't lead to problems, apparently, so long as the environment does not supply the kind of rare opportunities that might favour retroactive deliberation.

Hence it is tempting to suggest that the fixity of the past has the status of a useful piece of folk physics, deeply ingrained as our ancestors developed the cognitive framework that supports deliberation. With this hypothesis in place, subjectivists are free to help themselves to an asymmetry of counterfactual dependence, grounded on the (now explicable) convention that Lewis rejects; and hence, if they wish, to the resources of a counterfactual account of causal reasoning. They are also free to discuss possible modifications in the folk physics, e.g. to accommodate retrocausality in the kinds of cases to which Lewis himself calls our attention (see fn. 1).[25]

6.3 Interventionism

Much recent work has focussed on links between causation and what has come to be called *intervention*. Roughly, an intervention is a 'surgical' input into a system

of correlated variables, that sets the value of a particular variable, breaking the normal links between it and its causal 'parent'. As Woodward (2001) puts it:

> [T]he intervention disrupts completely the relationship between [a variable X] and its parents so that the value of [X] is determined entirely by the intervention. Furthermore, the intervention is surgical in the sense that no other causal relationships in the system are changed.

The basic proposal is then that the effects of X are the dependencies that survive when the value of X is fixed by an intervention of this kind. As Woodward goes on to note, this may be seen as a formalisation of the central idea of manipulability approaches to causation, such as that of Menzies and Price (1993):

> In this way, we may capture Menzies' and Price's idea that X causes Y if and only if the correlation between X and Y would persist under the right sort of manipulation of X.

It seems clear that this connection will be of great importance to any attempt to develop a subjectivist approach to the causal asymmetry. Ideally, the subjectivist will want to step into the interventionists' shoes - all the more so, now that Pearl, Woodward and others have shown us how far those shoes may

take us![26] On the face of it, the shoes seem to fit extremely well. The defining feature of an intervention is that it breaks 'upsteam' links, in a very similar manner, apparently, to the way in which we have seen that the mere presence of the deliberating agent breaks links to the past, in the case of disjunctive deliberation. In a sense, the main issue is who owns the shoes in the first place. Does deliberation need to be explained as a species of intervention, in other words, or is deliberation the primary notion?

Subjectivists put their money on the latter option, and we close by noting two sorts of argument they may offer, drawing on the conclusions of our earlier discussion. The first claims that only our contingent temporal asymmetry as agents can account for the fact that the class of interventions relevant to ordinary causal judgements are interventions 'from the past', not 'from the future'. (As before, subjectivists claim that their view explains an asymmetry that other views must treat as primitive, or simply leave unexplained.)

The second argument appeals to PRC, and turns the tables on popular objections to subjectivism in an interesting way. It is often objected that manipulability theories of causation will be circular, because manipulation is a causal notion. But we have now seen that deliberation can be characterised in a non-causal, epistemic fashion. As long as deliberation is construed in epistemic terms, in other words, it is simply not true that the manipulability theory relies on a causal notion at this point.[27] Whereas if intervention is the basic notion, then not only does it rely for its characterisation on causal notions, rendering circular any *analysis* of causation in interventionist terms;[28] but this also leaves it

vulnerable to the challenge of PRC. What is it about *that* causal notion - whatever it is - that renders it relevant to deliberation?[29]

6.4 Summary

There is a considerable consensus that there is no fundamental, intrinsic asymmetry of causation. To that extent, Hume and Russell seem to have been right: there is no asymmetric causation in Sellars' 'scientific image', at least at its most basic level. Concerning the 'manifest image' - the explanation of the asymmetry and temporal orientation of ordinary concepts and judgements about causation, and of related matters, such as deliberation and counterfactual reasoning - the most promising strategy seems to be to begin with the *de facto* asymmetry of human deliberation, characterised in epistemic terms, and to build out from there. More than any rival, this subjectivist approach promises to demystify the asymmetry, temporal orientation, and deliberative relevance of our causal judgements.

In a recent survey article about causation, much concerned with the issue of temporal asymmetry, Hartry Field (2003: 443) remarks:

> [W]e have a problem to solve: the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics. This is probably the central problem in the metaphysics of causation.

