Can We Reduce Causal Direction to Probabilities?

David Papineau

King's College London

This paper will be a mirror image of Huw Price's paper. Like Price, I shall first introduce the "fork asymmetry", then consider objections to using it to analyse causal asymmetry, and finally examine the relation between causation and human agency. But where Price rejects analyses of causal asymmetry in terms of the fork asymmetry, and so turns to human agency to ground an alternative account of both the causal and fork asymmetries, I defend the fork asymmetry analysis of causation, and correspondingly want to explain agency in terms of the fork asymmetry, rather than vice versa.

1. The Fork Asymmetry

The approach I shall be defending is a version of what Price calls the "third arrow" strategy. I want to analyse the direction of *causation* independently of the direction of *time* by finding some *further* objective asymmetry within time. (Though my view of the conceptual and metaphysical relationships between these three arrows is rather different from Price's. More on this in sections 2 and 3 below.)

The idea that we can use the fork asymmetry as such a third arrow is not a new one. It goes back to Reichenbach (1956), and relevant recent discussions include Lewis (1979), Ehring (1982), Hausman (1984), Papineau (1985), Horwich (1987) and Arntzenius (1990). In this first section I shall consider three different versions of the fork asymmetry, namely the asymmetry of overdetermination, the screening-off asymmetry, and the probabilistic graph asymmetry.

Perhaps the most easily graspable version of the fork asymmetry is David Lewis' *asymmetry of overdetermination*. Lewis (1979) has pointed out that causes are characteristically overdetermined by their effects, but effects are rarely overdetermined by their causes—any given cause will characteristically generate a large number of different chains of effects, any one of which provides grounds for thinking that the cause occurred earlier; but any given effect will normally only have one such chain of causes.

Of course, if this asymmetry is to ground an analysis of causal direction, we need to be able to phrase it without using the contrasting terms "cause" and "effect". Well, we could say that any given event is characteristically overdetermined by *later* events,

PSA 1992, Volume 2, pp. 238-252

but not by *earlier* events. But we need to say this without using earlier and "earlier" and "later" either, if we want the arrow of causation not to presuppose the arrow of time. So what we should say is this: given any event C, in one direction in time there will be many different sequences of events each of a type which is generally found with C, while in the other direction in time there will only be one such sequence. And then we can say that the former sequences of events are the effects of C, and the latter sequence its causes.

Let me note in passing that this Lewis asymmetry offers a natural explanation of why it is so easy to have knowledge of the past, but so hard to have knowledge of the future. Namely, that the present normally contains a large number of traces of the past, but only one set of circumstances that fixes the future. This is not the place to go into details, but I think this thought applies quite generally, explaining not only why the history archives tell us more about distant past battles than future ones, but also why human vision can tell us about our immediate past but not about our immediate future.

I now want to introduce the *screening-off asymmetry*, which can be thought of as a probabilistic version of Lewis' asymmetry. One advantage of putting the matter in probabilistic terms is that it will enable us to see more clearly the difficulties facing this kind of reduction of causation. (A proleptic remark. People often ask me what I mean by probability in this kind of context. I can't answer. Beyond the thought that my probabilities are objective, and not just subjective degrees of belief, I have no worked-out interpretation. I am talking about probabilities like the probability of 10p pieces coming down heads when tossed, the probability of getting cancer if one smokes forty a day, the probability of drawing to an inside straight. Even if we don't have an adequate philosophical interpretation of these probabilities, they permeate our lives, and we know lots about them. So there's no reason why we should not consider what follows from them.)

The *screening-off asymmetry* says that the joint effects of common causes are correlated, and that these correlations are screened off by the common cause; on the other hand, the joint causes of common effects are not correlated.

Let A and B the joint effects of common cause C. Then

Prob(B/A) > Prob(B) — A and B are correlated

But

Prob(B/A & C) = Prob(B/C) and
Prob(B/A & not-C) = Prob(B/not-C) — C "screens off" A from B.

On the other hand

if D and E are the joint causes of common effect F, then we won't find D and E correlated to start with.

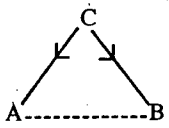
Again, if this asymmetry is to ground an analysis of causal direction, we need to be able to phrase it without using .the contrasting terms "cause" and "effect". So we might say this:

(S-O) Take any event C. Then among the events which are correlated with C will be some that are correlated with each other in such a way that their correlation

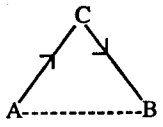
is screened off by C—these are C's effects; and among the events which are correlated with C will also be some that are not correlated with each other—these will be C's causes.

In a moment I shall show that this suggestion is too crude. But first let me point out that the screening-off asymmetry is very similar to Lewis' asymmetry of overdetermination. For the screening-off asymmetry implies that the various effects of a joint cause will tend to occur together (since they are correlated), and moreover that this tendency is due to their common tendency to occur when their common cause does (the common cause screens off their correlations). But this is just Lewis' thought that causes are characteristically followed by clusters of effects.

