Brit. J. Phil. Sci. **50** (1999), 521–583

Independence, Invariance and the Causal Markov Condition

Daniel M. Hausman and James Woodward[1]

ABSTRACT

This essay explains what the Causal Markov Condition says and defends the condition from the many criticisms that have been launched against it. Although we are skeptical about some of the applications of the Causal Markov Condition, we argue that it is implicit in the view that causes can be used to manipulate their effects and that it cannot be surrendered without surrendering this view of causation.

Whether or not one holds that causation is a deterministic relation, there appear to be connections between causation and probabilities. Causes are correlated with their effects. Effects of a common cause are unconditionally correlated with one another, but they are independent conditional on their common cause. Causal intermediaries screen off their effects from their causes. Events that are not related as cause and effect or as effects of common causes are uncorrelated. These claims are rough and, as just formulated, indefensible. But they point to an important connection between causation and probabilities. This paper is concerned with a promising formulation of this connection, which Peter Spirtes, Clark Glymour and Richard Scheines (hereafter SGS) call 'the

[1] This essay represents the product of more than a year of intense discussion which forced each of us not only to clarify but also to modify positions that we previously held. Some disagreements between us remain, which are registered in footnotes, but for the most part this paper presents our shared view, and virtually the whole of it was written and rewritten and rewritten still again by both of us.

Causal Markov Condition,' which was apparently first described by Kiiveri and Speed ([1982]). This condition plays a central role in recent literature on causal inference from statistical data. It is the keystone of SGS's work, and it is also central to the work of Judea Pearl and his collaborators (e.g. Pearl [1995], [1998]). Versions of this condition have also been extensively discussed in the statistical literature, although not always with an explicitly causal interpretation. In this essay, we examine what the Causal Markov Condition says, whether one should accept it, and what its limits may be. In particular we argue that there is a deep connection between the Causal Markov Condition (and hence claims concerning screening-off) and a plausible principle connecting causation and manipulation.

1 What is the Causal Markov Condition?

SGS define the Causal Markov Condition (**CM** from now on) as follows:

> Let ***G*** be a causal graph with vertex set **V** and ***P*** be a probability distribution over the vertices in **V** generated by the causal structure represented by ***G***. ***G*** and ***P*** satisfy the Causal Markov Condition if and only if for every X in **V**, X is independent of **V**\(**Descendants**(X) ∪ **Parents**(X)) given **Parents**(X).[2]

This definition obviously requires a bit of explication. A causal graph ***G*** is a set of vertices that represent the relata of the causal relation and a set of directed edges from one vertex to another. A causal graph can be used to represent token causal relations, but SGS are concerned instead to represent causal relations among types of events or among variables. For the moment we shall trust to the reader's own understanding of what it is for one type or variable to be a cause of another type or variable. To avoid cumbersome terminology and notation, we shall use the same characters to refer to types or variables and to the vertices that represent them. So **V** is both the set of vertices of the graph ***G*** and the set of variables whose causal relations are represented by ***G***. This conflation is, we believe, harmless. Following SGS, we shall use upper-case italics to represent variables and lower-case italics to represent their values.

The convention for representing causal relations is to draw a directed edge from vertex U to vertex V if and only if there is a 'direct' causal relation between U and V. A direct causal relation in this context is merely a causal relation that does not pass through any of the other vertices in the set **V**. A sequence of vertices $\{V_1, \ldots, V_n\}$ is a *path* from V_1 to V_n if and only if for all i ($i < n$), there is a directed edge from V_i to V_{i+1}. This essay considers only *acyclic* graphs—that is graphs in which no vertex appears more than once along any path. X is a *parent* of Y if and only if there is an edge from X to Y.

[2] We have made some slight changes in notation, substituting X for W.

X is an *ancestor* of *Y* if and only if there is a path from *X* to *Y*. *Y* is a *descendant* of *X* if and only if *X* is an ancestor of *Y*.

With these preliminaries, we are now ready to elucidate **CM**. *X* is a vertex (variable) in the vertex (variable) set **V**. **Descendants**(*X*) is the set of all the descendants of *X* in the graph—that is, all the effects of *X* in the set **V**. **Parents**(*X*) is the set of all the parents of *X* in the graph—that is the set of all the direct causes of *X* in the set **V**. Since **Parents**(*X*) is a subset of **V**, direct causes of *X* that are not in **V** will not be in **Parents**(*X*).[3] Independence is, of course, probabilistic independence. So ***CM** says that, conditional on the set of all its direct causes in **V**, X* is independent of all the variables in **V** that are not direct causes of *X* or effects of *X*. Since (trivially) *X* is independent of its parents conditional on its parents, one can simplify still further and formulate the Causal Markov Condition as follows:

> **CM** (*The Causal Markov Condition*): For all distinct variables *X* and *Y* in the variable set **V**, if *X* does not cause *Y*, then Pr(*X*/*Y*&**Parents**(*X*)) = Pr(*X*/**Parents**(*X*)).

CM says that conditional on its parents, *X* is independent of every variable in **V** except its effects. **CM** thus says that every variable is screened off by the set of all its parents from every other variable except its effects. Although SGS do not insist explicitly that the variables be distinct from one another, this restriction is required. When variables bear conceptual or logical connections to one another, or when their located values have parts in common,[4] then they may bear probabilistic relations to one another that have no causal explanation, and an unrestricted version of **CM** will fail. One can also think of **CM** as stating a sufficient condition for '*X* causes *Y*.' If *X* and *Y* are probabilistically dependent conditional on the set of all the direct causes of *X* in a probability distribution generated by the given causal structure among the variables in **V**, then *X* causes *Y*.

SGS call the converse of **CM**, 'the faithfulness condition.' It also plays an important role in their work and in the work of Judea Pearl, but we are not going to discuss it here. The Causal Markov and Faithfulness conditions together imply the biconditional, '*X* causes *Y* if and only if *X* and *Y* are probabilistically dependent conditional on the set of all the direct causes of *X* in a probability distribution generated by the given causal structure among the variables in **V**.' This biconditional does not provide any immediate reductive analysis of causation, since 'causes' appears on both sides of the biconditional.

[3] A direct cause of *X* that is not in **V** causes *X via* a path that does not pass through any of the variables in **V.**

[4] In our view, to hold that there are causal relations among variables is to assert modal generalizations concerning causal relations among token events. Token causal relations obtain among distinct events; that is, among distinct instantiations of properties at particular spatio-temporal locations or among spatio-temporally distinct located values of variables.

CM also has implications concerning *unconditional* probabilistic dependencies between X and other variables. If X and Y are not related as cause and effect and have no ancestors in common, **CM** implies that they are independent conditional on the empty set—i. e. unconditionally independent.[5] One can thus regard **CM** as a conjunction of two claims that it is analytically useful to separate. One concerns unconditional probabilistic dependencies and the other concerns screening off:

> **CM1** If X and Y are probabilistically dependent, then either X causes Y or Y causes X or X and Y are effects of some common cause Z in the set of variables **V**.[6]
>
> **CM2** If **Parents**(X)—that is, the subset of **V** containing all the direct causes of X in V—is non-empty, then conditional on **Parents**(X), X is probabilistically independent of every variable except its effects.[7]

CM2 implies that **Parents**(X) screens off X from its indirect causes and from other effects of **Parents**(X) that are not also effects of X. When **Parents**(X) is empty, **CM2** holds vacuously. **CM1** says in this case that X is unconditionally independent of every variable Y which is not an effect of X. **CM1** and **CM2** are both expressions of the idea that genuine probabilistic dependencies—statistical dependencies that do not arise from the unrepresentative character of particular samples—reflect causal relations. **CM1** says this about unconditional probabilistic dependencies, while **CM2** makes the analogous claim for conditional dependencies. Nonetheless, the conditional independencies asserted by **CM2** do not entail the unconditional independencies implied by **CM1**. Moreover, one might accept **CM1** but deny that causes always screen off in the manner required by **CM2**. This last position is explicitly endorsed by Lemmer ([1996]) and may be Cartwright's view as well. Lemmer holds that correlations that do not reflect direct causal connections are always due to common causes but, like Cartwright, he denies that the full set of common causes of two correlated joint effects (neither of which is a cause of the other) always screens off those effects from one another. Cartwright's arguments will be considered below in Section **9**. In Sections **7** and **8** we will present several arguments showing that in a deterministic context **CM1** plus additional premises implies **CM2**. In Section **10**, we will show how these arguments can be extended to indeterministic contexts.

[5] Consider the subgraph consisting of X and Y in the case where X and Y are not related as cause and effect or as effects of a common cause in **V**. If **CM** holds, then since **Parents**(X) and **Parents**(Y) both consist of the empty set and the set $\{X, Y\}$ is, in the sense to be explained shortly 'causally sufficient' if the original variable set **V** is, X and Y must be unconditionally independent.

[6] Reichenbach takes the biconditional consisting of **CM1** and its converse to define the notion of a 'causal connection' ([1956], p. 29). This biconditional is also one of the crucial propositions in Hausman's account ([1998]).

[7] Both **CM1** and **CM2** ignore the sort of nomic but non-causal correlations one finds in the EPR phenomena. See Section **9** below for further discussion.

While **CM1** and **CM2** are logically independent, it is hard to see on what grounds one might accept **CM2** and yet reject **CM1**. If one believed that there are unconditional probabilistic dependencies between variables that are not related as cause and effect or as effects of a common cause, then why should one believe that those dependencies should disappear when one conditions on direct causes? Conversely, if one thinks that unconditional dependencies can never arise by chance but only as a result of causal relationships, it is hard to see why one should believe that there should be residual or conditional dependencies that remain after causal structure is fully taken account of. The more natural view is that **CM1** and **CM2** stand or fall together.

We turn first to **CM1**. Whether one finds this condition plausible depends on what one takes 'correlation' or 'probabilistic dependency' to mean. It is (we believe) just coincidence that there is a correlation between whether a Republican or Democrat wins the US Presidency and whether the American or National League team wins the World Series in that year.[8] If **CM1** rules out this belief, then it is obviously unacceptable. The natural strategy for defending **CM1** from this sort of difficulty is to distinguish between sample frequencies and population probabilities and to take **CM1** to apply only to the latter. According to this strategy, the above correlation is merely a dependency in a sample drawn from some appropriate larger population. Such sample correlations can fail to reflect causal connections—indeed, the laws of probability ensure that it is overwhelmingly likely that this will sometimes happen. By contrast, assuming that there is no causal relationship between political and baseball outcomes, these variables will not be correlated in an appropriate larger population and hence there will be no violation of **CM1**.

Of course this defense of **CM1** raises a number of questions. What exactly is a 'population correlation'? When one is presented with a sample correlation, how does one identify the appropriate population from which it is drawn and how does one tell whether a corresponding correlation holds in this population? If one says that a population correlation obtains between X and Y just in case they are related as cause and effect or as effects of a common *cause*, then **CM1** turns out to be a definition of 'population correlation' rather than a

[8] One might say the same thing about Elliott Sober's famous example of the supposed correlation between water levels in Venice and grain prices in England ([1987], p. 465; [1988], p. 215), but we have our doubts that there is even an apparent correlation. What is at stake is not the co-occurrence of increases in both series, but the claim that (for example) $\Pr(p_{t+1} > p_t \,\&\, w_{t+1} > w_t) > \Pr(p_{t+1} > p_t).\Pr(w_{t+1} > w_t)$, where p is grain price and w is water level. And it is by no means obvious that this is true. If both prices and water levels are monotonically increasing, then all three probabilities will be one and the above inequality will not hold. More generally, since in most periods $p_{t+1} > p_t$ and $w_{t+1} > w_t$, most of the time when $p_{t+1} > p_t$, it will also be the case that $w_{t+1} > w_t$. But this fact does not imply that there is any *correlation* between the increases in one series and the increases in the other. We are indebted here to Forster ([1988]) and to David Papineau.

substantive claim about relations between probabilities and causation. If the notion of a population correlation is to be useful in this context it must be distinct from the notion of a sample correlation, and it must not have an explicitly causal definition. One also needs some justification for believing that there cannot be accidental probabilistic dependencies within whole populations. In what follows we will have little to say about the status of **CM1** or about these issues, which are deep questions for a theory of probability. We will assume that there is some suitable notion of a population correlation—if there isn't, the Causal Markov Condition is a non-starter—and we will ask whether, given this notion, there is any reason to accept the rest of what the Markov Condition asserts. Thus this essay will focus on the screening-off claims embodied in **CM2**.

2 Screening-off and the Causal Markov Condition

CM2 provides a promising formulation of the intuition behind the claim that common causes and causal intermediaries screen off. It is false to say simply that common causes screen off their effects or that causal intermediaries screen off their effects from their causes. A common cause of *X* and *Y* or a causal intermediary between them obviously does not screen them off when there is some independent causal relationship between them. Less obviously and more interestingly, a variable *Z* can fail to screen off *X* and *Y* even when the only causal connection between *X* and *Y* is *via Z*. Consider the causal graph shown in Figure 1a. There is a path between *X* and *Y* passing through *Z*, and the only edge out of *X* is into *Z*. *X* and *Y* have no common cause. Yet *X* and *Y* are not independent conditional on *Z*.

Why not? How is the fact that *Y* and *Z* have a common cause *W* (so that in conditioning on *Z*, one is failing to condition on all the parents of *Y*) relevant? Suppose that *X* measures the amount of acid in container 1 and *W* measures the amount of base in container 2, and suppose that a mixture of equal volumes of the acid and base is neutral. Container 1 drains into container 3. One third of the liquid in container 2 drains into container 3 and two-thirds into container 4. All the liquid in container 3 drains into container 4. *Z* is the pH of the liquid in container 3 and *Y* is the pH of the liquid in container 4. The causal set up is shown in Figure 1b. Suppose that one observes that the pH of container 3 remains at 7 (i.e. neutral) when the amount of acid in container 1 increases by 10 ml. In that case, the amount of base in container 2 flowing into container 3 must have increased by 10 ml or else *Z* would not remain unchanged. So *W*, the amount of base in container 2, must increase by 30 ml. So the liquid in container 4 will become more basic—*Y* will increase. *Z* does not screen off *X* and *Y*. The reference to the full set of direct causes in **CM2** is essential—anything less will fail to screen off (Figure 1).

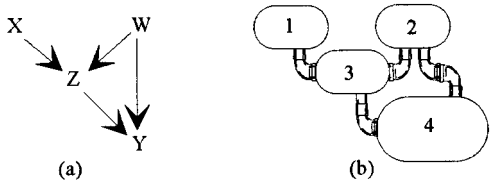

Fig. 1

CM2 remedies these difficulties with traditional formulations of claims about screening off, but it might appear to be obviously unacceptable. Suppose that *X* and *Y* are not related as cause and effect, but are effects of a common cause *Z*. Suppose further that *Z* is not in the set of variables **V** and that *Z* causes *X via* a causal pathway that does not pass through any of the variables in **V**. Then *X* and *Y* may well covary even though *X* does not cause *Y*. **CM2** appears to be false. One can rule out this difficulty by confining oneself to what SGS call 'causally sufficient' ([1993], p. 45) sets of variables—that is sets of variables that do not leave out relevant common causes. Of course, in many practical applications, a causally sufficient set of variables will not be known, and one will need a method for discovering causal relationships that does not rest on the assumption of causal sufficiency. SGS explore such methods and develop procedures for detecting latent common causes ([1993], Ch. 6). Since we are interested in the underlying rationale for **CM1** and **CM2** rather than in applications, we shall assume that **V** is causally sufficient.

Even when **V** is causally sufficient, **CM** has been subjected to a number of criticisms (see, for example, Sober [1988], Forster [1988] and Arntzenius [1993]). First, as already noted, it seems that a probabilistic dependency between *X* and *Y* might be merely accidental. Second, such a dependency may arise when one mixes two subpopulations in which *X* and *Y* are causally independent (Yule [1903]; SGS [1993], pp. 57–64). Third, dependencies that do not reflect causal connections may arise when the 'wrong' variables are measured, or the variables are not measured with sufficient accuracy or are not distinct (Yale [1903]; Salmon [1984]; Arntzenius [1993]). Fourth, there are correlations in quantum mechanics, which appear not to have any causal explanation. We shall return to the difficulties concerning quantum mechanics in Section **9** and confine ourselves here to discussing the second and third difficulties.

As an illustration of the problems created by mixing, suppose that one finds the following results from a study of the effectiveness of a new treatment:[9]

[9] This is borrowed from Glymour and Meek ([1994], p. 1012).

	Recovers	*Does not recover*
treated	240	140
untreated	260	350

Treatment and recovery are strongly correlated. Yet it could be that when one analyses the data further one finds the following:

		Recovers	*Does not recover*
women	*treated*	200	60
	untreated	100	30
men	*treated*	40	80
	untreated	160	320

Among men and among women there is no correlation between treatment and recovery. The correlation in the aggregate data results from the facts that women are more likely to recover and that a larger proportion of the women than the men in the sample are treated.[10] One might then diagnose the problem as a failure of causal sufficiency. If the relevant cause of treatment—gender—had been included in the graph, there would be no conditional probabilistic dependence between treatment and recovery. The real causal relations are as shown in Figure 2 where *G* is gender, *T* is treatment, and *R* is recovery. The probabilistic dependencies that arise from mixing populations create no new problems in principle—though they suggest that inferences from conditional probabilistic dependencies to causation are problematic in practice.

As an illustration of the difficulties that arise for the Markov Condition when variables are not accurately measured, consider a frequently cited example due to Wesley Salmon ([1984]). A cue ball collides with two other billiard balls. Call the variable that measures whether or not a collision occurs *C*, the variable that measures whether or not the first ball goes into the corner pocket *A* and the variable that measures whether or not the second ball goes in the corner pocket *B*. Each variable can take values *a*, $\sim a$; *b*, $\sim b$; *c*, $\sim c$. Because of the conservation of momentum, conditional on whether or not the collision (*c* or

[10] Such mixing has the same structure as Simpson's Paradox (Cartwright [1979], pp. 24f; Simpson [1951]). Bickel, Hammel, O'Connell, ([1975]) (cited in Cartwright [1979], p. 37) describe an example in which lower percentages of women than men are admitted to Berkeley, which makes it look as though Berkeley is practicing gender discrimination. However, department by department, the percentages of women admitted is just the same—it's just that women apply to departments that admit lower percentages of applicants.

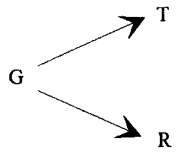

Fig. 2

$\sim c$) occurs, whether or not the first ball goes into a corner pocket (a, $\sim a$) provides additional information about whether the second ball does (b, $\sim b$). Hence although C is a common cause of A and B and A and B are not related as cause and effect; C fails to screen off A from B. The obvious response that defenders of the Markov Condition can make to this example is that a more precise and informative specification of the collision event (call it C^*)—'the exact momentum of the cue ball on striking the two target balls' (SGS, p. 63)—will be a screening-off common cause.