We have suggested, in effect, that the best option is to move the problem from metaphysics to pragmatics. So long as we see the problem as one of explaining the practical relevance of causal notions, in the lives of creatures in our situation, there is some prospect of reconciliation.

Further Reading

Reichenbach (1956) developed the first third arrow account, grounded in a probabilistic theory of causation. Horwich (1987) and Hausman (1998) are more recent theories developed along broadly similar lines. Lewis (1979) first proposed the counterfactual overdetermination account, to which Elga (2000) provides an important objection. Field (2003) is a useful survey covering all of these accounts. Recent examinations of causal asymmetry in the context of fundamental physical theories are Price (1996), Albert (2001) and Frisch (2005*b*). Finally, Price and Corry (2007) is a collection of recent papers on causation and physics, many of which address the issues at hand.

Notes

[1]In a closely related context (see §2.1 below) David Lewis (1986*a*: 40–41) puts the point like this:

> Careful readers have thought they could make sense of stories of time travel . . . ; hard-headed psychical researchers have believed in precognition; speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models models with closed timelike curves. . . . It will not do to declare [these phenomena] impossible *a priori*.

[2]As Michael Dummett (1954: 28) puts it:

> Why should we lay down temporal precedence as a defining property of a cause? If we can observe that an event of a certain kind is a sufficient condition of an earlier event of some other kind, it does not seem to matter much whether we choose to call the later event the "cause" of the earlier or not: the question rather is why we should not use this observed regularity as we use those that operate from earlier to later; why, when we do not know whether or not the earlier event has occurred, we should not bring about the later event in order to ensure that the earlier had occurred.

[3]Russell's main claim was that the evolution of physical systems can be fully described without employing causal vocabulary, but he also calls attention to the time-symmetry of the determination relations one finds in physics.

[4]See the chapter on anti-reductionism for more, and Tooley (1987, 1990) as an example of such a view.

[5]We emphasise that our objection here is not that causation is not directly experienced or that

it is not analysable in terms of experiences. Rather, it is that hyperrealism entails that causal facts are underdetermined by *all* available non-causal evidence.

[6]An early proponent of this view was F. Ramsey ([1929] 1931: 146), who says that "from the situation when we are deliberating seems to ... arise the general difference of cause and effect".

[7]He later revised the analysis (Lewis, 2000), but our remarks here are independent of these details.

[8]In what follows we gloss over a number of details that are irrelevant to our discussion. See the chapter on counterfactual analysis of causation for more.

[9]See fn. 1.

[10]Lewis did not give a precise definition of a miracle, but Frisch (2005*b*: 170–173) persuasively argues, partly by appealing to quotes from Lewis, and partly by responding on behalf of Lewis to an objection made by Arntzenius (1990), that Lewis requires a spatiotemporal locality condition on miracle size. We take this for granted in what follows.

[11]Albert argues that PH has the status of an additional physical *law* - a view supported by Callender (2004), for example.

[12]A similar conclusion follows from Frisch's (2005: Ch. 8) demonstration that there is no fundamental asymmetry of overdetermination in classical electromagnetism.

[13]We would become fatalists about the *distant* future, presumably, believing that nothing we could do would prevent the final state of which God had informed us. But this would have little or no immediate relevance, except under very unrealistic suppositions about what we would otherwise take ourselves to be able to achieve.

[14]It specifies our own future boundary!

[15]We suppose that we cannot tell which way Death is heading before we make our choice. (We know that he takes mortal form, but not *which* mortal form, so it wouldn't help to have CCTV footage of all the travellers currently moving in either direction.) We thus avoid the so-called bilking argument, exploiting the loophole identified by Dummett (1964).

[16]This kind of zig-zag is discussed by Kutach (2002).

[17]Cf. Horgan (1981). Horgan notes an apparently ineliminable circularity in the two-boxers'

attempt to justify two-boxing.

[18]The qualification is needed because a one-boxer may want to argue that ordinary causal intuitions are misleading in Newcomb cases: perhaps the decision to one-box should be seen as *retrocausing* the presence of §1,000,000.