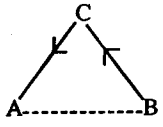
However, as I said, the simple screening-off account of causal asymmetry is too crude. It is not in fact true, as (S-O) claims, that whenever C screens off a correlation between two items A and B, then C is a common cause of A and B. For we will find exactly the same probabilistic structure if C is causally intermediate between A and B; that is, if A causes C, which causes B. What is more, we will find exactly the same probabilistic structure if B causes C, which causes A. That is, the following three causal diagrams (where arrows indicate causal links, and dotted lines screened-off correlations) are all consistent with C screening off a correlation between A and B.



diag 1

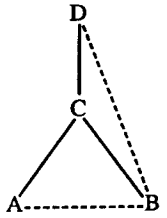


diag 2



diag 3

This problem is not unfamiliar to advocates of fork asymmetry accounts of causal direction, and it is clear enough in outline how to deal with it. We can observe that even if diagrams 1-3 are in themselves probabilistically indistinguishable, they can easily become distinguished when they are embedded in some wider structure. So, for example, if there is some further event D which is correlated with C and B, but not with A, and whose correlation with B is screened off by C, but whose correlation with C isn't screened off by anything, then it is intuitively clear that D and A are causes of C, and B its effect, that is, that diagram 2 is correct, rather than 1 or 3.



diag 4

(unbroken lines indicate unscreened-off correlations)

However, while this line of response has been gestured at in the literature, there has, so far as I know, been no systematic development of the idea.

Recent work by Spirtes, Glymour and Scheines (1992) shows how to fill this gap. These authors are themselves more interested in methodology than metaphysics. They want to show that surprisingly strong causal conclusions can often be derived from relatively meagre probabilistic data, contrary to received wisdom in econometrics and allied disciplines. But their arguments also establish the conclusion needed for a metaphysical reduction of causation: namely, that *every* causal conclusion will be fixed, given sufficient information in the form of a *probabilistic graph* detailing conditional and unconditional correlations between variables.

This is not the place to explore all the technical details of their analysis, both for reasons of space, and because I am not sure I fully understand them fully. But in outline their strategy is to make the following assumptions about causation:

(I) Pairs of events which have common causal ancestors or which cause each other are probabilistically correlated.

(II) Pairs of events which are not causally related as in (I) are probabilistically independent.

(III) Causal ancestors and causal intermediaries screen off the correlations required by (I).

Spirtes, Glymour and Scheines then show how these assumptions place constraints on the possible causal orderings that might obtain between probabilistically related variables. Of course, as diagrams 1-3 make clear, limited probabilistic information might fail to identify a unique causal structure. But then, as in diagram 4, further probabilistic information can. Spirtes, Glymour and Scheines show (cf their theorem 4.6):

For any set of probabilistically related variables, there is a possible wider such set such that assumptions (I)-(III) will fix the causal order of the original variables.

It is of course a contingent matter whether such a *possible* wider set is *actually* available in every case, that is, whether, for any causally ambiguous probabilistic structure (as in diagrams 1-3), there is always a wider structure (diagram 4) which disambiguates it. Let us assume that there is. Our entitlement to this assumption will be discussed further in the next section. But for the moment let me simply observe that, if we do make this assumption, then we have an effective reduction of causation to probabilities.

For this assumption means that the causal facts are always fixed by the probabilistic facts, in virtue of (I) - (III). We can thus take (I) - (III) to show us how there is nothing more to the causal facts than certain kinds of probabilistic facts.

Of course, (I) - (III) *use* the notion of cause, and it may not be immediately obvious in what sense they offer an analysis of it. But suppose (I) - (III) were framed, not as a set of statements about causes, but rather as a set of constraints on admissible ways of drawing arrows between probabilistically correlated sets of variables. The above points imply that there will only be one way of drawing arrows consistent with these constraints and the actual probabilistic facts. And this, I suggest, *is* causal direction. Causal direction is simply that ordering of correlated variables that satisfies (I) - (III).

(I) - (III) go beyond the simple principle that common causes screen off correlations between their joint effects. But they do include it, so in the next section I shall consider some of the standard objections to this principle.

However, let me first briefly observe that the common cause principle is not entirely independent of our other assumptions about causes. For if we assume determinism, then assumption (III) follows from (I) and (II). To assume determinism, in the present context, means that we observe conditional probabilities other than 0 and 1 only because we are conditioning on less than complete information. If we assume in addition that the further factors which make up complete causes are probabilistically independent of each other (cf assumption (II)), then the screening-off property of common causes follows. (Cf Papineau 1985; Horwich 1987; Cartwright 1989; Spirtes, Glymour and Scheines *op cit* 3.5.1.) With genuinely indeterministic systems, on the other hand, the screening-off property of common causes is not guaranteed, and so needs to be added as an independent extra assumption. (This then offers a possible *explanation* of the common cause principle: namely, that the causal systems with which we are familiar are (effectively) deterministic systems, in which the exogenous causes are probabilistically independent. I myself find this explanation plausible; but nothing except a couple of remarks right at the end of this paper will rest on it.)