This response strikes us as persuasive, but notice that it is not a defense of **CM**. **CM** *fails* with respect to the set of variables $\{A, B, C\}$. What this defense alleges is that **CM** holds for some *other* set of variables characterizing the same causal relations. The claim that is defended in the response is that if two variables are correlated and neither is a cause of the other, there must exist some set of screening-off common causes. As the above example illustrates, this is very different from saying that even the full set of all variables that people ordinarily describe as common causes will screen off their joint effects or that it will be possible to specify a set of screening-off common causes in terms of any particular framework or vocabulary for dividing up the world into alternative candidates for common causes. In the case under discussion, there will be no screening-off common cause if one is confined to coarse-grained variables such as 'collision' or 'no collision.' It is only at a more refined level of description that one will be able find a screener off. Similar limitations will apply in many other real-life attempts to apply the Causal Markov Condition. For example, it is a plausible guess that many of the variables that investigators are able to conceptualize and measure in the social and behavioral sciences are too crude in the way that the 'collision, non-collision' variable is. The Markov Condition may accordingly fail for these variables even if there is some other more informative set of variables for which it holds. What this shows is that what is defensible is really the following principle:

> **CM′** (*The Causal Markov Condition*): For any system of acylic causal relations, there exists some set of variables **V** such that for all distinct variables X and Y in **V**, if X does not cause Y, then $\Pr(X/Y\&\mathbf{Parents}(X)) = \Pr(X/\mathbf{Parents}(X))$.

Although it is necessary to distinguish **CM** from **CM′** in order to respond to difficulties like those under discussion, during the rest of the essay, we shall

assume that **V** includes the 'right' variables and stick to the simpler formulation (**CM**).

A similar analysis applies to another class of putative counterexamples to the Markov Condition involving what Arntzenius ([1993]) calls equilibrium correlations. Consider a collection of gases in different initial states which are not in thermodynamic equilibrium and suppose that each is allowed to reach an equilibrium temperature and pressure. Then, for each gas, there will be a correlation between the temperature or pressure in one region of the gas and the temperature and pressure in any other region. Nonetheless, the pressure or temperature of any one region does not cause the temperature and pressure in any other region. (For one thing, there is no basis for singling out one region as the cause and the other as the effect rather than conversely.) There is also no obvious candidate for a macroscopic common cause that explains these correlations. However, this is again no counterexample to the Markov Condition, if one is allowed to advert to a different level of description. There is a complicated causal story, involving collisions and transfers of momentum between huge numbers of individual molecules, that accounts for these correlations and that, because it is deterministic, must conform to the Markov Condition. (The Markov Condition holds trivially for deterministic systems for all variables X such that **Parents**(X) is the full set of determining causes of X, since $\Pr(X/Y \,\&\, \textbf{Parents}(X)) = \Pr(X/\textbf{Parents}(X)) = 1$.) The moral is again that the level of analysis at which the Markov Condition holds may be very different from the level one originally had in mind.

Consider next a different sort of case, also due to Arntzenius. A particle moves around in a box. There is a perfect (anti)correlation between whether it has the property A of being in the right-hand side of the box and the property B of being in the left-hand side of the box, but this correlation is not due either to a direct causal connection between A and B or to a common cause. Again, however, we do not see this as a counterexample to the Markov Condition. Because A and B are not distinct, this is not the sort of correlation that demands a causal explanation. The connection between A and B is logical or conceptual rather than causal. The Causal Markov Condition is only concerned with relations among *distinct* variables. Sections **6** and **9** will return to this issue at greater length and will attempt to make more precise what it is for two variables to be distinct.[11]

Finally, consider a theorem due to Elliot Sober ([1988]) which shows that, in

[11] Many of the other purported counterexamples to the Markov Condition discussed by Arntzenius also involve variables that are conceptually or logically connected. For example, Arntzenius considers a case involving a system that can be in any one of four states S1–S4, and notes that there need be no common cause explanation for the correlation the system exhibits between the property of being in state S2 or S3 and the property of being in state S2 or S4. Again, we would deal with this case by denying that such correlations fall within the scope of the Markov Condition.

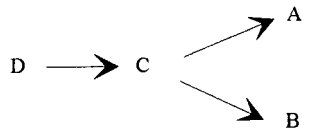

Fig. 3

indeterministic contexts in which all the probabilities involved are strictly between zero and one, if a direct (or, as Sober says, 'proximate') common cause screens off its causes from its effects and its effects from one another, then an indirect or 'distal' common cause of the those joint effects will not screen them off from each other. Sober shows that, in Figure 3, if *C* screens off *A* and *B*, then *D* does not. Again this is no objection to the Markov Condition which claims only that direct causes screen off (and Sober never suggests otherwise). It does, however, reinforce the point that one needs to apply the condition to the right variables. What looks like a failure of the condition may simply reflect the fact that one has conditioned on a distal cause such as *D* rather than a proximate cause like *C*.

These problems with the Markov Condition illustrate an important point to which we will return. This point is that one needs to know a great deal before one can justifiably assume that the Markov Condition is satisfied for some system. One needs the right variables or the right level of analysis—variables that are sufficiently informative and that are not conceptually connected. As we shall see, one also needs to know how to segregate the system correctly into distinct causal mechanisms. Apparent failures of the Markov Condition typically indicate limitations in background knowledge—that one is employing variables at the wrong level, or that one is failing to include relevant variables, or that one is treating variables or mechanisms as distinct when in fact they are not.

3 Factorizability

The Causal Markov Condition in the form stated by SGS is provably equivalent, in the case of directed acyclic graphs, to the following factorizability condition:

> Let ***G*** be a causal graph with vertex (variable) set **V** and *f* be a probability density function over the variables in **V** generated by the causal structure represented by ***G***. Then *f* is factorizable, that is
> $f(\mathbf{V}) = \Pi_{X \in \mathbf{V}} f(X/\mathbf{Parents}(X))$.

Π is the product. This is a paraphrase, not a quotation from SGS (see [1993], p. 33). **Parents**(*X*) is, as before, the set of all the direct causes of *X* in **V**, and in the case in which a variable *X* has no parents, the associated term in the

factorization is the unconditional density f(X). (Thus the factors include the marginal distributions of all the exogenous variables.) Factorizability states that the joint density of a set of variables is equal to the product of the density of each variable conditional on the set of all its direct causes. Conditional on all their direct causes, variables are all probabilistically independent of one another. In the case of discrete variables, factorizability says that $\Pr(X_1, \ldots X_n) = \Pr(X_1/\text{Parents}(X_1)) \ldots \Pr(X_n/\text{Parents}(X_n))$.

Here is a sketch of a proof of the equivalence of factorizability and the Causal Markov Condition. Define '**nd+***X*'—the non-descendants of X—to be the subset of variables in **V** that are not effects of (descendants of) X, and let '**nd–***X*' be the subset of **V** consisting of everything other than X and its descendants. **nd–***X* is, of course, **nd+***X*\{*X*}. Obviously, **Parents**(*X*) is a subset of **nd–***X* and **nd+***X*. If factorizability holds for the whole graph and the complete set of variables, it will also hold for the set **nd+***X* (for the subgraph formed by deleting all the effects of X) and for the set **nd–***X* (the subgraph formed by deleting X and all its effects). Factorizability implies,

1 $\Pr(\textbf{nd+}X) = \Pr(X/\textbf{Parents}(X)).\Pr(\textbf{nd–}X)$.

By the definition of conditional probability,

2 $\Pr([\textbf{nd+}X\backslash\textbf{Parents}(X)]/\textbf{Parents}(X)) = \Pr(\textbf{nd+}X)/\Pr(\textbf{Parents}(X))$.

nd+*X***Parents**(*X*) consists, of course, of the members of **nd+***X* that are not parents of X. (1) and (2) imply

3 $\Pr([\textbf{nd+}X\backslash\textbf{Parents}(X)]/\textbf{Parents}(X)) = \Pr(X/\textbf{Parents}(X)).\Pr(\textbf{nd–}X)/\Pr(\textbf{Parents}(X))$

and by the definition of conditional probability, one can derive

4 $\Pr([\textbf{nd+}X\backslash\textbf{Parents}(X)]/\textbf{Parents}(X)) = \Pr(X/\textbf{Parents}(X)).\Pr([\textbf{nd-}X\backslash\textbf{Parents}(X)]/\textbf{Parents}(X))$

4 says that X is independent of everything that is not a parent or a descendant conditional on the set of all of X's parents, which is SGS's formulation of the Causal Markov Condition, and which is, as explained in Section **1**, equivalent to **CM**. So factorizability implies **CM**.

To derive factorizability from **CM**, suppose **V** contains n variables $X_1, \ldots, X_n$. Let $\mathbf{X}_k$ be the set of all variables with indices k, and renumber the variables, so that if X_i is a direct cause of X_j, then $i < j$. Since the graph is acyclic, this can always be done. By the 'chain rule', which follows from the definition of conditional probability, one can derive:

(1) $\Pr(X_1 \ldots X_n) = \Pr(X_n/\mathbf{X}_{n-1}) \,.\, \Pr(X_{n-1}/\mathbf{X}_{n-2}) \ldots \Pr(X_1)$.

By construction, for all j **Parents**(X_j) is a subset of $\mathbf{X}_{j-1}$. Since no member of

$\mathbf{X}_{j-1}$ is an effect of X_j, **CM** implies

(2) for all j, $\Pr(X_j / \mathbf{X}_{j-1}) = \Pr(X_j / \mathbf{Parents}(X_j))$.

(1) and (2) then imply

(3) $\Pr(X_1 \ldots X_n) = \Pr(X_n/\mathbf{Parents}(X_{n-1})) \cdot \Pr(X_{n-1}/\mathbf{Parents}(X_{n-2})) \ldots \Pr(X_1/\mathbf{Parents}(X_1))$,

where $\mathbf{Parents}(X_1)$ and possibly some of the other sets of parents is empty. So CM implies factorization (3).

Although well known in the literature, the equivalence between **CM** and factorizability is remarkable. **CM** states that X will be independent of everything except its *effects* conditional on $\mathbf{Parents}(X)$. Factorizability, in contrast, says nothing about the effects of X. Instead it says that the joint distribution can be written as a product whose factors consist of each variable conditional on its direct causes. Yet these two claims turn out to be logically equivalent! In addition, this equivalence permits further arguments in defense of **CM**, since it may be possible to argue directly for factorizability.

4 Manipulability and causation

One crucial fact about causation, which is deeply embedded in both ordinary thinking and in methods of scientific inquiry, is that causes are as it were levers that can be used to manipulate their effects. If X causes Y, one can wiggle Y by wiggling X, while when one wiggles Y, X remains unchanged. If X and Y are related only as effects of a common cause C, then neither changes when one intervenes and sets the value of the other but both can be changed by manipulating C. For most scientists, the crucial difference between the claim that X and Y are correlated and the claim that X causes Y is that the causal claim, unlike the claim about correlation, tells one what would happen if one intervened and changed the value of X. It is this feature of causal knowledge that is so important to action.

This connection between causation and manipulability has been defended by a number of writers. For example, the economist Guy Orcutt writes, '[. . .] we see that the statement that [. . .] z_1 is a cause of z_2, is just a convenient way of saying that if you pick an action which controls z_1, you will also have an action which controls z_2' ([1952], p. 307). Similarly, Clark Glymour writes, 'One influential view about causation goes something like this: If in a system S with a set of variables $\mathbf{V}$, variable X is a direct cause of variable Y then an ideal intervention that changes X sufficiently (i) changes no variable that is not an effect of X, and (ii) changes Y' ([1997], p. 224; notation changed slightly). A number of philosophers, including R. G. Collingwood ([1940]), Douglas

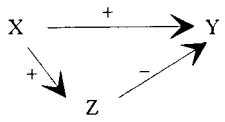

Fig. 4

Gasking ([1955]), G. H. von Wright ([1971]), and, most recently, Peter Menzies and Huw Price ([1993]; see also Price [1991], [1992], [1993]) have defended a more ambitious thesis. They have held, not just that there is a connection between causation and manipulation of the sort described above but that one can appeal to this connection to provide a reductive analysis of causation.

As reductive analyses of causation, manipulability accounts are, in our view, non-starters. First, although it is true that if interventions that change the value of *X* sufficiently change the value of *Y*, then *X* causes *Y*, the converse is not true, and so there is no analysis. Consider a causal arrangement like that shown in Figure 4.

The direct positive causal influence of *X* on *Y* is exactly canceled out by the indirect negative causal influence via *Z*. So even though *X* causes *Y*, changing the value of *X* sufficiently does not change the value of *Y*. Although, as SGS argue, one should expect such violations of 'faithfulness' to be rare ([1993], pp. 68–9), they are a serious difficulty for attempts to provide reductive analyses of causation in terms of manipulability.

Even if this problem could be solved, the biconditional, '*X* causes *Y* if and only if interventions that change the value of *X* sufficiently change the value of *Y*,' has no claim to be an analysis of causation, because its right-hand side is chock full of causal notions. Not only does 'change' here mean 'cause to be different,' but, as we shall see below, the notion of an intervention carries with it other sorts of causal commitments. Those who have tried to extract a reductive analysis of causation from the links between causation and manipulability have tried to evade these fatal difficulties by emphasizing the connection between causation and human action and by arguing that the notion of human action is prior to and independent of the notion of causation. Indeed Menzies and Price's call their account an 'agency theory' of causation. Their hope (which can be found in the work of Collingwood, Gasking, and von Wright as well) is that the notion of an intervention can be understood through its links to human agency, and that a non-circular (though highly anthropomorphic) analysis of causation can then be constructed on its basis. Menzies and Price, for example, maintain that *X* causes *Y* whenever agents could use *X* as an effective means to bring about *Y* ([1993], p. 187). In their view, *X* is an 'effective means' to bring about *Y* and only if the 'agent probability' of *Y* given

X is greater than the agent probability of *Y* given not-*X*. The agent probability of *Y* given *X* is the probability that *Y* obtains given that *X* is brought about '*ab initio*, as a free act of the agent concerned' ([1993], p. 190).

Although we place the greatest emphasis on the relationship between causation and manipulability—on the fact that in principle effects can be controlled through their causes—we wish to distance ourselves completely from agency or manipulability theories when they are understood in this way. We maintain that the notion of an intervention required to explicate the connection between causation and manipulation must be defined in explicitly causal terms and that for this reason the links between causation and manipulability could not possibly provide a reductive analysis of causation. In addition, we deny that the relevant notion of an intervention needs to be understood anthropomorphically, and we deny that there is anything anthropomorphic about the claim that *X* causes *Y*. For practical purposes, causal knowledge is especially important for human action, but the importance of a fact for human action does not show that there is anything anthropomorphic about the fact itself. For the purposes of finding out whether *X* causes *Y*, specifically human interventions are of special importance, because there are often good reasons to believe that actions that set the value of *X* bear no causal relations to *Y* except in virtue of influencing the value of *X*. But the relevant notion of an intervention that sets the value of *X* is (roughly) just that of something that is a direct cause of *X* and that bears no causal relations to the other variables under consideration except those that arise from its directly causing *X*. What this means is that the intervention *I* is not an effect of any variable in **V,** *I* does not cause any variable in **V** by a path that does not go through *X* first, and *I* is not caused by any variable that causes any other variable in **V** by a path that does not go through *I* and *X* first. Obviously this characterization makes no reference to human agency and allows a natural causal process in which human activity plays no role to count as an intervention as long as it has the right causal characteristics. So-called 'natural experiments' illustrate this possibility.

The connection between causation and manipulation we are going to rely on is a refinement of the claim that if one were to intervene to set the value of *X* within some range of values of *X* and the value of *Y* were to change, then *X* causes *Y*. In other words, if *X* does not cause *Y*, then the value of *Y* would not change if a possible intervention were to change the value of *X*.

The notion of an intervention is crucial to our formulation of the connection between causation and manipulation. In addition to causing a change in the variable *X*, an intervention on *X* must satisfy other conditions. Since changes in another variable *Y*, given an intervention with respect to *X*, are supposed to be produced through the change in *X*, a purported intervention that changes *X* must not directly change the value of *Y*. In addition, the process that changes the value of *X* must not also cause a change in other causes of *Y* and the change

in the value of *X* must not be correlated with such changes. If these conditions are not met, a change in the value of *X* could be accompanied by a change in the value of *Y* even though *X* does not cause *Y*. The rough idea of an intervention with respect to *X* is that of an exogenous change in the value of *X* that changes *Y*, if at all, only through *X* and not *via* some other causal route. Both of us have given more precise and slightly different versions of this intuitive idea elsewhere (Hausman [1998], sections 5.3* and 7.1*; Woodward [1997], p. S30), as have a number of other writers including SGS ([1993], sections 3.7.2 and 7.5) and Cartwright and Jones ([1991]). We shall not rehearse the details of these characterizations here, because nothing in the subsequent argument depends on them. The crucial point is that an intervention with respect to *X* is a direct cause of *X* that has no causal relations to any of the variables in **V** except in virtue of being a direct cause of *X*. Notice that it follows that an intervention with respect to *X* cannot also be an intervention with respect to any other variable.

To illustrate these ideas, consider a system in which two variables, *B* (e.g. the position of a barometer dial) and *S* (e.g. the occurrence or non-occurrence of a storm) are joint effects of a common cause *A* (e.g. atmospheric pressure), and in which neither *S* nor *B* is a cause of the other. As shown in Figure 5, an intervention on *B* with respect to this set of variables would consist of a process *Q* which changes the value of *B* but which is neither cause nor effect of *A* or *S* nor of any cause of *A* or *S*. (**CM1** implies that *Q* will be uncorrelated with *A* and *S*, their causes, and with any causes of *B* by a path that does not go through *Q*.) *Q* might be the action of a mechanical device that randomly sets the barometer dial at different positions. Our principle connecting intervention and causation maintains that if (as we have assumed) *B* doesn't cause *S*, then intervening to change the value of *B* in this way will not change the value of *S*. That is, *B* and *S* should be probabilistically independent under such interventions on *B*. Of course there are *other* ways of changing the value of *B* that will change the value of *S*. For example, processes that change the value of *B* by changing the value of *A* will also change the value of *S*. However, no such process counts as an intervention. It is the behavior of *S* under interventions on *B* and not the behavior of *S* under other sorts of changes in *B* that is crucial for the question of whether *B* causes *S*.