[19]In other words, we don't yet believe either disjunct, but would infer each from the falsity of the other. We have put this in terms of disjunctions rather than material conditionals simply to lessen the need to emphasise that the connectives in question *are* material.

[20]The claim that hypothetical reasoning is more fundamental than counterfactual reasoning - that 'counterfactuals are the price we pay for hypotheticals', as Alison Gopnik () puts it - is certainly not new. But the present argument suggests a new defense of this view, based on the argument that *only* an approach which grounds counterfactuals on hypotheticals can account for the asymmetry of deliberation.

[21]Note that this case provides an exception to TADD: the agent is presented as believing (a) that either he'll take both boxes or there is *already* $1,000,000 in the first; and (b) that he really has a choice as to whether to take one box or two. More on this below.

[22]See Eells (1981, 1982), Horwich (1987: Ch. 11), Price (1986, 1991) and Horgan (1981) for arguments of this kind.

[23]It is compatible with this account that PH might be needed to explain the existence of such asymmetric deliberators, as to explain the existence of very much else of a time-asymmetric nature in the world we observe. But this does not *reduce* the account here suggested to the AKL account.

[24]See Lewis (1980). Lewis himself thus counts as a subjectivist - even if surely at the objectivist end of the subjectivist spectrum! - in virtue of taking the Principal Principle to be an analytic element of any satisfactory theory of chance.

[25]Another advantage of this approach, compared to Lewis's own, is that it solves the problem of *transitions*. In Lewis's version the small miracle on which the time-asymmetry depends needs to be displaced somewhat from the antecedent in question - which implies that there will always be events between the miracle and the antecedent that turn out to counterfactually depend on the antecedent (e.g., if the vase had smashed later, it would have fallen through the air earlier). In the

hypothetical case no such problem arises: the transition is simply our deliberation itself, which we always think of as issuing in rather depending on the ensuing actions.

[26]See, e.g., Pearl (2000), Spirtes, Glymour, and Scheines (2000) and Woodward (2001).

[27]This is not to deny that an agent already equipped with causal concepts will regard her deliberations as causes of the acts to which they give rise, but only that the deliberations themselves *depend* on possession of causal concepts. The former circularity is not vicious, whereas the latter would prevent the proposed genealogy from leaving the ground.

[28]As writers such as Woodward (2001, 2003) have emphasised.

[29]Referring to Woodward's (2003) theory, Weslake (2006: 139) puts the point like this:

> [G]iven that any variety of counterfactual meeting the criteria of an intervention will give us a variety of manipulation, why is it only some subset of these that we are interested in? Why shouldn't we abandon counterfactual for counterfactual*, especially if counterfactual* will enable us to cause* past events? The answer . . . is that we can't, in fact, bring about counterfactual* antecedents (at least in all cases we know of) - but this is in part a fact about the sorts of agents we are.

References

Albert, David Z. 2001. *Time and Chance*, Harvard University Press, Cambridge MA.

Arntzenius, Frank. 1990. "Physics and Common Causes", in *Synthese*, Vol. 82, No. 1, January 1990, pp. 77–96. URL: http://dx.doi.org/10.1007/BF00413670.

Callender, Craig. 2004. "Measures, Explanations and the Past: Should 'Special' Initial Conditions be Explained?", in *British Journal for the Philosophy of Science*, Vol. 55, No. 2, pp. 195–217. URL: http://dx.doi.org/10.1093/bjps/55.2.195.

Campbell, Richmond, and Lanning Sowden. 1985. *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, University of British Columbia Press, Vancouver.

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies", in *Noûs*, Vol. 13, No. 4, November 1979, pp. 419–437.

Collins, John, Ned Hall, and L. A. Paul. 2004. *Counterfactuals and Causation*, MIT Press, Cambridge MA.

Dummett, Michael. 1954. "Can an Effect Precede Its Cause", in *Proceedings of the Aristotelian Society Supplement*, Vol. 28, No. 3, pp. 27–44.

————. 1964. "Bringing About the Past", in *The Philosophical Review*, Vol. 73, No. 3, July 1964, pp. 338–359.

Eells, Ellery. 1981. “Causality, Utility, and Decision”, in *Synthese*, Vol. 48, No. 2, August 1981, pp. 295–329. Reprinted in Eells (1982). URL: http://dx.doi.org/10.1007/BF01063891.