2. Objections to the Fork Asymmetry

(i) Pre-Established Harmonies

Aren't there plenty of correlations which don't have a common cause, like the correlation between the price of bread in Great Britain and the sea level in Venice, both of which have been increasing steadily for some time? (Cf Sober 1988.)

An initial response might be that this simply misses the target of the common cause principle, if this is understood as the claim that common causes always screen off correlations between their effects. For this claim only requires that, *if* there is a common cause, then it will screen off the correlation between its effects, not that every correlation has a common cause. So examples of correlations without common causes are beside the point.

But this would be too quick. For our discussion has now committed us to a stronger version of the common cause principle. The Spirtes-Glymour-Scheines assumptions do not only imply the weaker claim that common causes are screeners-off, but—see assumptions (I) and (II)—the stronger claim that two variables will be correlated if and only if they are causally connected, that is, if and only if one causes the other or they have a common cause. And so, whenever we have two correlated variables neither of which causes the other—such as, presumably, the bread price in the UK and the sea level in Venice—our assumptions do now require that there must be a common cause.

The right response to this problem is that the correlation between bread price and sea levels isn't the *kind* of correlation that demands a causal interpretation. There are various possible ways in which we might try to delimit the appropriate kind. An initial response would be that we should only be concerned with genuinely *general* correlations, correlations between *qualitatively specifiable* types of events that might occur in any spatio-temporal locations, and not with correlations between stages of *particular spatio-temporal* processes, like the bread price *in the UK* and the sea level *in Venice*.

But this would be too restrictive. What about the correlation between the level of oxygen in the earth's atmosphere and the level of plant life on its surface? This is a correlation between stages of particular spatio-temporal processes, but we don't want to deny it causal significance on that account.

The idea we want, I think, is that of correlations between types of events which are *not* simply due to the spatio-temporal relationships between successive instances of each type on its own. The sea level and bread price correlation fails this test, since it is entirely explicable by the fact that, for each process, there is a correlation between succeeding stages. This would be shown by the standard statistical procedures for analysing "time-series". The moral, then, is that correlations between the stages of *different* time-series are not to be counted as causally significant unless they display co-variation beyond that due to co-variation within *each* time-series.

(ii) Quantum Correlations

The well-known quantum correlations between spacelike separated measurements on joint systems also lack screening-off common causes. (It's not just that they *don't* have screening-off common causes. John Bell showed that their statistical structure means they *can't* have any.) These quantum correlations pose the same problem as the last objection. But they obviously doesn't have the same solution, since the correlated measurements aren't members of time-series.

The problem, to repeat, is that the Spirtes-Glymour-Scheines assumptions imply that if A and B are correlated, then either one causes the other, or they have a screening-off common cause. However, the quantum correlations can't have a screening-off common cause. But nor do we want to conclude that one measurement result causes the other, since, apart from anything else, this seems inconsistent with special relativity.

Perhaps there is some other way, apart from appeals to time-series, by which we can argue that these quantum correlations aren't the *kind* of correlations that demand a causal interpretation. One suggestion, due to Michael Redhead (1987, 1989), is that they are insufficiently *robust* to carry causal significance. By this he means the correlation between one measurement result and another isn't insensitive to how the former is brought about. If A really were a cause of B, then we would expect the influence of A on B to remain the same whether or not A or its absence is brought about by, say, D. In symbols, Prob(B/A&D) = Prob(B/A), and Prob(B/-A&D) = Prob(B/-A). But this robustness requirement is violated by the quantum correlations. If A is the measurement result on one wing, and B the result on the other, then the probabilistic dependence of B on A will alter if we alter circumstances so as to affect the probability of A's occurrence.

Whether this suffices to deal with the problem is a matter of some current controversy. Some commentators have pointed out that the probabilistic dependence of B on *A-and-the-prior-quantum-state-P* is robust with respect to how *that* putative cause is brought about (note that anything that changes the probability of A will change P), and so, even given Redhead's suggestion, this joint circumstance will qualify as the cause of B (Cartwright and Jones 1991; Healey 1992). In response, I have heard Redhead express doubts about whether quantum states qualify as the kind of entity that can enter into causal relationships.

I propose to leave this issue unresolved. We will know better what to say about the spacelike quantum correlations when we know better how to interpret quantum mechanics. Given our lack of current understanding of quantum mechanics, it would

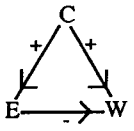
be premature, I think, to require a philosophical account of causation to deal with all quantum puzzles. (In this connection, note that the spacelike correlations pose a problem for not just for my specific approach, but for any account of causation that implies that, if Prob(B/A) > Prob(B), then either one causes the other or they have a common cause.)