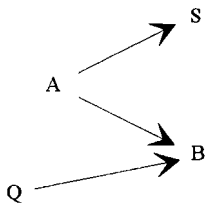

Fig. 5

We also emphasize that we are not claiming that causal claims hold only when interventions are *feasible* (or, to reiterate a point made above, only when they are carried out by human beings). Whether or not interventions that set the value of *X* are feasible and whether or not they have in fact taken place is irrelevant. The connection between causation and manipulability is counterfactual. If *X* does not cause *Y*, then *Y* would remain unchanged if any intervention were to occur that changed the value of *X* (within the relevant range of values of *X*).[12]

Note that the principle connecting causation and manipulation says that a sufficient condition for *X* to cause *Y* is that interventions on *X within a certain range* are associated with changes in *Y*. The italicized phrase is added, because even though *X* causes *Y*, interventions that change the value of *X* 'too much' may cause the relationship between *X* and *Y* to break down. For example, interventions that increase the extension *X* of a spring within some range of values of *X* may increase the restoring force *F* that it exerts, even though arbitrarily large extensions may have no effects or negative effects on *F*.

The connection between causation and manipulation we have pointed to apparently leaves open the possibility that *X* causes *Y* and that *Y* causes *X* in the circumstances in which an intervention on either variable leads to a change in the other. This can happen at the type-level, even when token causation is asymmetric. For example, an increase in the pressure of a gas can cause an increase in its temperature, and an increase in the temperature of a gas can cause an increase in its pressure. To simplify these issues and to keep this essay to manageable proportions, we shall assume that causal relations are always asymmetric. This is not as restrictive as it may appear. In the example just given, for example, one can restore asymmetry to the type-level relations between temperature and pressure by introducing time references.

[12] There are some delicate issues, which we lack the space to explore here, about the sense in which interventions must be *possible* if counterfactual claims about what would happen under those interventions are to have truth-values. For example, it seems plausible that there are real-life cases in which *X* causes *Y*, but in which there is no physically possible process that will satisfy the conditions for an intervention on *X*. All the physically possible processes that cause *X* might be, in Elliott Sober's words (private correspondence), too 'ham-fisted' to count as interventions. Consider the claim that the gravitational force exerted by one massive body causes some effect (such as tides) on a second. It is plausible that any physically possible process that would change the force exerted by the first on the second (by changing the position of the first) would require a change in the position of some third body and this would in turn exert a direct effect on the second, in violation of one of the conditions for an intervention. In this particular case, there is nonetheless an obvious basis for counterfactuals about what would happen under interventions that are minimal or surgical in the sense that they change only the position of the first body. Newtonian gravitational theory itself predicts what would be true under such interventions. That is, even if the interventions in question are not physically possible, Newtonian theory provides the resources for representing them and predicting their consequences. In other cases involving 'impossible' interventions, this will not be true. For example, as we argue below, quantum mechanics does not permit one to represent an intervention on one of the measurement results in the EPR experiment or allow calculations about what would happen under such an intervention.

This description of the connection between causation and manipulation is rough. To exhibit the connection between manipulability and the Causal Markov Condition requires a more precise formulation. Indeed two are needed: one for deterministic relations and one for probabilistic relations. The next sections will pursue the link between manipulability and the Causal Markov Condition given deterministic relations. Section **10** will pursue the link between manipulability and the Causal Markov Condition when the relations between cause and effect are indeterministic.

5 Level invariance and modularity in linear equation systems

To formulate the relations between causation and manipulability more precisely and to explore the relations between manipulability and the Causal Markov Condition (especially **CM2**), we shall begin by considering systems of deterministic linear equations and then extend the results to other contexts. We will maintain that the relation between causation and manipulability should be formulated as the claim that when such equations have a causal interpretation, they must satisfy certain requirements concerning invariance under interventions and independence of mechanisms. Consider first a regression equation of form

(1) $Y = aX + U$

where Y is some dependent variable of interest (e.g. the height of various individual plants in some population of plants), X the independent variable (e.g. amount of water each individual plant receives) and U is here an error term representing omitted causes. What conditions must be satisfied if (1) is to have its natural causal interpretation—i.e. for it to be a correct description of a quantitative causal relationship between X and Y? A natural thought is this: (1) is a correct description of the causal relationship between X and Y if and only if were one to intervene in the right way to change the value of X (if one were to wiggle X appropriately), then Y should change in the way indicated by (1)—i.e. the change in the value of Y in response to a change in x of magnitude dx should be given by adx. To employ terminology used by Woodward ([1997], [forthcoming a], [forthcoming b]), the functional relation (1) expresses should be *invariant*—it should remain stable or unchanged under such changes in the value of X. When this functional relation is invariant, then Y should change in the way indicated by (1)—i.e. the change in the value of Y in response to a change in x of magnitude dx should be given by adx. If, on the contrary, Y doesn't change in this way as a result of an intervention that changes the value of X—if the result of the intervention is that the relation between X and Y breaks down, then (1) will not be a correct description of the causal relationship

between *X* and *Y*. When a relationship such as (1) is invariant in this way, we will say that it is 'level invariant'—it is invariant under interventions that change the level (value) of the independent variable(s). Level invariance captures an intuitive idea about causation that is central to manipulability theories—that, whatever else may be true of causes, they should at least be potential handles or means for manipulating effects.

Although this idea is intuitively appealing, it needs to be stated more carefully. First, just as with the formulation of the relationship between causation and manipulation in Section **4**, the claim that a relationship is invariant under interventions is to be understood as a counterfactual claim about what would happen to the relationship if interventions (in the sense sketched above) were to be carried out. We reiterate that we are not claiming that the interventions in question need to be feasible. Our claim is rather that when (1) correctly describes a causal relationship between *X* and *Y*, then if one were to intervene to change *X* appropriately, the relationship (1) between *X* and *Y* would remain invariant, and the value of *Y* would change in accordance with it when interventions change the value of *X*.

Whenever an equation such as (1) is true of a set of variables, those variables will be regularly associated in a way that accords with (1). But the claim that (1) is level invariant says something stronger: Level invariance says that (1) would continue to hold for some values of *X* when those values are set by interventions. For example, the equation (1) might hold because the values of *X* and *Y* are always determined by a common cause and never by an intervention. In that case, (1) would break down if an intervention were to occur on *X*. Another way to see what level invariance adds to the simple claim that an equation holds true is to consider equations with more than one independent (right-hand side) variable. Nothing in an assertion of the truth of an equation implies that if an intervention were to change the value of one of the independent variables, then the values of the others would remain unchanged.

Another caveat reflects a point already noted above: Even if there is a causal relationship between *X* and *Y*, (1) will correctly describe how *Y* will change in response to interventions on *X* only for some range of values of *X*, and not for all values. For example, even if, as (1) implies, water (*X*) causes plant height (*Y*), it is not true that putting arbitrarily large amounts of water on the plants in P will cause them to grow arbitrarily tall. It is likely that (1) correctly captures the causal facts only for a relatively small range of changes in the values of *X*, and it will break down outside this range. Claims about invariance or concerning how one variable responds to interventions that change the value of another variable thus must be relativized to a range of values of variables that are set by interventions: Causal relationships such as (1) are invariant—they correctly describe how the value of a dependent variable changes in response to

interventions on an independent variable—only for a limited range of the values of variables set by interventions.[13]

Let us now consider a slightly more complicated causal structure, represented by two equations:

(1) $Y = aX + U$

(2) $Z = bX + cY + V$

where U and V are, as before, error terms. We interpret each equation as saying that every variable on its right-hand side is a direct cause of the variable on its left-hand side. Thus (1) says that X is a direct cause of Y and (2) says that X and Y are direct causes of Z. Z might for example be the weight of peas harvested from a plant in a given population of pea plants, which depends directly both on the height of the pea plant Y and on rainfall X. We can associate (1) and (2) with the graphical structure shown in Figure 6a. Notice that causal sufficiency plus **CM1** imply that the error terms are uncorrelated both with the other independent variables in the equations in which they occur and with each other.

Consider now the following equation (3), formed by substituting the right-hand side of (1) for Y in (2)

(3) $Z = dX + W$ (where $d = b + ca$ and $W = cU + V$)

This is associated with the graphical structure shown in Figure 6b.

Obviously (3) has exactly the same solutions in X and Z as (1) and (2). None the less, (3) is associated with a different graphical structure than (1)–(2), and by the rules given above represents a different system of causal relationships—(3) says that X is a direct cause of Z but omits all mention of Y, while (1) says that X is a direct cause of Y, and (2) says that X and Y are direct causes of Z. From the perspective of (1)–(2), (3) collapses two different mechanisms by

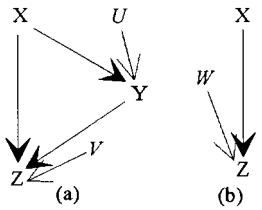

Fig. 6

[13] In speaking of a 'limited range of interventions' we mean 'interventions limited to those that set the values of variables within a specified range of values.' Although we speak here of a range of values of a particular variable for which an equation is invariant, the range of interventions on one variable for which an equation is invariant may depend on the values of other variables in the equation. The precise formulation of the relativization is thus rather complicated. Readers should understand our talk of invariance with respect to a 'limited range of interventions' as a short-hand for the precise formulation.

which X affects Z into a single mechanism. One way of bringing out this difference is again to appeal to the notion of a hypothetical intervention.

Suppose it were possible to intervene so as to change the value of Y. Since Y does not cause X, a genuine intervention that sets the value of Y will be probabilistically independent of X. (We might imagine, for example, that Y is changed by some randomizing device that is causally independent of X.) One can think of such an intervention as changing or replacing (1) with another equation (1*):

(1*) $Y = k$.

This implies that the value of Y is no longer caused by X but is set exogenously to some value k (by the intervention).

Implicit in the idea that the system (1)–(2) correctly captures the causal relations is the assumption that when (1) is disrupted by an intervention that sets the value of Y, (2) will remain invariant (provided, of course, that the intervention is truly an intervention and that the change in Y is within the relevant range of values). That is, one assumes that if the value of Y is changed by an intervention from its original value $Y = y$ by amount gy (so that (1) no longer holds and the new value of Y is now equal to $y + gy = k$, as indicated by (1*)), (2) will continue to hold, and z will change by cgy. Thus the equation system consisting of (1) and (2) differs from equation (3) in what it claims about what will or would happen under an intervention or exogenous change in Y. (1)–(2) makes a determinate prediction about what will happen in this case, while (3) does not. In this way we can capture the idea that, unlike (3), (1) and (2) describe two distinct mechanisms that can be separately disrupted.

To say that an equation such as (2) is level invariant with respect to interventions that set x is to say that when one intervenes and sets x and measures y and z, then x, y, and z satisfy equation (2). It is *not* to say that one can correctly calculate the value of Z by plugging in the value of X that is set by intervention and holding fixed the values of the other variables. This procedure will yield the correct value for Z only in the special case in which there are no causal relations among the right-hand side variables in the equation. Because, as (1) tells us, X causes Y, an intervention that changes the value of X by dx, will—if (1) is level invariant—change the value of Y by adx and hence—if (2) is level invariant—will change the value of Z by $bdx + cadx$, rather than by bdx. In other words, what level invariance requires in the case of an equation like (2) is simply that the values of the variables figuring in (2) should conform to (2) once those variables have been allowed to assume whatever values they take as a result of the intervention, where this is understood to include whatever changes may occur in those variables that are causally downstream of the variable intervened on.

This point can be made clearer by pointing out the relations between the idea

that different equations in a system of equations should correspond to distinct mechanisms that can be changed independently of each other and another idea, due to Pearl, Glymour and their associates, about how to represent interventions graphically: According to these writers, an intervention on a variable such as *Y* that sets its value breaks all of the edges in the graph directed into *Y* (i.e. all of the arrows whose heads are pointed into *Y*) and preserves all other edges in the graph, including all edges directed out of *Y* (*cf.* SGS [1993], Pearl [1995]). The breaking of the arrows represents the idea that the mechanism that determined the value of *Y* has been disrupted and the value of *Y* is now set by the intervention, rather than by the variables that were previously causes of *Y*, such as *X*. Interventions that set the value of a single variable thus disrupt only the single mechanism that previously determined the value of that variable. The preservation of other arrows in the graph, including those directed out of *Y*, corresponds to the idea that an intervention that only sets the value of *Y* leaves the other mechanisms alone—it does not alter the other causal relations in the graph, including the relation between *Y* and *Z* indicated by (2). The relationship represented by the arrow between *Y* and *Z* remains invariant under such an intervention, just as the relations between *X* and *Y* and between *X*, *Y* and *Z* remain invariant to interventions that set the value of *X*. Since each set of arrows directed into a variable corresponds to a distinct equation, this 'arrow-breaking' interpretation of interventions is an alternative way of expressing the independent changeability of equations described above. Saying that you can break the arrows directed into a variable while leaving the arrows directed into other variables unbroken is just another way of saying that you can disrupt the equation corresponding to the first set of arrows while leaving the other equations in the system undisturbed.[14]

The arrow-breaking interpretation of interventions implies both what we called above the 'level invariance' of each of the individual equations and a stronger condition that we shall call 'modularity.' Level invariance says that changing the level of the independent variable(s) in an individual equation (within some range of values) should not disrupt the equation. So if one changes the value of a variable *X via* an intervention within a specified range, whether this involves breaking arrows—as it will unless *X* is exogenous—or not, the arrows into the endogenous variables that depend on *X* will be unaffected. Modularity involves a stronger invariance condition that also applies between equations. It says that each structural equation in a system of structural equations that correctly captures the causal relations among a set

[14] One can also conceive of an intervention, as the addition of an intervention variable into the graph with an arrow into the variable that is manipulated that leaves all the arrows in the graph intact. An intervention might, for example, merely add to the value of a variable without canceling the effect of its causes. See Hausman ([1998], Ch. 5). The logic of the situation remains the same as above, however. For this conception to work, one still needs a condition like modularity to the effect that drawing a new arrow into a variable doesn't change other arrows.

of variables is invariant under interventions that disrupt other equations in the system by setting the values of their dependent variables within some limited range. Thus if the system (1)–(2) is modular, and one intervenes to set the value of the dependent variable *Y* in (1) or—what comes to the same thing—to replace (1) with (1*), (2) will continue to hold.

As a further illustration of what modularity means, suppose that we re-write the system (1)–(2) once again as:

(1) $Y = aX + U$

(2*) $Z = eY + R$, where $e = (b/a + c)$ and $R = V - (b/a)U$.

If (1)–(2) is modular, then (1)–(2*) will not be. As noted above when one disrupts (1) by intervening and setting the value of *Y*, *Y* becomes independent of *X*—it is no longer a function of *X*. One can thus think of this intervention as changing the coefficient *a* in (1) to zero and replacing (1) with an equation like (1*). Under such an intervention (2*) will not be invariant, because the coefficient *e* in (2*) is a function of *a* and changing the value of *a* will change *e*.

A numerical illustration may clarify matters. Suppose that $a = 2$, $b = 4$, and $c = 3$, and ignore the error terms. Then (1)–(2) say that

(1e) $Y = 2X$

(2e) $Z = 4X + 3Y$,

and (2*) says that

(2e*) $Z = 5Y$.

So if, for example, $X = 1$, (1e)–(2e) says that $Y = 2$ and $Z = 10$; and if $Y = 2$, then (2e*) says of course that $Z = 10$. Suppose now that one intervenes and sets the value of *Y* equal to 3. Then (2e) says that $Z = 4 + 9 = 13$, while (2e*) says that $Z = 15$. Obviously (2e) and (2e*) cannot both be invariant with respect to this intervention. If (2e) is invariant with respect to the intervention, then (2e*) is not. If (1)–(2) is modular, then (1)–(2*) is not.

From the perspective of (1)–(2), (2*) entangles distinct mechanisms. Disrupting the mechanism represented by (1) changes the behavior of the mechanism which, according to (2*), links *Y* and *Z*. According to (2*), there is a single mechanism linking the variables *Y* and *Z*, the operation of which is represented by the coefficient *e*. From the perspective of (1)–(2) however, (2*) really represents the combined impact of two distinct mechanisms, one of which links *X* to *Y* and one of which links *X* and *Y* to *Z*. This is reflected in the fact that *e* is a function of *a*, *b*, and *c* and hence that (1) and (2*) are not independently changeable.

To forestall any possible confusion, we want to emphasize that we are not

claiming that one can determine whether a system of equations is modular simply from its syntactic form. We are not claiming that the mere fact that *e* can be written as a function of *a*, *b* and *c* shows that (1)–(2*) is not modular. After all, it is equally true that each of *a*, *b* and *c* in (1)–(2) could be written as a function of *e* and other variables. As the example above illustrates, it is nature and in particular facts about what happens or would happen under interventions that determine whether a given system of equations is modular. If (1)–(2*) fails to be modular then the result of an intervention on *Y* that disrupts the mechanism represented by (1) will be a change in the relationship between *Y* and *Z*—that is, there will be a change in the value of the coefficient *e*. In the example, an intervention that sets the value of *Y* to 3 results in $Z = 13$, not $Z = 15$. It looks as though the intervention on the mechanism represented by (1) also changes the allegedly distinct mechanism represented by (2*)—which is really just to say that these two mechanisms are not distinct and that the rule that each equation should represent a distinct mechanism has been violated. Of course it might turn out instead that one could intervene to disrupt (1) without changing (2*) and vice-versa. This would then show that the system (1)–(2*) is modular, in which case an argument parallel to the one given above would show that (1)–(2) is not modular.

Modularity provides a natural explication of what it is for a variable to be a direct rather than an indirect cause. Consider again the difference between the equation system (1)–(2) and the equation system (1)–(2*). If (1)–(2*) is modular (and thus correctly represents the causal structure), then if one were to intervene and set the value of *X* (within the relevant range of values), equations (1) and (2*) would remain invariant, and the values of both *Y* and *Z* would change. So in (1)–(2*)—as in (1)–(2)—*X* is a cause of both *Y* and *Z*. But in (1)–(2*), unlike (1)–(2), the value of *X* affects the value of *Z* only *via Y*. If one intervenes and sets the value of *Y*, then, according to (1)–(2*), one breaks the connection between *X* and *Z*. Under this intervention, changes in *X* should no longer affect *Z*. However since *X* appears on the right-hand-side of (2), it is not true, according to (1)–(2) that under an intervention that sets *Y*, changes in *X* will have no impact on changes in *Z*. Thus the difference between claiming, as (1)–(2*) do that *X* is an indirect cause of *Z*, the influence of which is entirely mediated by Y and claiming, as (1)–(2) do, that *X* is a direct cause of *Z* corresponds to the different predictions these two systems of equations make about whether *X* and *Z* would be independent under a hypothetical intervention on *Y*.