———. 1982. *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Elga, Adam. 2000. “Statistical Mechanics and the Asymmetry of Counterfactual Dependence”, in *Philosophy of Science*, Vol. 68, No. 3, September 2000, pp. S313–S324. URL: http://dx.doi.org/10.1086/392918.

Feynman, Richard. 1965. *The Character of Physical Law*, MIT Press, Cambridge MA.

Field, Hartry. 2003. “Causation in a Physical World”, in Michael J. Loux and Dean W. Zimmerman (Ed.), *The Oxford Handbook of Metaphysics*, Oxford University Press, Oxford, pp. 435–460. URL: http://philosophy.fas.nyu.edu/docs/IO/1158/Cause.pdf.

Frisch, Mathias. 2005*a*. “Counterfactuals and the Past Hypothesis”, in *Philosophy of Science*, Vol. 72, No. 5, December 2005, pp. 739–750. URL: http://dx.doi.org/10.1086/508111.

———. 2005*b*. *Inconsistency, Asymmetry, and Non-locality: A Philosophical Investigation of Classical Electrodynamics*, Oxford University Press, Oxford. URL: http://dx.doi.org/10.1093/0195172159.001.0001.

———. 2007. “Causation, Counterfactuals, and the Past-Hypothesis”, in Price and Corry (2007).

Frisch, Mathias. forthcoming. “Does a Low-Entropy Constraint Prevent Us from Influencing the Past?”, in Gerhard Ernst and Andreas Hüttemann (Ed.), *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*, Cambridge University Press, Cambridge. URL: http://philsci-archive.pitt.edu/archive/00003390/.

Gibbard, Alan, and William Harper. 1978. “Counterfactuals and Two Kinds of Expected Utility”, in Clifford Hooker, James Leach, and Edward McLennen (Ed.), *Foundations and Applications of Decision Theory Foundations and Applications of Decision Theory*, D. Reidel, Dordrecht, Holland, pp. 125–162.

Gopnik, Alison. “Personal communication”.

Hausman, Daniel M. 1998. *Causal Asymmetries*, Cambridge University Press, Cambridge.

Hawking, Stephen W. 1994. “The No Boundary Condition and the Arrow of Time”, in Jonathan J. Halliwell, Juan Pérez-Mercader, and Wojciech Hubert Zurek (Ed.), *Physical Origins of Time Asymmetry*, Cambridge University Press, Cambridge, pp. 346–357.

Horgan, Terence. 1981. “Counterfactuals and Newcomb’s Problem”, in *The Journal of Philosophy*, Vol. 78, No. 6, June 1981, pp. 331–356.

Horwich, Paul. 1987. *Asymmetries in Time: Problems in the Philosophy of Science*, MIT Press, Cambridge MA.

Kutach, Douglas. 2001. “Entropy And Counterfactual Asymmetry”. PhD thesis. New Brunswick NJ: Rutgers University. URL: http://www.brown.edu/Departments/Philosophy/Douglas_Kutach/Kutach-Dissertation.pdf.

———. 2002. “The Entropy Theory of Counterfactuals”, in *Philosophy of Science*, Vol. 69, No. 1, March 2002, pp. 82–104. URL: http://dx.doi.org/10.1086/338942.

———. 2007. “The Physical Foundations of Causation”, in Price and Corry (2007). URL: http://www.brown.edu/Departments/Philosophy/Douglas_Kutach/Kutach-PhysicalFoundationsofCausation.pdf.

Lange, Marc. 2006. *Philosophy of Science: An Anthology*, Blackwell, Malden MA.

Lewis, David. 1973. “Causation”, in *The Journal of Philosophy*, Vol. 70, No. 17, October 1973, pp. 556–567. Reprinted in Lewis (1986*a*: 159–171). URL: http://dx.doi.org/10.2307/2025310.

———. 1979. “Counterfactual Dependence and Time’s Arrow”, in *Noûs*, Vol. 13, No. 4, November 1979, pp. 455–476. Reprinted in Lewis (1986*a*: 32–52); page references are to the latter version. URL: http://dx.doi.org/10.2307/2215339.