(iii) Future Screeners-Off in State Space

Frank Arntzenius (1990) has observed that in any macroscopically described type of deterministic system, if we have a common cause C which screens off a correlation between later effects A and B, then there will be always be some yet later type of "event" D for which C is both necessary and sufficient—namely, the "event" consisting of the set of points in phase space which the points in C will evolve into. But if this is right, then it would seem that no probabilistic relationships can possibly identify C rather than D as the common cause of A and B, since D will bear exactly same probabilistic relationships to everything that C does.

My response is that these later "events" will not be be events in any normal sense of the word, since they will invariably consist in an entirely disconnected set of phase space points, with nothing in common except that they have all evolved from points in C. There will certainly be no everyday, macroscopic description which picks them out. I take causation to be a relationship between events of a kind that can be described in such normal terminology. So I do not regard cooked-up events like D as candidates for entering into causal relationships. (Arntzenius' example of Cleopatra and the slaves is intended to provide a kosher macroscopic later D, namely death (*op cit* pp 86-7). But I would dispute whether this example properly instantiates the formal phase space idea, and would correspondingly argue that there are probabilistic relationships which can distinguish the later death D from the earlier poisoning C. But this is not the place to pursue details.)

(iv) The Spirtes-Glymour-Scheines assumption (I) says that if two factors are causally connected, then there will be a correlation between them. But there are counter-examples. For example, imagine that drinking cola (C) both stimulates people to exercise more (E), but also causes them to put on weight (W). And suppose further that exercise E independently has a negative influence on weight increase W, to just the extent required to cancel out the direct positive influence of C, and leave us with an overall zero correlation between cola C and weight increase W. The cola and the weight increase are causally connected, but aren't correlated—so they violate (I). (In discussion, Nancy Cartwright has stressed this difficulty facing the Spirtes-Glymour-Scheines approach.)



C and W uncorrelated

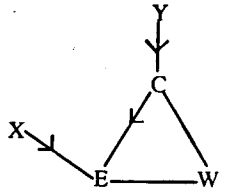
diag 5

Spirtes, Glymour and Scheines, if I understand them right, have a weaker variant assumption which is not falsified by such cases. This is the assumption of "minimality", which says only that every *direct* causal connection makes *some* probabilistic difference. This assumption is satisfied in the cola-exercise-weight example. The stronger requirement, which is violated by that example, and which they call "faithfulness", is in effect that these probabilistic differences should never cancel out in such a way as to leave us with a zero overall correlation between two variables where the causal structure would lead us to expect some correlation (*op cit* 3.4).

Despite the fact it is open to counter-examples, Spirtes, Glymour and Scheines, generally assume faithfulness rather than minimality in their analysis. This is because their interests are as much methodological as metaphysical, and any practical procedures for inferring causes from probabilities are likely to deliver definitely wrong answers in cases where faithfulness is violated. Thus, for example, in the cola-exercise-weight example, the correlations between these three variables unequivocally indicate the false verdict that cola C and weight increase W are independent causes (positive and negative respectively) of exercise E. For cola C and weight increase W are themselves probabilistically independent, yet each correlated with exercise E.

In defence of the faithfulness assumption which rules out the possibility of such misleading cases, Spirtes, Glymour and Scheines say that it would be a complete freak if there were ever in fact any perfect cancelling out. (Or, rather, they say (*op cit* 3.5.2) that, for any causal structure, the set of points in the parameter space that create misleadingly vanishing correlations have Lebesgue measure zero— but this comes to the same thing.) Given their methodological interests, this is not unreasonable. From a metaphysical point of view, however, it is less than satisfactory. If it is *possible* that there should be cancelling-out cases, as clearly seems to be the case, then the idea that causal direction *reduces* to probabilitistic asymmetries is in trouble.

Note, however, that the misleading conclusion indicated by the cola-weight-exercise correlations taken on their own could be overturned if we embedded these variables in a larger structure of variables. For the probabilistic relationships among such a larger structure of variables could well be inconsistent with the claim that weight W is a negative cause of exercise E, rather than vice versa, or that cola C exerts no causal influence on weight W. Thus, in the diagram below, a correlation between X and W would undermine the former claim, and a correlation between Y and W would undermine the latter.



diag 6

This is analogous to the earlier point made about the underdetermination of the causal diagrams 1-3 by probabilistic facts. In that case we saw how extra probabilistic facts from a wider network could resolve the *ambiguity* between diagrams 1-3. Similarly, in the present case we see how such extra probabilistic facts can overturn the definitely *misleading* verdict of the C-W-E correlations taken on their own.

Still, what if the extra probabilistic facts themselves violate faithfulness, and so continue to conceal the underlying causal facts? Couldn't the true causal structures be concealed by a kind of cosmic conspiracy, in which probabilistic correlations keep on cancelling out, however wide a network of variables we consider?

I deny that this is possible. I allow that *within* a framework of probabilistically connected variables we can sometimes find cancelling-out substructures which are causally misleading when considered on their own. But I do not accept that such failures of faithfulness can be universal. A world in which no probabilistic dependencies at all manifested some supposed causal structure would be a world in which that causal structure did not exist.