In the examples so far, systems of equations that are not modular are also not level invariant, but this is not always the case. Given any system of equations, one can always derive the associated set of what econometricians call 'reduced-form' equations. This set is observationally equivalent to the original system in the sense that it has exactly the same solutions for measured variables. One forms the reduced form equations by writing each endogenous

variable in the original system solely as a function of exogenous variables, with each endogenous variable appearing as a dependent variable on the left-hand side of exactly one equation. For example, since X is the only exogenous variable in (1)–(2), the reduced form equations associated with (1)–(2) are (1) $Y = aX + U$ and (3) $Z = dX + W$ (where $d = b + ac$, and $W = cU + V$). In general, each reduced form equation will be level invariant—that is, invariant under interventions on the variables on its right-hand side—if the equations in the original system are level invariant. If (1) and (2) are level invariant and modular, then (1) and (3) are level invariant, too. But the system (1)–(3) is not modular. This is because an intervention on the dependent variable Y in (1) will change the value of a and in doing so will change the value of d, since d depends on a. If one intervenes and changes the value of Y, then according to (1)–(2), the value of Z will change, while according to (1)–(3) it will not. So if (1)–(2) is modular, (1)–(3) cannot be. As this example shows, modularity is not implied by the requirement that each equation in a system be level invariant.

Like level invariance, modularity is not absolute—when one intervenes and sets the value of X (thereby disrupting the equation (mechanism) that determined the value of X), the other equations will remain invariant only for some range of values of X. We represent what modularity claims as follows:

> **MOD** For all subsets **Z** of the variable set **V**, there is some non-empty range **R** of values of members of **Z** such that if one intervenes and sets the value of the members of **Z** within **R**, then all equations except those with a member of **Z** as a dependent variable (if there is one) remain invariant.

When X is endogenous—that is, when X appears on the left-hand-side of some equation, then an intervention that sets the value of X disrupts that equation and **MOD** asserts the invariance of the other equations—which is what we called 'modularity' above. We emphasize that **MOD** requires that this be true only for interventions that set values of variables within some range of values, not for all possible interventions. However, **MOD** does require that every intervention that sets the value of a variable X within this relevant range leaves every equation invariant (except the one, if there is one, with X as a dependent variable). **MOD** implies in addition that all equations are invariant with respect to interventions that (within the relevant range) change the values of exogenous variables. When X is exogenous, there is no equation in which it appears as a dependent variable (though, of course, if one wished, one could add dummy equations stating that the value of X depends on some parameter). We thus have defined **MOD** so that it entails both what we called 'level invariance' and what we called 'modularity,' and we shall from here on mean **MOD** when we speak of 'modularity.' **MOD** is how we propose to make precise the idea that if one breaks the arrows into one variable in a graph by setting its value (within some range), then if the graph correctly represents causal relationships, all the arrows directed into other variables in the graph remain as they were.

One further implication of modularity is worth making explicit. If (1)–(2) satisfies modularity, then (2) is invariant both with respect to interventions that set the value of *X* and with respect to interventions that set the value of *Y* within the relevant range. It does not matter for the invariance of (2) (as long as we are within the range of values of *X* and *Y* for which (2) is invariant) whether the value of *Y* varies as the result of an intervention that sets the value of *Y* or as the result of an intervention that sets the value of *X* and permits the value of *Y* to be determined by (1).

A system of equations that lacks modularity will be difficult to interpret causally. Suppose, for example, that when one intervenes to change the value of *Y* (i.e. to disrupt the relation (1) between *X* and *Y*), equation (2) breaks down in such a way that the value of *Z* does not change. One would normally associate *Z*'s remaining unchanged when interventions change the value of *Y* with the absence of any causal relationship between *Y* and *Z*, but on a causal interpretation, equation (2) and the associated graphical representation assert that there is such a relationship. In this case, (1) and (2) do not fully capture the causal relationships in the system we are trying to model. In so far as modularity fails, the asserted causal structure fails to mirror what will happen under hypothetical interventions and, on a conception that connects causal claims to predictions about what will happen under such interventions, fails to represent correctly the causal structure of the system. Similarly, suppose that interventions on *Z* that disrupt (2) also disrupt (1), so that the value of *Y* changes for a fixed value of *X*. Again this is just the sort of behavior that one would ordinarily take to show that *Z* causes *Y*, but (1) denies this. As these examples illustrate, when modularity does not hold for a system of structural equations, the equations will fail correctly to describe the causal structure of the system they purport to represent.

A similar point can be made about the use of causal graphs. Modularity says that it should be possible to break the arrows into one variable, thereby changing its value within some range (or—within a similar range—to change the value of an exogenous variable) without changing the other arrows directed into any other variable in the graph. Like level invariance, which it subsumes, **MOD** implies that it should be possible to change the value of a variable *X* (within some suitable range) without making arrows into variables that depend on *X* appear or disappear. If these requirements are violated, then the graph does not represent the causal structure correctly.

6 Modularity, linearity and mechanisms

So far we have argued that, for systems of equations like those considered above, modularity holds—that is, disrupting the relations expressed by one equation by setting the value of its dependent variable leaves the other

equations alone—if and only if breaking the arrows into any vertex in the associated graph leaves the arrows into all other vertices undisturbed. Only in this case will the equations and the graph provide clear and correct answers to questions about what will happen under hypothetical interventions.

In the case of linear, or more generally additive relations, one can impose an even stronger invariance condition. One can require that the individual coefficients be *separately* invariant under interventions that change the value of other coefficients. Let us call this 'coefficient invariance.' Coefficient invariance should hold if each *term* in each equation represents a distinct causal mechanism (and thus acts independently of the mechanisms registered by other terms). For example, in a linear equation such as (2), coefficient invariance requires that the coefficient *b*—the direct effect of *X* on *Z*—be invariant under some suitable range of interventions that alter how *Y* affects *Z*—that is, change the coefficient *c* and *vice versa*. If the mechanism by which *X* affects *Z* is genuinely distinct from the mechanism by which *Y* affects *Z*, then the mechanisms should operate independently and it should be possible to interfere with one without interfering with the other. Like modularity, the coefficient invariance condition has a simple graph-theoretic interpretation: When one has additive relations, one can interpret individual edges as representing separate causal mechanisms. Breaking any single arrow anywhere in the graph, leaves all the other arrows undisturbed, including other arrows directed into the same variable as the broken arrow.

Coefficient invariance is a restrictive condition: it will be violated whenever additivity is, or when the causal relationship between two variables depends on the level of a third variable. If one thinks of individual causes of some effect *Y* as conjuncts in a minimal sufficient condition for *Y* (or the quantitative analogue thereof)—that is, as 'conjunctive causes'—then the relationship between an effect and its individual causes will not satisfy coefficient invariance. Removing the arrow between an individual cause *X* and one of its effects *Y* will not leave the coefficients relating *Y* to its other causes unaffected. Coefficient invariance can be expected to hold only when the vertices in a graph that represent causes of *Y* in fact represent (components of) *separate* minimal sufficient conditions—i.e. 'disjunctive causes.' Thus, for example, one would expect coefficient invariance to hold in (2) if the mechanism by which rainfall directly influences the pea harvest is independent of the mechanism by which plant height influences the harvest. Otherwise one would expect coefficient invariance to fail. Notice that (unlike models that explicitly represent the functional form of the relationship between variables by means of equations) the graphical representation is insensitive to the difference between conjunctive and disjunctive causes. Consider, for example, the graph shown in Figure 7.

X and *Y* are both causes of *Z*. Suppose that the graph represents the relationship between the acceleration of a falling steel ball bearing (*Z*), the strength of

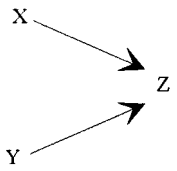

Fig. 7

the magnetic field (*X*) and the strength of the gravitational field (*Y*). In this case (which, like a linear structural model, involves 'disjunctive causes'), changing the value of *X* or disrupting the relationship between *X* and *Z* leaves the arrow between *Y* and *Z* unaffected. But in the causal modeling literature *Z* might also represent the illumination of a light that is controlled by a dimmer switch that pushes in to turn the light on and rotates to control its intensity. *X* could be a binary variable representing whether the switch is pushed in or not. *Y* would be a continuous variable representing how far the switch is rotated. With such 'conjunctive causes,' changes in the value of one variable or disruption of the relationship between one variable and the effect, may break the other arrow. Nothing in Figure 7 indicates which sort of causal relations is being depicted. Coefficient invariance is an implication of the independence of distinct mechanisms in the special case of additive relations, in which case the individual terms in an equation can be taken to represent separate mechanisms. It is not a general condition that one would expect causal relations to satisfy. 'Interaction effects'—that is, failures of coefficient invariance—are in fact common. Coefficient invariance obviously implies modularity, but modularity permits the causal factors appearing in any single equation to interact. Modularity only requires that the mechanism described by each individual equation be distinct from the mechanisms described by the others.[15]

Modularity thus does not presuppose linear or additive relations. Suppose that, instead of (1) and (2) above, one had

$(1')$ $Y = F(X) + U$

$(2')$ $Z = G(X, Y) + V$

where *F* and *G* are arbitrary functions. In the case of $(1')$–$(2')$, **MOD** says that *F* and *G* are invariant to interventions that set the value of *X* within some range of values of *X* and that *G* is invariant to interventions that set the value of *Y*

[15] Although we speak of the mechanism an equation captures, we do not interpret modularity as ruling out the possibility that a single structural equation might describe the operation of more than one mechanism. In the case of additive relations a single structural equation may express several mechanisms that are distinct from one another and which can be separately disrupted. When coefficient invariance does not hold, the mechanisms an equation describes will not be distinct in this way, and we treat them as one. See below, Section **7**.

within some range of values of Y. Whether this condition is satisfied is completely independent of whether F and G are linear. Equations $(1')$ and $(2')$ still assume that the contribution of the error terms is additive, and this assumption might not be dispensable, but since it is always possible to expand the graph and incorporate more variables, this restriction on the generality of our claims concerning modularity would be a mild one.[16]

The central presupposition underlying this discussion of modularity and coefficient invariance is that if two mechanisms are genuinely distinct it ought to be possible (in principle) to interfere with one without changing the other. Conversely, if there is no way, even in principle, to decouple mechanisms—to interfere with one while leaving another alone—then the mechanisms are not distinct. We take this as a criterion for counting mechanisms. The mechanisms by which gravitation and friction influence the acceleration of a block sliding down an inclined plane are distinct. By greasing the plane one can alter the relation between the velocity of the block and the frictional force that resists its movement without altering the relation between gravitation and acceleration. There are thus two mechanisms involved. Greasing the plane changes the equation or term relating velocity and frictional force without changing the equation or term relating acceleration and gravitational force. By contrast, it is plausible that the mechanism by which water influences plant growth is not distinct from the mechanism by which fertilizer influences growth. Probably, the relationship between water and plant height varies depending on the level of fertilizer, so that an intervention on the latter will alter the former relationship. If this is the case, there is a single mechanism mediating the combined influence of both water and fertilizer on height and to express that relation with a linear equation would be misleading about the true causal structure: the equation would not be invariant to an intervention that changed the level of one of the variables.

This understanding of distinctness of mechanisms plus the assumption that each equation expresses a distinct mechanism implies modularity: it is, in principle, possible to intervene and to disrupt the relations expressed by each equation independently. Similarly, this convention concerning distinctness of mechanisms plus the assumption that the whole set of arrows into each vertex represents a distinct mechanism implies that an intervention that breaks an arrow into one vertex should leave arrows into other vertices unchanged. In the case in which each arrow (or each term in the equations) represents a distinct mechanism, the convention concerning the distinctness of mechanisms implies coefficient invariance.[17]

The view that mechanisms are distinct if and only if it is in principle possible

[16] We are indebted here to Nancy Cartwright's comments.

[17] We do not maintain that the invariance conditions we have discussed exhaust all those that one would expect to find, and indeed below in Section **10** we shall discuss some others.

to interfere to disrupt one while leaving the other alone is not an arbitrary stipulation on our part. On the contrary, it is implicit in the way people think about causation, and, in different contexts, both of us have argued that this sort of independence is essential to the notion of causation. Causation is connected to manipulability and that connection entails that separate mechanisms are in principle independently disruptable. Causes could not be levers for moving their effects if the linkages between X and the set of its (conjunctive) causes could not in principle be disentangled from the linkages between other effects and their causes.

For example, when one thinks of atmospheric pressure as a common cause of barometer readings and storms, one supposes that the mechanism by which atmospheric pressure influences the barometer reading is distinct from the mechanism by which it influences the onset of the storm—this appears to be part of what it means to say that atmospheric pressure has two joint effects rather than just one.[18] If these mechanisms were not distinct, it would be more appropriate to think of atmospheric pressure as producing a single composite effect, of which barometer readings and storms are components. If these mechanisms are distinct, there must in principle be some way of bringing on or impeding storms without interfering with the mechanism by which atmospheric pressure causes barometer readings. Such interference may not be feasible, but if it were not even conceivable, there would be only one mechanism, not two. Similarly, there must be some way to set the barometer reading without interfering with the mechanism whereby atmospheric pressure causes storms. Manually moving the barometer dial to a new position constitutes such a interference: This operation alters the relation between the atmospheric pressure and the barometer reading without altering the relation between atmospheric pressure and storms. To count as a distinct mechanism at all, a putative mechanism must exhibit some degree of robustness or insensitivity to context in its operation, and this includes insensitivity to changes in other mechanisms.

As additional motivation for this idea, we should also note that without some such specification of what it is for mechanisms and causal relationships to be distinct, it is easy both to trivialize the Markov Condition and to generate spurious counterexamples to it, as the following examples show. First, suppose that X and Y are correlated, with X caused by X^* and Y caused by Y^*. The correlation between X and Y results from the fact that X^* and Y^* are themselves correlated. Even if there is no variable that, intuitively, would count as a common cause of X and Y, one can always preserve the Causal Markov

[18] Since we are identifying a mechanism with an equation with a single dependent variable on its left-hand side or, equivalently, with the set of arrows into a particular variable or vertex, there will always be exactly as many distinct mechanisms as there are distinct effects. We are indebted here to Nancy Cartwright's queries.

Condition by taking the composite event X^*Y^* as the common cause of X and Y. However, this trivializes the condition. If **CM** is to have any substance, one needs some way of capturing the idea that X^*Y^* is not a genuine common cause. One can avoid this sort of trivialization and capture what it means to say that X^*Y^* is not really one event but rather two events by noting that it is possible to intervene to change the value of Y^* independently of the value of X^* and hence independently of the mechanism by which X^* causes X and similarly possible to change X^* independently of Y^* and hence of the mechanism that links Y^* to Y. Since it is in principle possible to intervene to set the value of Y^* without changing X^* and since the intervention leaves X unchanged, X^*Y^* is not a single composite event which is a common cause of X and Y.

Related remarks can be made about events or variables that are logically or analytically rather than causally connected. Suppose that X and Y are statistically independent. From these variables one can always construct other variables—for example, $Z = X + Y$ or $W = X - Y$—that are statistically dependent on X and Y. However, we do not want to interpret the Markov Condition in such way that this sort of dependence establishes that X and Y cause Z or W. It is, of course, easy enough—and perfectly reasonable—to stipulate that causal relations obtain among *distinct* events or variables; but logical or conceptual distinctness may not always be as easy to judge as it is in this case. The analytical connection between X, Y and Z or W will, however, also show itself in the fact that there is no conceivable way to intervene to change the value of Z or W that is not simultaneously an intervention to change the value of X or to change the value of Y, and there is no way to intervene to change the value of X or to change the value of Y that is not simultaneously an intervention to change the value of W or Z. A similar sort of analysis may be applied to the cases of logical or conceptual dependence, described in Section **2**, which have been advanced as counterexamples to **CM**.

As a final illustration consider the following putative counterexample to **CM**. Suppose that radium atoms have a probability of 0.5 of emitting an alpha particle (composed of two protons and two neutrons) during a certain time interval. Call the decay event C and consider the event A consisting of the emission of two neutrons and the event B consisting of emission of two protons during this time interval. $\Pr(A/C) = 0.5$, and $\Pr(B/C) = 0.5$, but $\Pr(A\&B/C)$ is also 0.5 rather than 0.25. Does this represent a failure of **CM**? Intuitively, A and B are not distinct events, produced by distinct mechanisms. There is just one event here—the emission of the alpha particle, produced by just one causal mechanism. This is again reflected in the fact that there is no conceivable way of interfering in the mechanism that produces A while leaving the mechanism that produces B undisturbed or *vice versa*. Since this is thus a case in which a single stochastic cause produces a single effect, it is not a counterexample to the Markov Condition.

On the other hand, if this treatment is not allowed, then violations of the Markov Condition will be ubiquitous—all cases of stochastic causality, whether microscopic or macroscopic, in which the effect event is decomposable into parts (which is to say virtually all cases of stochastic causality) will count as counterexamples. Again we see that sensible application of the Markov Condition rests on prior assumptions about when variables and mechanisms are distinct. We will return to this theme in Section **9** when we discuss the EPR phenomenon.

We conclude these remarks with a final observation. The central ideas described in this section and the previous one—modularity and distinctness of mechanisms—are strongly modal or subjunctive. They have to do with whether equations or arrows *would* remain unchanged if one *were to* intervene to disrupt equations or break arrows. When we say that equations in a system of equations are modular we are making claims about relationships between different systems of equations or between different graphs and probability distributions. Thus if the system (1)–(2) and the corresponding graph capture the causal relations among rainfall (X), plant height (Y), and harvest (Z) correctly, this tells one not just how to represent the situation in which Y directly depends on X, and Z directly depends on both Y and X, but also how to represent the new situation that would result if an intervention were set the value of Y and break the arrow between X and Y.

7 Modularity, mechanisms, and an argument for the Causal Markov Condition

This long discussion of level invariance, modularity, and the distinctness of mechanisms may appear to have strayed far from the Causal Markov Condition, but in fact modularity turns out to bear a close and suggestive relationship to **CM**. Given determinism, modularity (**MOD**), and causal sufficiency, the value of any variable Y in **V** can change with an intervention with respect to a distinct variable X in **V**, only if the value of one of the parents of Y changes. The value of one of Y's parents can change only if the value of one of its parents changes, and so forth. Since **V** is a finite set, **MOD** implies that either X is an ancestor of Y or the value of Y does not change with an intervention with respect to X. In other words, if X does not cause Y, then the *value* of Y remains unchanged given an intervention that sets X. This (of course) restates the fundamental principle relating causation and manipulation that we introduced above.