———. 1980. “A Subjectivist’s Guide to Objective Chance”, in Richard C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Volume 2, University of California Press, Berkeley, pp. 263–293.

Reprinted with postscripts in Lewis (1986*a*: 83–132). URL: http://dx.doi.org/10.1093/0195036468.003.0004.

———. 1981*a*. "Causal Decision Theory", in *Australasian Journal of Philosophy*, Vol. 59, No. 1, March 1981, pp. 5–30. Reprinted in Lewis (1986*a*: 305–339); page references are to the latter version.

———. 1981*b*. "'Why Ain'cha Rich?'", in *Noûs*, Vol. 15, No. 3, September 1981, pp. 377–380.

———. 1986*a*. *Philosophical Papers*, Volume II, Oxford University Press, New York.

———. 1986*b*. "Postscripts to "Counterfactual Dependence and Time's Arrow" ", in Lewis (1986*a*), pp. 52–66.

———. 2000. "Causation as Influence", in *The Journal of Philosophy*, Vol. 97, No. 4, April 2000, pp. 182–197. Reprinted in Collins, Hall, and Paul (2004: 75–106) and Lange (2006: 466–487). URL: http://dx.doi.org/10.2307/2678389.

Loewer, Barry. 2007. "Counterfactuals and the Second Law", in Price and Corry (2007), pp. 293–326.

Menzies, Peter, and Huw Price. 1993. "Causation as a Secondary Quality", in *British Journal for the Philosophy of Science*, Vol. 44, No. 2, June 1993, pp. 187–203.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice", in Nicholas Rescher (Ed.), *Essays in Honor of Carl G. Hempel*, Reidel,

Dordrecht, pp. 114–146. Reprinted in Nozick (1997: 45–74) and Campbell and Sowden (1985: 107—133).

———. 1997. *Socratic Puzzles*, Harvard University Press, Cambridge MA.

Pearl, Judea. 2000. *Causality*, Cambridge University Press, Cambridge.

Price, Huw. 1986. "Against Causal Decision Theory", in *Synthese*, Vol. 67, No. 2, May 1986, pp. 195–212. URL: http://dx.doi.org/10.1007/BF00540068.

———. 1991. "Agency and Probabilistic Causality", in *British Journal for the Philosophy of Science*, Vol. 42, No. 2, June 1991, pp. 157–176. URL: http://dx.doi.org/10.1093/bjps/42.2.157.

———. 1996. *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*, Oxford University Press, Oxford.

Price, Huw, and Richard Corry. 2007. *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford.

Ramsey, Frank Plumpton. [1929] 1931. "General Propositions and Causality", in Richard B. Braithwaite (Ed.), *The Foundations of Mathematics and other Logical Essays*, Kegan Paul, Trench, Trübner, London, pp. 237–255. Reprinted in Ramsey (1978: 133–151). Page references are to the latter edition. URL: http://www.dspace.cam.ac.uk/handle/1810/194722.

Ramsey, Frank Plumpton. 1978. *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, D. H. Mellor (Ed.), Routledge, Kegan Paul, London.

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

Russell, Bertrand. 1912–1913. "On the Notion of Cause", in *Proceedings of the Aristotelian Society*, Vol. 13, pp. 1–26.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction and Search*, 2nd edition, MIT Press, Cambridge MA. URL: http://cognet.mit.edu/library/books/view?isbn=0262194406.

Tooley, Michael. 1987. *Causation: A Realist Approach*, Clarendon Press, Oxford.

————. 1990. "Causation: Reductionism versus Realism", in *Philosophy and Phenomenological Research*, Vol. 50, Supplement, Autumn, 1990, pp. 215–236.

Weslake, Brad. 2006. "Review of *Making Things Happen*", in *Australasian Journal of Philosophy*, Vol. 84, No. 1, March 2006, pp. 136–140. URL: http://dx.doi.org/10.1080/00048400600571935.

Woodward, James. 2001. "Causation and Manipulability", in Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*, Stanford University, Stanford. URL: http://plato.stanford.edu/entries/causation-mani/.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, New York. URL: http://dx.doi.org/10.1093/0195155270.001.0001.