I do not necessarily want to deny that such conspiratorial worlds are *conceivable*. I am happy to accept our *idea* of causation is sufficiently detached from ideas of probabilistic dependence and independence for us to be able to *imagine* that there are causal connections which never show up probabilistically, however wide a set of variables we consider. But I deny that such imagined situations are *possible*. I think that the relevant probabilistic connections constitute a *metaphysical* reduction of the causal relationship, just as, say, molecular motion constitutes a metaphysical reduction of temperature. The probabilistic connections provide the metaphysical essence of the causal structure. So a conspiratorial world may be conceivable, but it is not metaphysically possible.

This answers one of Price's main objections to the fork asymmetry version of the "third arrow" strategy I am defending. He complains that the fork asymmetry has inadequate scope, in that there are cases of causal asymmetry where the fork asymmetry is absent (this volume, his section 3). For example, agues Price, a possible world in thermodynamic equilibrium would lack the fork asymmetry, as would any microscopic realm in the actual world which had no macroscopic features. I agree. But I say that these situations would lack causal structure too. Even if we can conceive of them as being causally ordered (which is in effect all Price assumes), it does not follow that they really would be.

I do not of course want to deny that worlds which lack the fork asymmetry, like worlds in thermodynamic equilibrium, are possible in themselves. What are not possible are worlds which both lack the fork asymmetry but contain causal asymmetry.

(These last few paragraphs have argued that the relationship between *causal asymmetry* and the *fork asymmetry* is metaphysical rather than conceptual. There is also the question of the relation between these two arrows and the arrow of *temporal asymmetry*. On this question, I would be prepared to argue that the direction of time is *conceptually* dependent on the direction of causation. So, ascending the explanatory order, the fork asymmetry first yields a metaphysical explanation for causal asymmetry, and this in turn yields a conceptual explanation for temporal asymmetry.)

3. Causation and Action

In the first instance a metaphysical reduction, as opposed to a conceptual reduction, does not require any *a priori* grounding, but only *a posteriori* evidence that the relevant probabilistic asymmetries in fact coincide with our pre-theoretical judgements of causal asymmetry. But we standardly hope for something more from a reduction than just such an empirical demonstration of coextensiveness. We also want a reduction to *explain* why the reduced phenomenon has the manifest features by which we pick it out and which make it interesting in the first place. Thus the kinetic reduction of temperature explains why hot things feel hot, why increases in temperature lead to increases in pressure, and so on.

I take it that the manifest feature by which we identify causal direction, and in virtue of which it is of such interest to us, is its *instrumentality*. We human beings can *use* directed causal relationships to manipulate things. If A causes B, and we can bring about A, we can thereby influence B—which is important to us, since there are lots of Bs we care about.

In this final section I want to show how the reduction I have proposed can explain the instrumental power of directed causal relationships. As it happens, I shall be appealing to pretty much the same probabilistic phenomena that Huw Price uses to explain the direction of causation. But my attitude to these probabilities is different. Where Price thinks that these probabilities are somehow peculiar to the agent's perspective, and that causal asymmetry is therefore a "projection" of the agent's perspective onto the world, I think that these probabilities and the resulting causal structure are entirely objective features of the world, whose relationship to human agency is only incidental. Even if human agents had never existed, there would still have been objective causal asymmetry. The connection with agency is simply that this objective probabilistic-causal asymmetry enables us human beings to influence events. Still, as I said, it is this connection with agency that makes causal asymmetry especially interesting to us. So we would like our reduction to explain it.

At first sight this might seem trivial. Our reduction hinges *inter alia* on the assumption that if A causes B then (unfaithful freaks aside) Prob(B/A) > Prob(B). So doesn't it immediately follow that, if we can control A, we can thereby influence B, at least to the extent of increasing its probability?

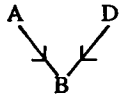
The trouble with this explanation is that it works too well. For if A causes B, then not only is Prob(B/A) > Prob(B), but also Prob(A/B) > Prob(A), since these inequalities are trivially equivalent. So the above explanation threatens also to imply that anybody who can control some effect B will thereby be able to influence its cause A, which would be absurd.

So we need some way of avoiding this overkill. We need to identify some feature that differentiates the "backwards" conditional probabilities from the "forwards" ones and explains why the latter alone are good for acting on.

I think the idea we need here is *robustness*, in Redhead's sense. It follows from the reduction that I am proposing that the "backwards" conditional probabilities, unlike the "forwards" ones, are not so robust. The idea of robustness, recall, was that genuine causal conditional probabilities should be stable under variations in the way the antecedent condition is brought about. If A causes B, then we should find that Prob(B/A) = Prob(B/A&D), and Prob(B/-A) = Prob(B/-A&D), for all the different Ds which might antecede A. That is, the causal dependence of B on A and its absence should itself be independent of where A or its absence have come from. This requirement, as we saw, was violated by the quantum correlations between spacelike separated measurements. It is also violated, I shall now show, by the "backwards" conditional probabilities which at first sight might seem to suggest that we can use effects to influence causes.