Assume in addition that an intervention i_X that sets the value of a variable X can be treated as a random variable. Since this intervention is not an effect of any variable in **V** and is causally related to variables in **V** only by virtue of being a direct cause of X, **CM1** implies that it is probabilistically independent of all other interventions and of everything that is not an effect of X. Rather

than referring to intervention variables explicitly, we shall make use of some useful terminology proposed by Pearl ([1995]) and represent the random variable whose values are the values of X when these are set by intervention as 'set-X.' So an intervention that sets the value of X at x is treated as the random variable set-X assuming the value set-$[X = x]$.

Given **CM1** and the assumption that interventions can be treated as (independent) random variables, **MOD** thus implies by the argument given immediately above:

> **MOD*** For all distinct variables X and Y in **V**, if X does not cause Y, then $\Pr(Y\&\text{set-}X) = \Pr(Y).\Pr(\text{set-}X)$.

MOD* formulates the basic principle relating causation and manipulation as the claim that whenever X does not cause Y, then Y and set-X are probabilistically independent. To recapitulate, we get to this claim in the following way. Modularity (**MOD**) implies that the *equations* for all variables other than X are invariant to interventions that set-X. This implies that the *values* of all variables that are not effects of X are not changed by interventions with respect to X. If interventions can be treated as random variables, one arrives at **MOD***, the probabilistic independence of set-X of all variables that are not effects of X.

We shall now argue that, given **CM1**, causal sufficiency and the assumption that there are unrepresented causes, **MOD*** holds if and only if **CM** does. (Given causal sufficiency, the unrepresented causes cannot of course be common causes.) The independent disruptability of each mechanism turns out to be the flip side of the probabilistic independence of each variable conditional on its direct causes from everything other than its effects. Here is the argument. Suppose that X is a dependent variable. By assumption, in addition to its represented direct causes (**Parents**(X)), it has some unrepresented causes whose effect is summarized by the error term U_X in the equation for X. By definition, these are direct causes, and, given causal sufficiency, they bear no causal relationship to any variable other than X apart from those which result from their being direct causes of X. U_X thus satisfies the definition of an intervention with respect to the X, and given **CM1**, U_X has no probabilistic relations to any variables, except for those that arise from its being a direct cause of X. In particular, U_X is probabilistically independent of **Parents**(X).

Since, conditional on **Parents**(X), the only source of variation in X is U_X, any variable Y in **V** distinct from X can covary with X conditional on **Parents**(X) if and only if it covaries (unconditionally) with U_X. (To see that this is true, recall that the distribution of X conditional on any set of variables **W** is the distribution X would have if the variables in **W** were fixed. Since $X = f(\textbf{Parents}(X)) + U_X$, the conditional distribution of X/**Parents**(X) would

be degenerate—X would have an unvarying value—were it not for the variations caused by U_X. So Pr(X/**Parents**(X & Y) = Pr(X/**Parents**(X)) if and only if Pr(Y & U_X) = Pr(Y).Pr(U_X).)

Since U_X satisfies the definition of an intervention, one can infer the following claim:

> (=) Pr(Y & set-X) = Pr(Y) . Pr(set-X) if and only if Pr(X/**Parents**(X) & Y) = Pr(X/**Parents**(X)).

(=) says that Y is independent of set-X if and only if it is independent of X conditional on the set of direct causes of X. Given causal sufficiency, the existence of unrepresented causes, **CM1**,[19] the definition of an intervention, and the assumption that interventions can be represented as random variables, the independence of X and Y conditional on **Parents**(X) is a proxy for Y's independence of interventions that set the value of X. One can 'read off' which variables are independent of interventions and which are not by examining the probabilities conditional on the direct causes. For example, if, as Fisher hypothesized ([1959]), smoking does not cause cancer but is related to it only as an effect of a common cause, then if one brought about people's smoking by intervention, one should find no more lung cancer than among a control group who are not subject to the intervention. For moral reasons, this experiment cannot be performed. Given (=), one can find out whether cancer is independent of set-smoking (and thus whether it is caused by smoking) by determining whether cancer is independent of smoking conditional on the direct causes of smoking.

The rest of the argument for the equivalence of **MOD*** and **CM** (conditional, of course, on the assumptions used to derive (=)) is trivial. **MOD*** says that the left-hand side of (=) is satisfied whenever Y is not an effect of X. **CM** says that the right-hand side of (=) is satisfied whenever Y is not an effect of X. So if one accepts (=), then one can accept **MOD*** if and only if one accepts **CM**. In relying on (=), one is of course implicitly invoking the assumptions used in the proof above: **CM1**, causal sufficiency, the existence of unrepresented causes, and the treatment of interventions as random variables. Given these assumptions, the Causal Markov Condition states how probability distributions

[19] One might object that we are here relying on something stronger than **CM1**, which only applies to unconditional dependencies. It is arguable that we are here supposing that if X and Y are dependent conditional on some set $\mathbf{V}_X$ of variables, then this conditional dependency must itself have a causal explanation in the sense that either X causes Y, Y causes X or the conditional dependence is due to some common cause of X and Y that is not in $\mathbf{V}_X$. But all claims about causal relations at the type level must be relativized to some set of circumstances. So it seems fair to consider what **CM1** implies in circumstances in which (for example) the members of **Parents**(X) are unchanging. Alternatively, consider the variable set **V****Parents**(X). If **V** is causally sufficient, then in the circumstances in which there is no variation in any of the variables in **Parents**(X), **V****Parents**(X) must be causally sufficient as well. **CM1** can then be used to infer unconditional probabilistic dependencies and independencies among the variables in **V****Parents**(X).

reflect invariance conditions such as modularity. The Causal Markov Condition translates the relation between causation and manipulability into a claim about the relation between causation and probability distributions.

Some readers may not find this argument from modularity to the Causal Markov Condition compelling, because they find the assumptions it requires, such as **CM1**, causal sufficiency, and the existence of unrepresented causes, as questionable as the Causal Markov Condition itself. It might seem particularly odd that we need to assume that the set **V** does not include all the causes of any variable. The reason this assumption is needed is that with deterministic relations, probabilities would otherwise become one or zero. Under determinism the 'only if' part of (=) would fail, since one would have Pr(X/**Parents**(X) & Y) = Pr(X/**Parents**(X)) = 1 even if X causes Y and Y and set-X are not independent. The 'only if' part of (=) requires that there be an unrepresented source of variation. When we retrace this argument in indeterministic circumstances, we will not need this assumption. Even readers who question our assumptions, should find this long discussion of modularity and its relations to **CM** of interest for the tight links it reveals between **CM** and the view that causes can be used to manipulate their effects. If conditional independencies echo independencies given interventions, then **CM** follows from the view that causes can be used to manipulate their effects.

8 Strong independence and the Causal Markov Condition

Although the relations between a manipulability view of causation and the Causal Markov Condition developed in the last three sections are central to this essay, there are easier ways of arguing for **CM**. For example, consider what SGS call 'pseudo-indeterministic' systems, where the values of left-hand side variables are determined by the values of the variables on the right-hand side and error variables. SGS claim that in pseudo-indeterministic systems, **CM** must hold ([1993], p. 57). If one assumes that there are unrepresented causes of all the variables, then causal sufficiency and **CM1** imply that each error variable, which summarizes the effects of the unrepresented causes is probabilistically independent of the variables on the right-hand side of the equation in which it appears and of all other error variables. The Causal Markov Condition then follows trivially. Any variable Y in **V** distinct from X can covary with X conditional on **Parents**(X) if and only if it covaries (unconditionally) with the error term in the equation for X. Since the error term is independent of all variables that are not effects of X, X must be independent of everything except its effects conditional on its parents.

Like all the other arguments in defense of **CM**, this one depends on very strong independence assumptions. The causes that are not represented in the graph have to be probabilistically independent of one another, and they have to

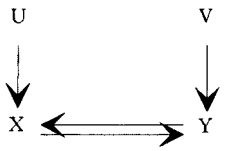

Fig. 8

be independent of all the represented causes of the same variable. Such assumptions are not always plausible—for example, they may fail to hold when the situation under investigation is represented by a non-recursive model. Consider, for example, the causal structure represented by the following two equations, $X = aY + U$ and $Y = bX + V$, with U and V as error terms. This corresponds to the graphical structure shown in Figure 8. Because U is a direct cause of X and X is a direct cause of Y, U will be correlated with the variable Y, which is a cause of X in the first equation. For systems of this sort, the Causal Markov Condition as formulated here does not hold, although related conditions can be defended (see Spirtes [1995], and Koster [1996]).[20]

In Chapter 12 of *Causal Asymmetries*, Hausman offers a more elaborate argument for **CM** which relies on the notion of a 'non-accidental connection' rather than on the notion of invariance to intervention. If, as we believe, the understanding of causation requires some sort of counterfactual or modal notion, which bears no simple relation to observation, then (as Cartwright [1989] has argued) it is by no means obvious that the best way to proceed is to follow neo-Humeans and to take the notion of a law of nature as one's only primitive. One possibility, explored above and developed in more detail in Woodward ([1997], [forthcoming b]) is to rely on the notion of invariance under intervention. Another possibility, which Hausman explores, is to take the notion of a non-accidental dependency or connection as a modal primitive. Both the notion of a non-accidental dependency and the notion of a relationship that is invariant under interventions represent different attempts to rehabilitate the idea, repudiated by Hume, that causation involves 'necessary connections.' Although we disagree about which of these alternatives is the most promising path to follow, we both maintain that the modal component in the notion of causation is not best understood in terms of the notion of a law of nature.

The strategy of this section is to begin with Hausman's notion of a non-accidental dependency and show how **CM1** arises naturally from the rough correspondence between non-accidental dependencies and probabilistic dependencies. We will then show how **CM2** can be deduced from **CM1** and an independence condition that Hausman maintains is a boundary condition for

[20] David Papineau ([1985], p. 283) has also sketched an argument for binary variables that strong independence assumptions imply that deterministic common causes screen off their separate effects.

the applicability of causal explanations, but which also derives from assumptions we have already made.

On Hausman's approach, the key notion is that of a 'non-accidental' dependency or connection between properties or between property instantiations. Non-accidental connections are not directly observable, but they are not ineffable either. They are fallibly—but reliably—indicated by sample correlations. Just as empiricists are prepared to accept the existence of unobservable entities when their postulation makes an observable difference, so they should be prepared to accept the existence of non-accidental connections between properties, because these too make an observable difference. The discovery of mechanisms linking apparently correlated variables often elucidates the nature of postulated non-accidental connections. For example, the non-accidental connection between pressing keys on a keyboard and characters appearing on a computer screen can be associated with a complicated mechanism. But unlike Cartwright ([1997], p. 344) we deny that non-accidental dependencies can always be linked to mechanisms. Those between variables in fundamental laws may, in contrast be 'brute.'[21]

Unconditional sample dependencies fallibly indicate non-accidental connections. Most of these non-accidental connections among properties or variables obtain because their instances (or located values) are related as cause and effect or as effects of a common cause. When two light bulbs are wired in parallel into the same circuit, so that they both go on when a switch is thrown, the correlation between the switch position and the lighting of either bulb and the correlation between the lighting of the two bulbs both arise because the switch position and the bulb lightings are linked by causal mechanisms. In contrast, when two switches are wired in serial into the same circuit, so that they both must be closed for a bulb to light, no mechanism connects the switch positions and no (unconditional) dependency is forged between their positions. Quantum mechanics apparently presents us with additional (symmetrical) non-accidental connections. Coincidence is a fourth variety of sample dependency. When one has a sample dependency that is not coincidental, then one has a probabilistic dependency as well as a non-accidental connection. The notion of a probabilistic dependency and of a non-accidental connection thus largely coincide.[22] In this way one arrives back at **CM1**. But it seems that one has merely followed a long detour to arrive back where one began. There is no

[21] These questions do not need to be answered in this essay. Notice that a non-accidental dependency may reflect the operation of multiple mechanisms. As argued above, the mechanisms by which a cause Z produces two effects X and Y, where X and Y are not also related as cause and effect should be regarded as distinct. The non-accidental relationship between X and Y obtains because the mechanism that gives rise to X shares a component (Z) with the mechanism that gives rise to Y, not because a single mechanism links X and Y. See Sections **5**, **7** and **9**.

[22] The coincidence is not perfect because causal influences along different pathways can cancel out. See Hausman ([1998], Chs 10 and 12).

independent argument here for **CM1**. The discussion merely restates the intuition behind **CM1** in a different way.

With the help of two additional elements, this framework does, however, permit an argument for **CM2**. The first of these two additional elements is a strong independence condition: Every variable X in **V** has some cause that is not in **V**, which is distinct from everything in **V** and bears no causal relations to any variable in **V** apart from those that result from its being a direct cause of X. Although the formulation of this condition is different, it obviously is closely related to the claim that every variable in **V** has an intervention variable directed into it—i.e. that interventions are possible with respect to every variable. The strong independence assumption is a basic axiom for Hausman, but it may also be derived from (a) the assumption that variables in **V** have unrepresented causes and (b) causal sufficiency. Causal sufficiency implies that the unrepresented causes of X can have no causal relations to any other variables except those that result from their being causes of X. Since the graph represents only the members of **V**, the graph does not represent this independent source of variation. **CM1** implies that the unrepresented causally independent source of variation in X is probabilistically independent of all of X's represented causes. Although this account does not rely explicitly on any claims concerning the relations between manipulability and causation, there are important implicit connections (see Hausman [1998], Chs 5, 7). Hausman maintains that a more general formulation of the strong independence condition is a boundary condition for the applicability of causal concepts and for the possibility of specifically causal explanation. When it apparently breaks down—as it does in the case of the EPR phenomena—then causal explanations are out of place.[23]

Second, we shall rely on the assumption that a conditional sample dependency or partial correlation operationalizes a counterfactual concerning what non-accidental connections would obtain if the condition were met. The existence of a sample partial correlation between X and Y conditional on some value of Z (*not* conditional on an intervention that sets the value of Z) is evidence that in circumstances in which Z does not vary X and Y would be probabilistically dependent. **CM1** then justifies the conclusion that they would be related as cause and effect or as effects of a common cause—that is that there would be a non-accidental connection between X and Y if the value of Z were observed to remain unchanged at z.

Two examples may help clarify what is meant. Suppose that a switch is connected in parallel to two light bulbs, *B1* and *B2*, so that when it is thrown they both go on. By the strong independence assumption, each of the two

[23] See Hausman ([1999]). As argued below in Section **9**, a similar conclusion follows in Woodward's framework, since it is impossible to carry out an intervention on either of the two separated particles in the EPR experiment.

variables representing whether the bulbs are illuminated has its own independent source of variation (such as, for example, frayed wiring or a battery backup). So in a possible world in which the switch was held fixed in its 'on' position, the illumination of *B1* would depend only on causes that are independent of the causes of *B2*. So there would be no non-accidental dependency between the illumination of the two bulbs—just as there is no conditional probabilistic dependency.

Consider next a light controlled by two switches, like the light in the upstairs hallway of many houses. If only one of the two switches is 'up,' then the light is on. If both of the switches are up or both down, the light is off. There is no unconditional correlation between the position of the switches. Consider then a world in which the light is always on. In this world, the positions of the switches must be perfectly negatively correlated, and by **CM1** the negative correlation in this world must have a causal explanation. There would have to be some sort of mechanism linking the switch positions so that any variation affecting the position of one influenced the position of the other, too.[24]

Given strong independence and this way of operationalizing counterfactuals about what non-accidental connections would obtain, one can argue as follows in defense of **CM2**. Assume first that the causal graph under consideration is causally sufficient and that causation is deterministic. Suppose then that X and Y are probabilistically dependent conditional on **Parents**(X)—that is, conditional on all the direct causes of X that are in the set **V**. From the counterfactual operationalization, it follows that X and Y would be non-accidentally connected if all the parents of X were unchanging. So if all the parents of X were unchanging, then (by **CM1**) X and Y would be connected as cause and effect, or they would be connected as effects of a common cause, or they would be connected in that remarkable way that the spins of electron pairs are connected in EPR phenomena. If all the represented direct causes of X were unchanging, then X and Y would not be effects of a common cause because the only source of X's variation, by the strong independence condition, bears no causal relation to Y except via causing X. Nor could Y cause X, because all the causal influences on X apart from its unrepresented independent cause have been frozen. So either one has EPR-like phenomena or X causes Y. The EPR correlations are symmetrical, and so the first possibility can be eliminated if

[24] As Paul Humphreys correctly pointed out to us, all the sources of variation in the positions of the two switches cannot be *probabilistically* independent of one another if the positions of the switches are perfectly correlated. This observation could be understood as a criticism of either the counterfactual operationalizing assumption or the strong independence condition. We believe that it shows that the counterfactual operationalizing assumption must be revised before it applies to causal relations involving *cycles*, but that Humphreys' observation constitutes no objection to either strong independence or the counterfactual operationalizing assumption when the relations are acyclic. In order to represent a system that includes a device that ties the switches together so that whenever one is thrown the other is too, a causal graph would have to contain a cycle, even though at the token level the causal relations are asymmetric. We do not have space here to discuss the appropriate probabilistic conditions to impose on graphs containing cycles.

X and *Y* are not correlated conditional on all the parents of *Y*. Thus one can derive the claim that if *X* and *Y* are probabilistically dependent conditional on all the parents of *X*, and they are probabilistically independent conditional on all the parents of *Y*, then *X* causes *Y*. This condition is logically weaker than **CM2** and **CM**. If one sets aside the possibility of EPR-like phenomena (as one should in deterministic circumstances), one derives **CM** from **CM1**.

Many readers may feel that **CM** and **CM2** are more plausible than the metaphysical assumptions in the arguments in this section in support of them. This is not the occasion to assess this framework, which is developed at length in Hausman's *Causal Asymmetries*. The critical role played by the independence assumption is, however, worthy of some comment. Because every variable has a strongly independent cause, one could in principle intervene independently with respect to each variable. If one intervenes with respect only to *X*, and *Y* wiggles, too, and wiggling has a causal explanation, then *Y* cannot cause *X* and *X* and *Y* cannot be related only as effects of a common cause. If the joint probability distribution over the variables in the graph was generated by deterministic processes that include such 'wiggling,' then **CM** should hold. As the last section demonstrated, the connection between **CM** and a manipulability conception of causation is very close.

9 Cartwright's objection

The long arguments given thus far for **CM** suppose that causation is a deterministic relation. But some causal relations in fundamental physics appear to be indeterministic, and (at least at the level of analysis at which they are stated) so do many causal relations in the social, behavioral, and biomedical sciences. Suppose, for example, that one is interested in the relationship between schooling and earnings in the contemporary US. Even if this system is deterministic at some sufficiently fine-grained level, it is unlikely that there exist deterministic relationships between education, earnings and other variables such as social class or family background. Is there any reason to believe that indeterministic causal relations should also satisfy **CM**?