Consider the situation where the effect of some cause is also partially controllable by human decision. For example, let the cause be high atmospheric pressure, A; let the effect be a high reading on the barometer, B; and suppose that we can also decide to raise the barometer reading artificially, D, by fiddling with the dial, say, or putting

the barometer in pressure cooker. I say that we cannot increase the atmospheric pressure A by deciding D to increase the barometer reading B, despite the correlation between B and A, because this latter correlation is not robust with respect to D. It doesn't remain stable in those cases where B occurs as a result of your decision D.



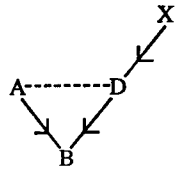
diag 7

In order to show this, let me assume for the moment the special case where D is probabilistically independent of A—the probability of someone deciding to fiddle with the barometer is independent of the atmospheric pressure. Given this assumption, we can't possibly have robustness. For robustness tells us that the B/A correlation doesn't depend on whether B or its absence comes from D—that Prob(A/B) = Prob(A/B&D), and that Prob(A/-B) = Prob(A/-B&D). But these are just the requirements that B screens D off from A. And this screening-off requirement, together with the fact that both A and D are certainly unconditionally correlated with B, would imply that D is correlated with A, contrary to my assumption at the beginning of this paragraph. (In effect robustness would mean that B stood probabilistically to A and D as common causes stand to their joint effects—and this probabilistic structure implies that such joint effects are correlated. For a proof, see Reichenbach 1956, ch 19.)

This establishes the non-robustness of "backwards correlations" for the special case where D and A are themselves uncorrelated. But we can imagine situations where this assumption is violated. Can we establish backwards non-robustness for this more general case?

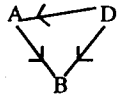
Well, if D is correlated with A, then either one causes the other or they have a common cause. Let us consider separately the situations where: (a) either A causes D or A and D have a common cause, and (b) D causes A. (Note that all cases where A is in the *past* will be of type (a).)

In case (a) we need to retreat to a broader causal framework, and find some further cause X of D which is independent of A (see diagram 8) and run the argument as before. That is, if A could be influenced backwards by B—the possibility we need to discredit—then this influence ought to be robust in the situation where B comes from X—Prob(A/B&X) = Prob(A/B) and Prob(A/-B&X) = Prob(A/-B). But this would mean that B screens X off from A, and once more this, together with the overall A/B and X/B correlations, would be inconsistent with the fact that X is independent of A.



diag 8

For case (b) we need something different. For if D, in addition to causing B, is also a genuine cause of A (diagram 9), then there won't be any causes X of D which are independent of A. But here we can appeal to a result of Spirtes, Glymour and Scheines. They show that if we start with an overall causal structure satisfying (I)-(III), then, in any sub-structure derived by conditioning on some factor D, the "forwards" conditional probability Prob(B/AD) will remain equal to the original Prob(B/A) (cf *op cit* theorems 3.6, 7.1). But it is not difficult to show that if this "forwards" conditional probability remains constant in this way, then in general the "backwards" conditional probability Prob(A/BD) will not. So in this kind of set-up we find once more that the "backwards influence" of B on A is not robust in the way that the forwards A-B influence is.



diag 9

Let me conclude this paper with three further observations on the line of thought developed in this section.

(i) This kind of argument used for case (a) is not new. It goes back to Dummett's classic paper "On Bringing About the Past" (1964). Variations on Dummett's theme are found in Mellor (1981) and Price (1992). These writers, however, use various assumptions about human freedom and knowledge to ground the assumption that D will be probabilistically independent of A (which assumption they then show is inconsistent with the robustness of a backwards influence from B to A). In outline, they note that, if A is past, then an agent can *know* whether A has happened, and they then observe that such agents can *freely* decide whether D, both when A has happened, and when it hasn't.

I have no objections to such arguments. But I think the approach adopted here yields a broader perspective. Rather than simply taking certain assumptions about human agents as basic, I prefer to view human agents as themselves part of the overall probabilistic structure of the world. From this perspective, the existence of circumstances in which D and A are probabilistically independent is a *consequence* of the overall structure, not an independent datum. And this is important, because it shows that the asymmetry of causation is not merely a projection of human experience, but a general phenomenon of which human agency is just one instance. (To show exactly how the Dummett-Mellor-Price assumptions about agency follow from the overall probabilistic structure of the world would be a more detailed enterprise than is possible here. But remember the observation in section I that our ability to remember the past is itself an upshot of the fork asymmetry.)

(ii) Nancy Cartwright has argued (in discussion again) that the failure of robustness for the "backwards" conditional probabilities does not suffice to explain why we cannot manipulate effects so as affect causes. For the failure of robustness only shows that, when we condition on D, the B/A correlation is *different* from the unconditioned correlation. It doesn't show that the correlation disappears. That is, B might still make some difference to A in the presence of D; non-robustness only requires that this be a different difference from the difference B makes generally. (We estab-

lish non-robustness by showing B *fails* to screen off D from A, that is, that B doesn't make the same difference to A when D is present; but perhaps a better way of showing no backwards causation would be to show that D *does* screen off B from A, that is, that B doesn't make *any* difference to A given D.)

There are a number of possible responses to this objection. One line is to appeal to the assumption of human freedom used in the Dummett-Mellor-Price tradition. Mellor, for example, effectively assumes a deterministic link between D and B—we *can* bring about B if we decide to. This removes the difficulty, for, given such a deterministic D/B link, the assumption that D is uncorrelated with A implies that D&B is uncorrelated with A too. However, this deterministic assumption seems a special case. In connection with the more general case, Dummett observes that, in order for B to continue to make a difference to A in the presence of D, even though A and D are themselves uncorrelated in general, there would need to be a systematic tendency for you to *fail* to bring about B by deciding D in just those cases where A has *not* happened. And this tendency might itself be thought to be in conflict with the idea that you are free to bring about B by deciding A.

My hope is to be able to explain human freedom as itself an aspect of the overall probabilistic structure of the world. So from my point of view these appeals to more detailed assumptions about human freedom would incur more detailed explanatory burdens. It may be preferable to adopt a different strategy, and seek to uphold robustness as a necessary condition of causality after all. If we can do this, then we continue to maintain that the failure of B/A robustness indeed shows that the B/A link is not causal.

Here is one possible line of argument. The link between causation and decision requires more than just that a cause makes *some* probabilistic difference to its effect. Causation, I am assuming, is an objective phenomenon. But we are interested in it because knowledge of causal connections enables us to choose means appropriate to our ends. Such decisions, however, are characteristically quantitative. We want to know *how* likely it is that E will follow C, so as to be able to compare the overall advantage expected from C with those from other courses of action. But this means that a C/E link that had different strengths in different circumstances would not qualify as a causal connection. Just knowing that C makes *some* probabilistic difference to E is unhelpful in most real-life decisions. We need to know *how much* difference it makes. So a link, like the B/A link in our example, that oscillates in strength depending on what the background to B is, could on this account be argued not to be a genuine causal connection.

(iii) My aim in this section has been to explain, in the light of the probabilistic reduction of causal direction I have proposed, why directed causal connections are good for acting on. My answer has been that it follows from this reduction that "forwards" conditional probabilities are robust, but "backwards" ones are not. However, Frank Arntzenius (1990, section 5) argues that this kind of contrast in robustness is itself the essence of causal direction, quite independently of the fact that it follows from our assumptions (I)-(III). In particular, Arntzenius would like to detach the robustness asymmetry from the the commitment to the existence of screening-off common causes contained in assumption (III).

As you might expect, given my overall argument in this paper, I am in some sympathy with Arntzenius' stance. But there remain questions about the possibility of detaching the robustness asymmetry from the principle of the common cause, and I would like to conclude by briefly mentioning some of these.

For a start, we will still need the full set of assumptions (I)-(III) to fix causal structure in general. Even if the robustness asymmetry is the specific consequence of (I)-(III) which explains why effects can't be used to influence their causes, it won't on its own suffice to fix directed causal relationships among any set of probabilistically related variables. For without the principle of the common cause, we will have to allow that C may be the common cause of an otherwise unconnected A and B, yet C not screen off the A/B correlation. But then we won't be able to tell whether such unscreened-off correlations indicate direct A/B connections, or merely that A and B are joint effects of C.

But perhaps this isn't conclusive. While we clearly won't be able to fix causal direction if we have *no* assumptions about the probabilistic structure of common causes, we may be able to manage with *different* assumptions from the standard screening-off assumptions about common causes. Thus some people want to think of prior quantum states as the common causes of the quantum correlated spacelike separated measurements. These prior quantum states don't screen off the quantum correlations. But quantum theory does specify an alternative probabilistic structure. Perhaps this specification, when conjoined with the probabilistic independence of causally unconnected variables, and the screening-off property of intermediate causes, will give us an alternative way of fixing causal order for such systems to that provided by (I)-(III).

There remains another question, it seems to me, of whether we *in fact* ever find the robustness asymmetry without the normal screening-off property of common causes embodied in (III). I am happy to accept that there is no mathematical block to such uncoupling of robustness asymmetry from screening-off common causes. While assumptions (I)-(III) imply the robustness asymmetry, I agree that the converse implication does not hold. But the fact that there is mathematical space for certain kinds of structures—those with the robustness asymmetry but without screening-off common causes—does not mean that this space is ever physically occupied. (In which case I would say once more that the real essence of causal direction requires screening-off common causes, even if it is conceivable, in the way Arntzenius has in mind, that it might not.)