Nancy Cartwright has argued that the answer is 'no.' She offers the following example:

> Two factories compete to produce a certain chemical, which is consumed immediately in a nearby sewage plant. The city is doing a study to decide which to use. On Mondays, Wednesdays and Fridays chemicals are bought from factory Clean/Green. On Tuesdays, Thursdays and Saturdays, from Cheap-but-Dirty. Cheap-but-Dirty employs a genuinely probabilistic process to produce the chemical: the probability of actually getting it on any day the factory operates is only about 80%. So on some days the sewage does not get treated, but the method is so cheap the city is prepared to put up with that. What they object to are the terrible pollutants that are emitted as a by-product.

> That's what's really going on. But Cheap-but-Dirty will not admit to it. They suggest that it must be the use of the chemical in the sewage plant itself that produces the pollution. Their argument relies on the 'common-cause condition'. If there *were* a common cause (*C*), producing both the chemical (*X*) and the pollutant (*Y*) then conditioning on the information about which factory was employed (that is looking separately at the data from Mondays, Wednesdays and Fridays and that from Thursdays, Thursdays and Saturdays) should screen off the chemical from the pollutant. We should then expect
>
> P(X/C.Y) = P(X/C).
>
> But it does not.
>
> We know *why* it does not. Cheap-but-Dirty's process is a probabilistic one. Knowing that the cause occurred will [not] tell us whether the product resulted or not. Information about the presence of the by-product will be relevant since this information will tell us (in part) whether, on a given occasion, the [producing] cause actually 'fired' [. . .]
>
> What has gone wrong? [. . .] My own hypothesis is that our intuitions are still too deterministically schooled. For a deterministic case, the occurrence of the cause is co-extensive with its operation to produce its effect. So a very important question concerning the total set of causes and effects is concealed. That is the question of what the relationships are amongst the operations to produce the different effects . . .
>
> Consider a simple case of a cause *C* with two separate effects *X* and *Y* [. . .] Looking at the effects, we have an event space with four different outcomes:
>
> +X,+Y; ¬X,+Y; +X,¬Y; ¬X,¬Y.
>
> If causality is to be fully probabilistic, nature must set probabilities for each of these possible outcomes to result. That is, conditional on *C*, nature must fix a joint probability over the whole event space [. . .] Nothing in the concept of causality or probabilistic causality constrains how it should be done. The so called 'common cause condition' is satisfied only for a very special case, the one in which:
>
> $P_c(+X+Y).P_c(\neg X\neg Y) = P_c(+X\neg Y).P_c(\neg X+Y)$.
>
> Put in my earlier informal language, what is being fixed here is the relationship between *C*'s operation to produce *X* and its operation to produce *Y* (Cartwright [1993], pp. 115–6).

These remarks raise a number of issues. Conditioning on the day of the week is tantamount to conditioning on factory type (Clean/Green or Cheap-but-Dirty). This is presumably coarser-grained than a variable that reflects the full details of the operation of the specific chemical processes in the two factories involved in the production of X and Y. (It is essential to distinguish between the chemicals and the events of their production and between the factory-type and the details of the operation of the production process employed in the two factories. Let C be factory-type and X and Y the chemicals, while *C* is

the operation of the production process, *X*, the production of chemical X, and *Y* the production of chemical Y.) One possible interpretation of Cartwright's example is then that the common cause *C* has been characterized in insufficient detail. Understood in this way the example illustrates a point to which we have already agreed: In many practical applications researchers may be forced to work with variables that are imprecisely defined or measured and because of this, the Causal Markov Condition may fail with respect to those variables. However, it seems that Cartwright intends the example to motivate a more fundamental objection to the Causal Markov Condition. She denies that there is *any* screening-off common cause, regardless of how the variables are designated or measured.

As an initial observation, note that this denial sits uneasily with the sort of probabilistic theory of causation that Cartwright advocated in her ([1979]) and that many others have since advocated. According to these theories, the causes of an event of kind *Y* are those prior factors that are probabilistically relevant to *Y* in causally homogeneous background conditions. In Cartwright's example, *X* remains probabilistically relevant to *Y* even after one conditions on *C*—which by hypothesis is the only cause of *X* and *Y*. So, if *X* and *Y* are not simultaneous or if one relaxes the requirement that causation requires the temporal priority of the cause, it follows within the framework of standard probabilistic theories of causation that *X* is a cause of *Y* or *vice versa*.[25]

Consider now a significant feature of Cartwright's description of this example. She speaks of C's 'operation' or 'firing' to produce X and Y. This description plays an important rhetorical role in making it seem intuitively plausible that there is no reason why the firings to produce chemicals X and Y must be uncorrelated when one controls for the common cause. However, on closer examination, the description seems to undermine Cartwright's treatment of the example. If the firing of C is an event of some kind (and what else could it be?), one is faced with the question of its relationship to *C*, *X*, and *Y*. For example, if one conceives of the firing of C as a single event *F*, then does the event variable *C* cause *F*, which in turn causes *X* and *Y*, as in Figure 9a? If so, then there is no counterexample to the Markov Condition, since by hypothesis *F* is a deterministic cause of *X* and *Y* (once the firing *F* occurs *X* and *Y* will occur). It is true that *C* doesn't screen off *X* from *Y*, but *C* is not a direct cause of *X* and *Y*—*F* is the direct cause, and it does screen off. A similar conclusion follows if one thinks of *F* not as an effect of *C* but rather as a more precise

[25] Cartwright is aware of this point. In her ([1989]) she rejects the probablistic theory of causation and the associated principle **CC** advocated in her ([1979]) on a number of different grounds, including the possibility of non-screening off common causes. Although critical of Cartwright's conclusions, our discussion of Cartwright's critique of the Causal Markov Condition owes a great deal to lengthy discussions with her, to material presented by her in a joint course taught with Woodward at Caltech in Spring 1997, and to her extensive criticisms of an earlier draft of this essay.

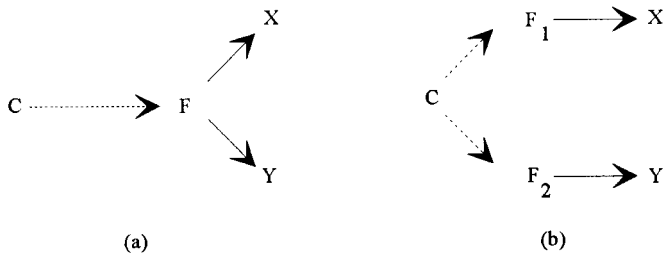

Fig. 9

specification of the state of the factory. Now the problem is that the variable *C* was not specified correctly or in sufficient detail. It should be replaced with the more informative variable *F*, and when one does so, screening off is restored.

Suppose, on the other hand, there are two firings—the firing F_1 of the factory C to produce chemical X and the firing F_2 of C to produce chemical Y. If F_1 and F_2 are more precise specifications of the state of the factory, it again seems that *C* is simply an insufficiently specified variable which should not be regarded as a cause of F_1 and F_2. Instead, the relevant specification of the state of the factory should replace *C* with F_1 and F_2. Suppose instead that F_1 and F_2 are distinct but correlated events that are not merely more precise specifications of the state of the factory. In this case the causal structure is as in figure 9b. Now there is no direct common cause of *X* and *Y* and the correlation between *X* and *Y* derives from the correlation between F_1 and F_2. But where does *that* correlation come from? Presumably F_1 and F_2 are not themselves related as cause and effect. If they also do not have a common cause, **CM1** is violated. If they have a screening-off common cause, one again does not have a counterexample to the Causal Markov Condition. If F_1 and F_2 have a non-screening-off common cause, then one has simply replicated the structure of the original example. In this last case, on pain of a vicious regress, one had better not invoke the 'firing' of their common cause to give some intuitive sense to how it is that F_1 and F_2 can be correlated even after one controls for *C*. We suspect that Cartwright would simply deny **CM1** and defend the possibility of non-causal correlations among purely chance occurrences.

In Cartwright's example, the correlation between *X* and *Y* arises because of their common cause *C*, yet that correlation does not disappear when one controls for the action of *C*, even though one ordinarily thinks that controlling for *C* mimics a situation in which they have no common cause at all. To see more clearly what is so strange about this, consider what happens if one intervenes to bring about *Y*. One possibility is that intervention is impossible—not merely infeasible for human beings or impossible because of special physical contingencies (see fn. 12), but that there is no way, even in principle, to disrupt the mechanism linking *C* to *Y* without also disrupting the mechanism linking *C* to

X—that is, the mechanisms are entangled, inseparable, or identical. Perhaps this is what Cartwright has in mind, because she goes on to write:

> The crucial point is that nature needs to set the joint probabilities; and the case that satisfies the 'common-cause condition' is just one very special way of doing so [. . .] For it is the case in which there is as little overlap as possible between *C*'s operation to produce *X* and its operation to produce *Y*. By contrast, my example of Cheap-but-Dirty's process was one of complete overlap—the cause operated to produce the intended product *X* just in case it also operated to produce the unwanted by-product *Y* ([1993], p. 117).

One way of interpreting this passage is that *X* and *Y* and the mechanisms linking them to *C* 'overlap' in the sense that they are not fully distinct—there is no possibility in principle of disrupting the mechanism connecting *C* to one of the effects without disrupting the link to the other. This would certainly make it intelligible how the correlation between Cheap-but-Dirty's production of chemical X and its production of chemical Y is accomplished, but it also means that the example is not correctly described as one in which *C* is a common cause of *X* and *Y*. Given the connection between causation and manipulation (which Cartwright might, of course, deny), for this description to apply, *C* must be linked to two distinct events *X* and *Y* via two distinct mechanisms. Recall the example used in Section **7**: if storms and barometer readings are both effects of atmospheric pressure, then there must in principle be a way to interfere with the mechanism connecting atmospheric pressure to the barometer reading without interfering with the mechanism linking the pressure to the occurrence of storms. If interventions that disrupt the *C*-*X* or *C*-*Y* mechanisms separately are not possible in Cartwright's example (as they are not possible in the EPR experiment—see below), then the case is not a counterexample to the Markov Condition.

In principle there seem to be two possible ways in which independent disruptability might fail. One is that *X* and *Y* are not distinct events. Consider again the example from Section **7** in which a radioactive atom emits an alpha particle. The event *E* consisting of the emission of two protons and the event *F* consisting of the emission of two neutrons are not related as cause and effect and are not independent conditional on the decay event. Yet it would be a mistake to view this as a counterexample to the Markov Condition, because there is really just a single effect, which is the emission of a particle consisting of two protons and two neutrons and a single mechanism which associated with the decay. A similar point applies to macroscopic indeterministic phenomena which are produced by a single causal process, if there are any.

A second possibility, which some would argue is illustrated by the EPR phenomenon, is that *X* and *Y* are distinct events, but they are not probabilistically dependent on one another in virtue of being cause and effect or

effects of a common cause. Instead they bear a different kind of non-causal (but non-accidental) relation to one another.[26] If this view of the measurement results as distinct and non-accidentally, but non-causally connected events is the correct interpretation of EPR, then—as noted in Section **8** in the argument from strong independence to the Causal Markov Condition—the Causal Markov Condition needs to be reformulated. Rather than saying that *X* causes *Y* if *Y* is probabilistically dependent on *X* conditional on all the direct causes of *X*, one should say that *X* causes *Y* if this condition obtains and it is also not the case that *X* is probabilistically dependent on *Y* conditional on all the direct causes of *Y*. Notice that this involves a revision of **CM1**. However, this is not the sort of case that Cartwright has in mind. She intends to present a case in which *C* is an indeterministic common *cause*, and does not argue that *X* and *Y* bear the sort of relationship to one another that measurements in the EPR experiments do. Moreover, Cartwright has argued elsewhere ([1989], Ch. 6; Chang and Cartwright [1993]) that the separated particles in the EPR experiment are (or at least may be) causally related. She does not make the analogous claim about *X* and *Y* in the chemicals example.

Since the EPR experiment has loomed large in recent discussions of the Markov Condition and since many philosophers regard the measurement outcomes in the EPR experiment as related as cause and effect or as effects of a common cause (or both), we should say a little more about why we believe that these interpretations of the EPR phenomenon are mistaken. Our two main reasons have to do with the impossibility of carrying out an intervention that sets or influences the value of one measurement result independently of the other or that interferes with one of the arrows or mechanisms characterizing the common cause structure independently of the other. First, as a number of writers have observed (e.g. Skyrms [1984]), once the orientation of the measurement apparatus is determined, there is no physically possible way to alter or fix the value of either measurement result. Indeed it is not even possible to intervene to change separately the probability distribution of either measurement. The experimenter can manipulate the orientation of the measurement apparatus, but neither she (nor any possible causal process) can, for a given orientation of the apparatus, manipulate the measurement outcome or its probability distribution. So it is impossible to produce a change in one measurement outcome by literally setting the other.

Second, as we noted in our discussion in Section **4**, for the notion of an intervention on *X* with respect to some set of variables **V** to be well defined, it is essential that there be a contrast between on the one hand intervening with respect to *X* only and changing *Y* if at all only through this change in *X* and on the other hand directly changing both *X* and *Y*, so that the change in *Y* is brought

[26] Woodward's view, to be defended below, is that the measurement results are not distinct events produced by distinct mechanisms.

about directly and not as a result of the change in *X*. The reason for this is that if the putative intervention process that changes *X* also directly changes *Y* but not through *X*, the direct effects of the process on *Y* and the effects if any of *X* on *Y* will be confounded. For example it is a familiar concern that, in an experimental investigation of the effect of a drug on recovery, the very act of administering the drug may exert a direct influence on recovery *via* a placebo effect that will be conflated with the pharmacological effects of the drug. The point of administering a placebo to a control group is to allow us to separate out these two effects and to measure the treatment effect alone.

The notion of an intervention with respect to one of the measurement events is not well-defined in the EPR phenomena, because the distinction between intervening with respect to *X* and acting directly on both *X* and *Y* cannot be drawn. The reason given for this within the standard interpretation is that the correlated particles are in a so-called non-separable or entangled state. In some way that is difficult to understand, the two particles constitute a single composite object, even though they may be at spacelike separation from each other. The measurement result on one wing is not really a distinct event from the result on the other wing but rather both comprise a 'single, indivisible non-local event' (Skyrms [1984], p. 255).[27] For this reason, it is wrong to think of the measurement process performed on one particle as directly affecting only the state of that particle and affecting the other particle if at all only through the change it produces in the first particle. Because the particles are in a non-separable state and because there is no well-defined notion of directly interfering with either of the particles without directly interfering with the other (or of disrupting the mechanism governing the behavior of one particle without disrupting the other) it is inappropriate to think of the EPR experiment as involving a causal structure in which there is a common cause of the measurement results and/or a direct causal link between the measurement results.[28]

This explanation for why it is that the notion of an intervention with respect to either of the measurement results is not well defined is controversial, although it is worth remarking that it is widely accepted among physicists. Moreover, because it allows one to avoid the conclusion that the measurement

[27] This is Skyrms' formulation of an idea he attributes to David Lewis.

[28] Another way of putting the point we are trying to make is to appeal to the distinction between conditioning and intervening (*cf.* Glymour and Meek [1994]). As we noted above, the probability of *Y* conditional on observing the value of *X* is in general different from the probability that *Y* would have if we were to intervene to set *X* to that value. The former involves conditioning, the latter intervening. The former has to do with the information observing *X* reveals about *Y*: the latter with the causal relationship, if any, between *X* and *Y*. In the EPR experiment the probability of one of the measurement results conditional on the other is different from the unconditional probability of the first result, but there is no operation of setting one of the measurement results that will change the probability of the other measurement result.

results are related as cause and effect, it allows one to avoid positing superluminal causal connections, symmetric or frame-dependent causal connections and other sorts of causal pathologies that are frequently appealed to in discussions of the EPR experiment.

A point which is less controversial and which we take to be illustrated by the preceding remarks is this: the application of the Markov Condition to any system, quantum mechanical or otherwise, requires a prior understanding of how that system should be segmented into distinct components or mechanisms. Those who have regarded the EPR experiment as a counterexample to **CM** have reasoned in the following way: since the measurement results are not related as cause and effect and are not independent conditional on any common cause, the Causal Markov Condition must be false. If the measurement results were distinct, the EPR experiment would be a genuine counterexample, but one needs a reason to accept the antecedent of this conditional and we have argued that there is none. If the measurement results are not distinct events or do not result from distinct mechanisms, then it must be mistaken to think of the EPR experiment as having a common cause structure or as having a structure in which the measurement results are related as cause and effect. This general point should survive any uncertainty the reader may feel about the right way to understand what is going on the EPR experiment.

We conclude that both in Cartwright's example and in the EPR experiment, if the causal relationships linking the supposed joint effects to their common cause are not independently disruptable, the system in question does not have a common cause structure and hence does not represent a counterexample to the Causal Markov Condition. We turn now to the alternative possibility, which is that it *is* possible to intervene and set the value of Y without disrupting the relationship between C and X. There are then three cases. Either 1. $Pr(X/C \& \text{set-}Y) = Pr(X/C \& Y)$, 2. $Pr(X/C \& \text{set-}Y) = Pr(X/C)$, or 3. $Pr(X/C \& \text{set-}Y)$ has some other value. Since case 3 raises no issues that do not show up in cases 1 and 2, we shall discuss only the first two cases.

Case 1: $Pr(X/C \& \text{set-}Y) = Pr(X/C \& Y)$. Since in Cartwright's example $Pr(X/C \& Y) > Pr(X/C)$, it follows that $Pr(X/C \& \text{set-}Y) > Pr(X/C)$ and given the connection between causation and manipulation, Y turns out to be a cause of X. Again the case does not have the structure intended by Cartwright, in which there is no causal connection between X and Y.[29] The case is closely analogous to the deterministic example, described in Section **6** above, in which an intervention on one of the joint effects A produced by a common cause C

[29] In Section **10** we introduce a probabilistic analogue to modularity which we call 'probabilistic modularity' (**PM**). This says that $Pr(Y/\mathbf{Parents}(Y)) = Pr(Y/(\mathbf{Parents}(Y) \& \text{set-}\mathbf{Z}))$ where **Z** is any set of variables distinct from Y and the values to which the variables in **Z** are set lie within the relevant range. **PM** is violated on this interpretation of Cartwright's example since $Pr(X/\mathbf{Parents}(X))$ is not the same as $Pr(X/(\mathbf{Parents}(X) \& \text{set-}Y))$.

alters the relation between *C* and the other effect *B*, in violation of modularity. Similarly, in the case under discussion, the probability with which *C* produces *X* varies depending on whether or not an intervention is carried out on *Y* and on what value *Y* is set to under the intervention, despite the fact that the mechanism linking *C* to *Y* is supposedly distinct from the mechanism linking *C* to *X* and that supposedly the only causal relationships are between *C* and *X* and *C* and *Y*. Just as in the deterministic case, it is reasonable to take this sort of behavior as an indication that the causal structure of the example is misspecified—either there is a causal link from *Y* to *X* or perhaps the causal relationships between *C* and *Y* and *C* and *X* are not really distinct (in which case the example collapses into the sort of analogue to the EPR phenomena already discussed above). On either interpretation, one does not have an example of a common cause structure that violates the Markov Condition.