Consider, by way of analogy, the link between determinism and screening-off common causes mentioned at the end of section 1. Underlying determinism plus assumptions (I) and (II) imply the screening-off feature of common causes. But the converse is not true. The screening-off feature of common causes does not require underlying determinism. So there is mathematical room for structures with screening-off common causes but where the link between causes and effects is fundamentally chancy. However, it is unclear whether such mathematical structures are ever physically actualized. When we study those quantum systems in which we are sure we have genuinely chancy links, we find that the correlations between coordinated results have a structure which precludes any possibility of screening-off common causes. Put it like this. Underlying determinism *forces* the existence of screening-off common causes. But in the absence of determinism, there is room for nature to avoid screening-off common causes—and it seems that nature takes it.

I am suggesting that screening-off common causes are in fact only found where there is underlying determinism (or determinism-to-a-high-degree-of approximation). An analogous conjecture would be that the robustness asymmetry is similarly only found where there is underlying determinism, and hence screening-off common causes. To make good this conjecture, we would need to show that underlying determinism allows a natural explanation for the robustness asymmetry (like the explanation it allows for the screening-off asymmetry) and that, in cases where this explanation is not available, the robustness asymmetry disappears (as screening-off common causes disappear in quantum situations). But that would take us beyond the issues discussed in this paper.

References

Arntzenius, F. (1990), "Physics and Common Causes", *Synthese* 82: 77-96.

Cartwright, N. (1989), *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

Cartwright, N. and Jones, M. (1991), "How to Hunt Quantum Causes", *Erkenntnis* 35: 205-231.

Dummett, M. (1964), "Bringing About the Past", *Philosophical Review* 73: 338-359.

Ehring, D. (1982), "Causal Asymmetry", *Journal of Philosophy* 79: 761-774.

Hausman, D. (1984), "Causal Priority, *Nous* 18: 143-54.

Healey, R. (1992), "Causation, Robustness and EPR", *Philosophy of Science* 59: 282-292.

Horwich, P. (1987), *Asymmetries in Time* . Cambridge, Ma: MIT Press

Lewis, D. (1979), "Counterfactual Dependence and Time's Arrow", *Nous* 13: 455-476.

Mellor, D. (1981), *Real Time* . Cambridge, UK: Cambridge University Press.

Papineau, D. (1985), "Causal Asymmetry", *British Journal for the Philosophy of Science* 36: 273-289.

Price, H. (1992), "Agency and Causal Asymmetry", *Mind* 101: 501-520.

Redhead, M. (1987), *Incompleteness, Non-Locality and Realism* . Oxford: Oxford University Press.

Redhead, M. (1989), "Nonfactorizability, Stochastic Causality, and Passion-at-a-Distance", in *Philosophical Consequences of Quantum Theory* , J. Cushing and E. McMullin (eds.). Notre Dame: University of Notre Dame Press, pp. 145-153.

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

Sober, E. (1988), "The Principle of the Common Cause" in *Probability and Causality: Essays in Honour of Wesley Salmon* J. Fetzer (ed.). Dordrecht: Reidel, pp. 211-228.

Spirtes, P., Glymour, G. and Scheines, R. (1992), *Causation, Prediction and Search*. Typescript, Department of Philosophy, Carnegie Mellon University.

The Direction of Causation: Ramsey's Ultimate Contingency

Huw Price

The University of Sydney

1. Introduction

Our present concern originates with two uncontroversial observations about causation: the causal relation is asymmetric, so that if A is a cause of B then B is not a cause of A; and effects never (or almost never) occur *before* their causes. Uncontroversial as they may be, these features of causation are far from unproblematic. A philosophical theory of causation thus has these two non-trivial tasks, among others: to explicate the difference between cause and effect—to reveal the true content of the "arrow" of causation, so to speak—and to explain why the arrow of causation is so well aligned with the arrow of time.

Note that the latter task permits two readings, depending on whether the temporal reference is read rigidly. On the stronger rigid or de re reading, the question is why the causal arrow points in *this* particular temporal direction, thought of as fixed independently of our disposition to treat the direction in question as that of the future rather than the past. On the weaker non-rigid or de dicto reading, the issue is that as to why we take the cause–effect arrow to point towards what we think of as later times; this allows that in a world in which we ourselves had the opposite temporal orientation, we might take both arrows to point in the opposite direction. (We might also consider rigid and non-rigid readings of the reference to the causal relation, which would yield four possible variants in all.)

An extreme form of a de dicto answer is that offered by Hume, who takes the two arrows to be related by definition. Thus the fundamental causal relation is held to be symmetric in time, and the claim is that we simply use the different terms "cause" and "effect" to distinguish the earlier and later members of a pair of events so related—the phrase "is a cause of" is simply shorthand for something like "is earlier than and causally related to".

It has often been noted that there is a heavy price to be paid for this convenience, however. For example it makes simultaneous causation and backward causation logical impossibilities, which seems too strong, and it precludes the project, attractive to many, of explicating temporal order in terms of causal order. (These arguments and

---

PSA 1992, Volume 2, pp. 253-267