Case 2: Pr(*X*/*C* & set-*Y*) = Pr(*X*/*C*). *Y* does not count as a cause of *X*, and it only carries information concerning the occurrence of *X* when it occurs as a consequence of *C*. Unlike case 1, one cannot manipulate Pr(*X*) or Pr(*X*/*C*) by intervening on *Y*. Informally the case might be described as one in which *C* produces *X* with the same probability both when it is allowed to produce Y and when the production of Y instead is suppressed or disrupted or due to some event distinct from *C*, and in which it also somehow happens that *C* causes *X* when and only when it causes *Y*. Nonetheless, there is no causal connection between *Y* and *X*—this is shown by the fact that one cannot manipulate Pr(*X*) (or Pr(*X*/*C*)) by intervening on *Y*. What we have is a kind of epistemic or informational relevance between *Y* and *X* (knowing the value of *Y* provides information that is relevant to the probability of *X* even after the value of *C* is taken into account) but no causal dependence.

The first thing to note is that the only real-life cases that appear to have this sort of structure involve coarse macro variables and are consistent with the satisfaction of the Causal Markov Condition at a more refined level of description. In particular, the possibility envisioned under (3) is *not* a macroscopic analogue of the EPR phenomenon since in the former but not in the latter it is possible to carry out an intervention that disrupts the correlation between the joint effects. This fact by itself provides reason to dismiss the example, absent some further argument for taking it seriously.

Nonetheless, it is worth trying to isolate more precisely what makes this example seem so at variance with ordinary causal intuitions and expectations. We begin by considering the nature of the dependence between *X* and *Y* that remains after one conditions on *C*. This is not supposed to be a sample coincidence but rather to be stable under repeated experiments in which *C* is allowed to produce *X* and *Y*—it is somehow the result of the causal or nomological structure of this system. Moreover, the conditional dependence

is also not the result of a primitive nomological connection between *X* and *Y*, because it is disrupted by interventions on *Y*. (Thus, for example, it is not like the conservation law that governs the results of the spin measurements in the EPR experiment.) How, then, is the coordination between *X* and *Y* effected? *Ex hypothesi*, it does not arise by accident and it is hard to understand the suggestion that it arises in some spontaneous but non-accidental way. (At the very least, spontaneous coordination in the behavior of *C* to produce *X* and *Y* would seem to violate the spirit of **CM1**.) The only remaining alternative appears to be that this dependence must be produced by means of some common state of *C* that affects both *X* and *Y*. But then shouldn't the dependence between *X* and *Y* disappear when one conditions on this common state? If one conditions on some candidate for this state and the dependence does not disappear, why isn't this just an indication that one has picked the wrong candidate?

A slightly different way of putting the point is this: The fact that the correlation between *X* and *Y* is only present when *C* causes both of them and disappears when one intervenes to set the value of *Y* means that what is informationally relevant to *X* is not the value of *Y per se* but rather how *Y* came to have that value. When *Y* is produced by an intervention, *Y* carries no information about whether *X* will occur, but when *Y* is produced by *C* it does carry such information. Thus knowing that *Y* is produced by *C* provides information about *X* over and above any information that is contained in the full specification of *C* and of *Y* itself. It is hard to understand how this is possible. If what is informationally relevant to *X* is the fact that *Y* has been caused by *C*, and this is not a fact represented in the state of *Y*, then it looks like it can only be a fact about some feature of (the causal structure or behavior of) *C* and hence a fact that one ought to take account of when one conditions on the common cause.[30] But when one does this, shouldn't one expect *X* and *Y* to be independent conditional on *C*? It is easy to understand how when *Y* is caused by *C*, it can provide information about *C* and how *Y* can thereby also provide information about the state of *X*, but how can *Y* tell us more about *X* than full knowledge of *C* can?

Whatever the intuitive force of these considerations, they are little more than a re-assertion of the naturalness of the Markov Condition and do not constitute an independent argument. But their intuitive force places the onus on Cartwright to explain why one should take seriously the possibility that there are real-life cases in which *X* and *Y* are joint effects of *C*, *X* and *Y* are not independent conditional on *C*, and yet an intervention with respect to *Y* leaves

[30] As Nancy Cartwright pointed out to us, the proposition that *C* causes *Y* will screen off *X* and *Y*, but *C* causing *Y* is a dubious candidate for an event type and is, in any case, not in the relevant sense distinct from *Y*. In addition *C* causing *Y* is not itself a cause of *X* or of anything else.

Pr(*X*/*C*) unchanged. After considering in the next section how the connection between causation and manipulability should be formulated in indeterministic circumstances, we will make an additional attempt to identify rigorously what it is about this case that is in tension with how people ordinarily think about causation.

Cartwright is certainly not the only writer who claims that the Markov Condition often fails in indeterministic contexts. Similar claims have been made by Salmon and Arntzenius. However, these writers appeal to examples, like those discussed in Section **2**, in which the wrong variables are measured or in which the variables are not distinct. Such examples can be dealt with along the lines described in Section **2**.

10 Indeterminism and the Causal Markov Condition

Although the last section answers Cartwright's critique, it does not address directly the question of whether the Causal Markov Condition should hold in indeterministic circumstances. Does the Causal Markov Condition break down when causation is not deterministic? Do any of the arguments for the Causal Markov Condition presented above in Sections **7** and **8** carry over to indeterministic circumstances? In this section, we shall make two arguments for the conclusion that **CM** holds in indeterministic circumstances. The first attempts to show that if **CM** holds in deterministic circumstances, then, given a plausible assumption about what indeterministic causation consists in, it must hold in indeterministic circumstances as well. The second retraces essentially the argument given above in Section **7** in a form that is appropriate for indeterministic causal relations.

One way to extend the arguments we have already made to indeterministic relations is to maintain that probabilistic causation is deterministic causation of probabilities. By this we mean that *X* is a probabilistic cause of *Y* if and only if *X* is a deterministic cause of the chance of *Y*, ch(*Y*), where this is identified with the objective probability of *Y*. In other words, the full set of probabilistic causes of *Y* are the full set of variables that deterministically cause the objective probability of *Y*. This view is defended at length by Hausman ([1998], Ch. 9) and has also been defended by David Papineau ([1989], p. 320) and Paul Humphreys ([1989]). It is also arguable that it is implicit in most probabilistic theories of causation. As such theories are usually understood, a probabilistic cause *X* is a variable that causally contributes to the chance of *Y* (or a variable, changes in the value of which change the chance of *Y*) and this contribution to or change in the chance of *Y* is thought of as determined by *X*. This is not the occasion for an extended defense of this view. We are interested instead in how one can use this view to argue that the Causal Markov Condition should hold when causation is not deterministic.

In this essay, as in the causal modeling literature generally, we have been concerned with causal relations among the variables in some specified set **V**. When one assumes that the relations among variables are deterministic, one assumes that there are other causal factors that are not in **V** that explain why the value of X is not a constant when the values of the parents of X are unchanging, or that explain why it is that Pr(X) is not equal to one for some particular value of X and zero for all other values. If causal relations are indeterministic, one does not have to assume that there are causes of X that are not contained in **V** in order to get non-degenerate probabilities, but, in many cases, it will reasonable to suppose that variation in the value of X conditional on its direct causes represented in **V** is due to causes that are not represented in **V** in addition to pure chance variation. Recall that **Parents**(X) was defined to be a subset of **V**. **Parents**(X) thus consists only of the direct causes of X that are represented in **V**. X may have other direct causes that are not represented in **V** and are not members of **Parents**(X).

As we have argued, to take probabilistic causation to be deterministic causation of probabilities is to say that ch(Y) is determined by the probabilistic causes of Y. To say this does not commit us to saying that ch(Y) is determined by **Parents**(Y). (In just the same way, in treating causation as a deterministic relation, one does not have to take Y to be determined by **Parents**(Y).) Since **Parents**(Y) is a subset of **V**, there may be direct causes of ch(Y) that are not contained in **Parents**(Y). So ch(Y) may be determined by all the direct causes of Y, even though it is not determined by **Parents**(Y).

To conclude that **CM** holds when causation is indeterministic if it holds when causation is deterministic, notice first that some changes in the values of X and Y may be deterministically caused, and some may involve chance. **CM1** says that unless X and Y are related as cause and effect or have some common cause, they will be probabilistically independent. This implies (plausibly) that to the extent that variations in X and Y are due purely to chance, they will not be correlated. Suppose then that X and Y bear probabilistic rather than deterministic causal relations to one another and that

(1) X is not a cause of Y.

Then from the claim that probabilistic causation is deterministic causation of probabilities, one can infer that

(2) X is not a cause of ch(Y).

If **CM** holds for deterministic relations, it follows that

(3) Pr(X/[**Parents**(X) & ch(Y)] = Pr(X/**Parents**(X)).)

Since (by **CM1**) Y will be independent of everything that ch(Y) is independent of, one can conclude for indeterministic variables X and Y, that if X does not

cause *Y*, then Pr(*X*/[**Parents**(*X*) & *Y*] = Pr(*X*/**Parents**(*X*)).) So if **CM** holds for deterministic causal relations, it holds for indeterministic relations as well.

One way to avoid this conclusion is to reject the assumption that probabilistic causation is deterministic causation of probabilities. However this assumption is rather deeply entrenched in the way philosophers have thought about probabilistic causation. Indeed, it appears to be presupposed in Cartwright's ([1979]) discussion of the relationship between probability and causation and by her well-known unanimity condition **CC**, in which all and only those factors that uniformly increase the chance of an event, conditional on the various possible combinations of other causal factors for the event, are taken to be probabilistic causes of the event. This brings out how deeply violations of the Markov Condition conflict with standard assumptions about probabilistic causation. Once these assumptions are accepted, one cannot consistently maintain that **CM** breaks down only in indeterministic circumstances.[31]

We do not know how much weight to place on this argument, because some critics of the Causal Markov Condition would find the construal of probabilistic causation as deterministic causation of probabilities as questionable as the Causal Markov Condition itself. Notice, however, that the argument only relies on the claim that if *X* is not a probabilistic cause of *Y*, then it is not a deterministic cause of ch(*Y*), not the full equivalence of probabilistic causation and deterministic causation of probabilities. Once this claim is accepted (and once one grants that ch(*Y*) can be an effect of other variables), this argument poses a serious challenge to those who maintain that **CM** breaks down only in indeterministic circumstances.

A second line of argument emerges if one considers what a manipulability view of causation implies in indeterministic circumstances. This argument closely follows the argument given above in Section **7**. The only differences are that the relation between manipulability and causation needs to be reformulated, and the assumption that each variable has an unrepresented cause is replaced by the assumption that either it has an unrepresented cause or its value varies spontaneously. The reformulation of the relation between manipulability and causation is basically just that instead of claiming that when *X* does not cause *Y*, the value of *Y* would remain unchanged given an intervention with respect to *X*, we shall claim that when *X* does not cause *Y*, then the probability distribution of *Y* will remain unchanged given an intervention with respect to

[31] One might try to avoid this conclusion by distinguishing between two kinds of probability—probability in the sense of chance and probability in the sense of informational relevance. Thus it might be said that, for example, the chance of *X* in Cartwright's chemical-factory story is entirely determined by *C* and is unaffected by the occurrence of *Y*, although the occurrence of *Y* is of course informationally relevant to *X*. Needless to say, this would require a fundamental revision of the framework assumed in standard probabilistic theories of causation. Moreover, as we shall see below, **CM1** commits one to denying that probabilities can be informationally relevant without also being causally relevant.

X. As before, we assume causal sufficiency, **CM1**, and that interventions can be treated as independent random variables.

The assumption that the probability distribution of *Y* will remain unchanged, given an intervention with respect to *X* plus the assumption that interventions can be treated as random variables can be restated formally as:

> **PM** (*Probabilistic modularity*) Pr(*Y*/**Parents**(*Y*)) = Pr(*Y*/**Parents**(*Y*) & set-**Z**) where **Z** is any set of variables distinct from *Y*, and the values to which the variables in **Z** are set lie within the relevant range.

PM expresses the idea that each conditional probability Pr(*Y*/**Parents**(*Y*)) corresponds to a distinct mechanism that expresses itself in a determinate conditional probability[32] and in addition that it should be possible to disrupt each such mechanism independently, without affecting the other mechanisms in the system. Thus it should be possible to intervene to set the values of the variables in any set of variables **Z** that does not contain *Y* and hence to disrupt the mechanism that connects those variables to their parents without disturbing Pr(*Y*/**Parents**(*Y*)). **PM** states roughly the same condition for indeterministic relations that **MOD** states for deterministic causal relations (and indeed it holds under pseudo-indeterminism as well as in genuinely indeterministic circumstances). **MOD** says every equation relating *Y* to its direct causes remains invariant with respect to interventions that set any subset **Z** whose members are distinct from *Y*. **PM** says that each conditional probability, Pr(*Y*/**Parents**(*Y*)) is invariant with respect to interventions that set **Z** (where *Y* is not in **Z**)—in other words, that *Y* and set-*Z* are probabilistically independent conditional on **Parents**(*Y*).

It is worth explicitly noting that it is not true in general that Pr(*Y*/**Parents**(*Y*)) & set-**Z**) = Pr(*Y*/**Parents**(*Y*) & **Z**) or that Pr(*Y*/**Parents**(*Y*)) = Pr(*Y*/**Parents**(*Y*) & **Z**) for all **Z**, even in systems that satisfy the Markov Condition. For example this equality will not hold for **Z** that contain descendants of *Y*—the values of such descendant variables may convey information about *Y* even if one

[32] We concede that it is at least logically possible for the condition that mechanisms express themselves in determinate conditional probabilities to fail. This would happen if intervening to change the value of **Parents**(*Y*) changed the probability of *Y*, so that **Parents**(*Y*) qualified as direct causes of *Y*, but the conditional probabilities Pr(*Y*/**Parents**(*Y*) differed in value on different occasions even for the same values of **Parents**(*Y*). As a concrete illustration, suppose that *C* is the only cause of *E*, that (1) Pr(*E*/*C*) > Pr(*E*/-*C*) but that the quantitative values of the conditional probabilities Pr(*E*/*C*) and Pr(*E*/-*C*) vary on different occasions of the occurrence of *C* and -*C*, although always in such a way that the inequality (1) is respected. Causes that operate in this way are called 'irregular' in Woodward ([1993]). Irregularity will sometimes arise because the specification of a cause 'averages' over heterogeneous micro-states. Cases of this sort may be dealt with by demanding a more precise specification of the cause, as in our treatment of Salmon's collision example. We know of no real-life examples of irregular causation that do not arise from heterogeneous micro-states and that are, so to speak, irreducibly irregular. At any event, it is widely assumed in philosophical discussion that when causation operates probabilistically it does not operate in an irreducibly irregular manner but rather in accord with determinate conditional probabilities, and our discussion above adopts this assumption.

conditions on **Parents**(*Y*). Again, the equality **PM** holds because of the special features possessed by interventions and the set-*Y* operation. By definition an intervention on *Y* renders *Y* independent of its causes in **V**. Hence even when *Z* contains descendants of *Y*, set-**Z** will be independent of *Y* even though **Z** itself will not be.

Since Pr(*Y*/**Parents**(*Y*) & set-**Parents**(*Y*)) = Pr(*Y*/set-**Parents**(*Y*)), **PM** implies:

PLI (*Probabilistic level invariance*) Pr(*Y*/**Parents**(*Y*)) = Pr(*Y*/set-**Parents**(*Y*)).

(Given the way that 'set-**Z**' is defined, Pr(*Y*/**Z** & set-**Z**) = Pr(*Y*/set-**Z**).) **PLI** expresses the idea that if **Parents**(*Y*) genuinely represents the causes of *Y*, then the conditional probability Pr(*Y*/**Parents**(*Y*)) should be invariant under interventions that change the value of any of the variables in **Parents**(*Y*). It corresponds, for indeterministic contexts, to the requirement that for (1) $Y = aX + U$ correctly to represent a causal relationship between *X* and *Y*, (1) must be invariant under interventions that set the value of *X* within some range. Note that it is not in general true that for any two variables *X* and *Y*, Pr(*X*/*Y*) = Pr(*X*/set-*Y*). For example, as we have noted, this equality will not hold if *X* and *Y* are the correlated effects of a common cause, since intervening on *Y* will make the correlation between *X* and *Y* disappear. Probabilistic level invariance claims that this equality will hold in the special case in which *Y* is the only direct cause of *X*.

As an intuitive illustration of **PM** and its implication **PLI**, suppose that *C* is the only parent—i.e. the only cause—of *X* and *Y*. Corresponding to the causal relationship or mechanism between *C* and *X* there will be a characteristic conditional probability Pr(*X*/*C*) and another conditional probability Pr(*Y*/*C*) will correspond to the causal relationship between *C* and *Y*. **PM** implies in this special case in which *C* is the sole cause of *X* and *Y* that intervening to increase the frequency of *C* from Pr(*C*) to Pr*(*C*) should not change the conditional probabilities Pr(*X*/*C*) or Pr(*Y*/*C*), but these should instead remain invariant under this change, so that the new distributions Pr*(*X*, *C*), Pr*(*Y*, *C*) are now given by Pr*(*X*, *C*) = Pr*(*C*). Pr(*X*/*C*) and Pr*(*Y*,*C*) = Pr*(*C*). Pr(*Y*/*C*).[33] Similarly, **PM** says that since the term Pr(*X*/*C*) describes the mechanism

[33] It is worth emphasizing that **PM** does *not* imply that when *X* causes *Y*, then Pr(*Y*/*X*) is invariant under all changes in the frequency of *X* . For example, if *X* and *Y* are also related as effects of a common cause, then an increase in the frequency of *Y* that results from an increased frequency of the common cause will have different consequences for the conditional probability than an increase in the frequency of *Y* that results from an intervention on *X*. In general, which quantities are invariant will depend on the causal structure of the situation with which one is concerned. For example, in the case discussed in the text, it is crucial that *C* is related to *X* and *Y* only as a cause of each and that *X* and *Y* bear no other causal relation to one another and have no other common cause. If instead *X* causes *Y* and *C* were the only common cause of *X* and *Y*, the Pr(*Y*/*X* & *C*) rather than Pr(*Y*/*C*) would be invariant under interventions on *C*. Kevin Hoover explores the possibility of using the invariance of conditional probabilities to intervention as a practical method for making causal inferences in macroeconomics; see Hoover ([forthcoming], Chs 8–11).

linking C to X and since this is distinct from the mechanism, described by Pr(Y/C), linking C to Y, it should be possible to interfere with the former mechanism without affecting the latter and *vice versa*. Thus, for example, one might set the value of X *via* an intervention to some value $X = x'$ so that the value of X is no longer influenced by C and Pr(X/C) is disrupted. The result of this should be that Pr(Y/C), as well as Pr(C) is unaffected. If instead we had said that Pr(C/XY) was invariant under interventions on Pr(X) or Pr(Y), this would be a way of encoding or representing a different set of claims about causal structure—that X and Y are causes of C, and that Pr(C/XY) represents a distinct mechanism.

It is important to distinguish the claim that a probability distribution can be factored in a certain way from **PM**. **PM** says instead that certain terms (and the variables or mechanisms corresponding to them) can be changed without disrupting other terms (variables and mechanisms). Any probability distribution can be factored in many different ways but at most one of these factorizations will consist of terms, any one of which may be changed independently of the other terms in accordance with **PM**. As an illustration, consider the simplest possible case. Given the definition of conditional probability and some elementary mathematics, the joint distribution Pr(A, B) can be written as either Pr(A). Pr(B/A) or as Pr(B). Pr(A/B). However, this fact does not show which if either of these factorizations consists of terms that may be changed independently of the others. If, for example, A is the sole cause of B, then from **PM** one should expect that Pr(B/A) will be stable under interventions that change Pr(A), but that Pr(A/B) will not be stable under interventions that change Pr(B). Similarly if B is the sole cause of A, then **PM** implies that Pr(A/B) but not Pr(B/A) should be stable under interventions on Pr(B), while if neither A nor B is a cause of the other, neither conditional probability should be invariant under interventions on either Pr(A) and Pr(B). Indeed it is easy to show that if Pr(B/A) and Pr(B/-A) are not zero and are invariant in this way to interventions that set A, then the inverse conditional probabilities Pr(A/B) and Pr(A/-B) cannot be invariant. By Bayes' theorem, Pr(A/B) = [Pr(A) . Pr(B/A)]/[Pr(A) . Pr(B/A) + Pr(-A) . Pr(B/-A)]. If Pr(B/A) and Pr(B/-A) are invariant under interventions that change Pr(A), then Pr(A/B) must change for different values of Pr(A). Alternatively, if it is not true that either A causes B or that B causes A, then an intervention that changes the value of either leaves the distribution of the other unchanged and hence disrupts the conditional probabilities Pr(A/B) and Pr(B/A) (for a similar argument, see Sober [1994], pp. 234–7; Hoover [forthcoming], Ch. 8).

Implicit in this way of looking at matters is the more general idea that there is more content to claims about the causal structure of some set of variables than that the joint distribution of those variables factors in a certain way. One way of capturing this additional content in both deterministic and indeterministic causal relations is by requiring that one of these factorizations satisfies

an invariance requirement such as **PM**. It is that factorization, if any, in which the terms are independent in the sense that each is invariant under changes in the other, that represents its causal structure. In a similar way, although different systems of equations may be compatible with the same probability distribution over the measured variables, it will be those equations (if any) that satisfy modularity that will capture the causal structure of the system under investigation.

We are now ready to argue for **CM** in much the same way as in §7. If *X* is distinct from *Y*, then from **PM** one can infer that Pr(*Y*/**Parents**(*Y*) & set-*X*) = Pr(*Y*/**Parents**(*Y*)). If, in addition, *X* does not cause *Y*, **PM** implies that for all ancestors *Z* of *Y* Pr(*Z*/**Parents**(*Z*) & set-*X*) = Pr(*Z*/**Parents**(*Z*)). So if *X* is distinct from *Y* and not a cause of *Y*, then an intervention with respect to *X* leaves invariant (a) the distribution of the exogenous variables that *Y* depends on, (b) the conditional and hence the marginal distribution of the ancestors of *Y* that have only exogenous variables as direct causes, and, hence (*c*) the marginal distribution of *Y* itself. So **PM** and the assumption that interventions can be treated as random variables imply the condition that in Section **7** we called **MOD***:

> **MOD*** For all distinct *X* and *Y* in **V**, if *X* does not cause *Y*, then Pr(*Y* & set-*X*) = Pr(*Y*). Pr(set-*X*).

Not only does **MOD*** hold in indeterministic as well as deterministic circumstances, so does (=). Consider the circumstances in which all the direct causes of *X* are observed to be unchanging. In those circumstances, (by assumption) the value of *X* varies either spontaneously or because of causes of *X* that are not represented in **V**. Given causal sufficiency, none of the unrepresented causes of *X* can cause any other variable by a path that does not go through **Parents**(*X*) and none of the causes of any other variables can cause *X* by a path that does not go through **Parents**(*X*). So in the circumstances in which **Parents**(*X*) is unchanging, either *X* varies spontaneously or because of causes that have no causal relation to any other variables except in virtue of causing *X*. In both cases, given **CM1**, changes in *X* count as interventions with respect to *Y*. Since changes in *X* conditional on **Parents**(*X*) count as interventions with respect to *Y*, changes in *X* conditional on **Parents**(*X*) must be independent of the same things that set-*X* is independent of. This is what the condition (=), which was proven in Section **7**, says:

> (=) Pr(*Y* & set-*X*) = Pr(*Y*) . Pr(set-*X*) if and only if Pr(*X*/(*Y* & **Parents**(*X*)) = Pr(*X*/**Parents**(*X*)).

Since, as we just argued, in the conditions in which the direct causes of *X* in **V** are unchanging, the spontaneous variation of *X* or the unrepresented causes of *X* satisfy the definition of an intervention, (=) must hold. **MOD*** says that the left-hand side of (=) holds when *X* does not cause *Y*, and (=) then implies that

the right-hand side must hold when *X* does not cause *Y*, which is exactly what **CM** says. So **PM** plus the assumptions needed to derive (=) implies and is implied by **CM**. In indeterministic as well as deterministic circumstances, the link between causation and manipulability implies and is implied by **CM**.

Given this argument why **CM** should hold in indeterministic circumstances, we can now say more precisely what is peculiar about the interpretation of Cheap-but-Dirty's production process in which $\Pr(X/C)$ and $\Pr(Y/C) = 0.8$, $\Pr(X/(Y \& C)) = 1$, and $\Pr(X/(\text{set-}Y \& C)) = \Pr(X/C)$. Since $\Pr(X/(\text{set-}Y \& C)) = \Pr(X/C)$ and *Y* does not cause *X*, the argument above for **MOD*** establishes that $\Pr(X \& \text{set-}Y) = \Pr(X) \cdot \Pr(\text{set-}Y)$. Since in addition $\Pr(X/(Y \& C)) \neq \Pr(X/C)$, one has a violation of (=). Since (=) follows from **CM1**, causal sufficiency, the assumption that interventions can be treated as random variables, and the existence of unrepresented causes or spontaneous variation, this interpretation of Cartwright's case must violate one of these conditions. In particular, in the circumstances in which Clean-but-Dirty's process is constantly in operation, *X* and *Y* have only a 0.8 chance of occurring, but whenever one occurs, so does the other. Since by assumption, there is no other common cause, it must be that the purely chance occurrences happen invariably to line up—in violation of **CM1**.

We earlier described this interpretation of Cartwright's example as a case in which informational and causal relevance fail to coincide—the value of *Y* (when caused by *C*) is informationally relevant to *X*, after conditioning on *C*, but there is no causal connection between *X* and *Y*. But given **CM1,** it is not possible for informational and causal relevance to come apart in this way. On reflection this conclusion ought to be unsurprising. Someone who thinks that causal and informational relevance can fail to coincide in cases like Cartwright's should also believe that they can come apart in the sort of case directly covered by **CM1**—that it is possible for *X* and *Y* to be correlated even though they are not related as cause and effect or as effects of a common cause. In other words, the connection between causal and informational relevance is built into **CM1** and once we accept this connection, **PM** and other plausible assumptions imply **CM**.

There is yet another way to diagnose what it is about Cartwright's case that is so at odds with common beliefs about causation. Return for a moment to the deterministic case, and consider again the equations:

(1) $Y = aX + U$

(2) $Z = bX + cY + V$

Modularity says that it should be possible to set the value of *Y* by an intervention and hence to disrupt (1) without disrupting (2). Indeed, for a given value of *Y*, (2) should hold regardless of whether that value is produced by setting *Y* or by letting *X* produce *Y* in accord with (1). A constraint on the causal

interpretation of (1)–(2) is that it should not matter to the value of *Z* whether a particular value of *Y* was fixed via an intervention or whether it came about *via* the mechanism described in equation 1. If this were not the case—if one found that the observed value of *Z* depended not just on the values of *X* and *Y*, but also on whether the value of *Y* was caused by the value of *X*, then one would take this to indicate that the system (1)–(2) is misspecified or incomplete in some way. *Z* would behave as though it were a function of *X* and of some more complex variable *Y+* that takes different values depending on how *Y* is produced.

It is natural to impose a condition in the probabilistic case that is analogous to this implication of modularity. If *X* is a probabilistic cause linked to *Y* by one mechanism and there is a distinct mechanism linking *X* and *Y* as probabilistic causes to *Z*, then the probability of *Z* given an intervention on *X* (which results in a new value of *Y* in accordance with equation (1) should be just the same as it would be if an intervention had set *Y* to that value. This aspect of modularity is not captured by **PM**, which concerns the independence of *X* and set-*Y* conditional on **Parents**(*X*).

How can we express this additional condition? We propose

> (**PM2**) When *X* and *Y* are distinct, Pr(*X*/set-[**Parents**(*Y*)] & *Y*) = Pr(*X*/set-[**Parents**(*Y*)] & set-*Y*).[34]

In other words, it should make no difference to the value of *X* whether we set *Y* or observe *Y*, once we set parents(*Y*). Moreover, this should be the case even if *Y* is a descendent or parent of *X*.

The interpretation of Cartwright's example in which Pr(*X*/*C*) = Pr(*X*/(*C* & set-*Y*)) violates **PM2**. The two joint effects *X* and *Y* have one parent *C* but Pr(*X*/(*Y* & set-*C*)) is not equal to P(*X*/(set-*C* & set-*Y*)). In violation of **PM2**, it makes a difference to the probability P(*X*/(set-*C* & *Y*)) whether the value of *Y* is produced exogenously *via* an intervention that disrupts its connection to *C* or endogenously through *C*, even when one sets *Y*'s parent (*C*). We think this captures at least part of the sense in which, without simply assuming the Markov Condition, one can say that the behavior in Cartwright's example seems to violate a widely accepted intuition about causality.

It is worth mentioning one other argument here. Consider, in place of **PM2** the following condition:

> **PM3** If *X* does not cause *Y*, then Pr(*X*/(**Parents**(*X*) & set-*Y*)) = Pr(*X*/(**Parents**(*X*) & *Y*)).

PM3 says that when *X* does not cause *Y*, then, conditional on **Parents**(*X*), set-*Y* and *Y* have the same probabilistic impact on *X*. Unless *X* causes *Y*, it should not matter to the probability distribution of *X* conditional on **Parents**(*X*), whether

[34] We owe this formulation to Judea Pearl.

one merely observes some value of Y or whether one sets it by intervention. The interpretation of Cartwright's example we have been considering obviously violates **PM3**, since $\Pr(X/(C \,\&\, Y)) > \Pr(X/(C \,\&\, \text{set-}Y))$. Since **PM3** is closely related to **CM**, we doubt that this point represents any additional strong criticism of Cartwright, but it is worth mentioning both because of the plausibility of **PM3** and because **PM** and **PM3** jointly provide a short and simple proof of **CM**. (Since **PM** implies that for Y distinct from X, $\Pr(X/(\mathbf{Parents}(X) \,\&\, \text{set-}Y)) = \Pr(X/\mathbf{Parents}(X))$, the proof is trivial.

11 Conclusion

In its simplest garb, our central argument for the Causal Markov Condition begins with a counterfactual sufficient condition for causation:

I. If one were to intervene with respect to X and the value of Y were to change, then X causes Y.

An intervention here is a direct cause of X that is not an effect of any variable in **V**, does not cause any variable in **V** by a path that does not go through X first, and has no cause that causes a variable in **V** by a path that does not go through the intervention. If one supposes:

1. that the value of X when it is set by intervention—that is, set-X—can be treated as a random variable,

then I implies:

II. If X does not cause Y, then Y is probabilistically independent of set-X.

If one assumes in addition

2. that causal independence implies probabilistic independence (**CM1**).
3. that there are no unrepresented common causes—that is, causal sufficiency—and
4. that variables have unrepresented causes,

then one can conclude:

III. Y is probabilistic independent of set-X if and only if Y and X are independent conditional on the direct causes of Y.

II and III then imply that if X does not cause Y, then Y and X are independent conditional on the direct causes of Y—which is essentially what the Causal Markov Condition maintains. In indeterministic circumstances, the general outline of the argument is just the same.

Although in Section **8** we explored other arguments in defense of the Causal Markov Condition, the one summarized above is at the core of this paper. Once

one accepts the connection between unconditional probabilistic dependencies and causation stated by **CM1**, the sufficient condition for causation proposed by those impressed by the link between causation and manipulability maps directly on to the Causal Markov Condition. That condition holds if causes are levers for moving their effects and probabilities are generated by processes in which causes that possess the structural properties of interventions are at work. For the reasons we have already canvassed, causal inferences based on the Markov Condition may readily go awry, but the condition itself follows from deep intrinsic features of causation.

The real problem with the Markov Condition is not, as Cartwright claims, that there are circumstances in which it breaks down, but rather that in order to apply the condition to a system of interest, one needs a great deal of knowledge—indeed much more knowledge than may be available. It is often far from obvious how to divide some system of interest into distinct causal mechanisms or networks of causal relationships. When the Markov Condition appears to lead one astray, one has evidence that either one has not succeeded in correctly segregating the system into distinct mechanisms, or one has failed to measure the full set of causally relevant variables with sufficient accuracy. The Markov Condition thus can play an important heuristic role in discovering causal structure, in the sense that its apparent failure suggests that one has left out causally relevant information. This may be its least contentious role, because in practice it is often immensely difficult to identify the mechanisms and to determine all the factors that are relevant to the operation of each mechanism, and these limitations to our knowledge often make it problematic to assume that the Markov Condition is satisfied for some system of interest. But to recognize that causal inference requires a great deal of knowledge is not to say that there are causal structures in which the Causal Markov Condition is violated.

Department of Philosophy
University of Wisconsin-Madison
Madison, WI 53706, USA
dhausman@facstaff.wisc.edu

Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125, USA
jfw@hss.caltech.edu

Acknowledgements

We are grateful to Nancy Cartwright, Ellery Eells, Malcolm Forster, Clark Glymour, Alan Hajek, Chris Hitchcock, Judea Pearl, Elliott Sober and anonymous referees for extremely generous and helpful comments.

References

Arntzenius, Frank [1993]: 'The Common Cause Principle', in *PSA 1992*, Vol. 2, pp. 227–37.

Bell, John [1964]: 'On the Einstein–Podolsky–Rosen Paradox', *Physics*, **1**, pp. 195–200.

Bell, John [1966]: 'On the Problem of Hidden Variables in Quantum Mechanics', *Reviews of Modern Physics*, **38**, pp. 447–52.

Bickel, Peter, Hammel, Eugene and O'Connell, J. William [1977]: 'Sex Bias in Graduate Admissions: Data from Berkeley', in William Fairley and Frederick Mosteller (*eds*), *Statistics and Public Policy*, Reading, MA: Addison-Wesley.

Cartwright, Nancy [1979]: 'Causal Laws and Effective Strategies', *Nous*, **13**; rpt. and cited from *How the Laws of Physics Lie*, Oxford: Oxford University Press, pp. 21–43.

Cartwright, Nancy [1989]: *Nature's Capacities and Their Measurement*, Oxford: Clarendon Press.

Cartwright, Nancy [1993]: 'Marks and Probabilities: Two Ways To Find Causal Structure', in F. Stadler (*ed.*), *Scientific Philosophy: Origins and Development*, Dordrecht: Kluwer.

Cartwright, Nancy [1997]: 'What Is a Causal Structure?' in McKim and Turner (*eds*), pp. 343–57.

Cartwright, Nancy and Jones, Martin [1991]: 'How to Hunt Quantum Causes', *Erkenntnis*, **35**, pp. 205–31.

Chang, Hasok and Cartwright, Nancy [1993]: 'Causality and Realism in the EPR Experiment', *Erkenntnis*, **38**, pp. 169–90.

Collingwood, R. G. [1940]: *An Essay on Metaphysics*, Oxford: Clarendon Press.

Cushing, James and McMullin, Ernan (*eds*) [1989]: *The Philosophical Consequences of Quantum Theory*, Notre Dame: Notre Dame University Press.

Fisher, Ronald [1951]: *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Fisher, Ronald [1959]: *Smoking: The Cancer Controversy*, Edinburgh: Oliver and Boyd.

Forster, Malcolm [1988]: 'Sober's Principle of Common Cause and the Problem of Comparing Incomplete Hypotheses', *Philosophy of Science*, **55**, pp. 538–59.

Gasking, Douglas [1955]: 'Causation and Recipes', *Mind*, **64**, pp. 479–87.

Glymour, Clark [1997]: 'A View of Recent Work on the Foundations of Causal Inference', in McKim and Turner (*eds*), pp. 201–48.

Glymour, Clark and Meek, Christopher [1994]: 'Conditioning and Intervening', *British Journal for the Philosophy of Science*, **45**, 1001–21.

Hausman, Daniel [1998]: *Causal Asymmetries*, New York: Cambridge University Press.

Hausman, Daniel [1999]: 'Lessons from Quantum Mechanics', *Synthese*.

Hoover, Kevin [forthcoming]: *Causality in Macroeconomics*, Cambridge: Cambridge University Press.

Humphreys, Paul [1989]: *The Chances of Explanation*, Princeton: Princeton University Press.

Kiiveri, H. and Speed, T. [1982]: 'Structural Analysis of Multivariate Data: A Review', in S. Leinhardt (*ed*) *Sociological Methodology*, San Francisco: Jossey-Bass.

Koster, J. [1996]: 'Markov Properties of Nonrecursive Causal Models', *The Annals of Statistics*, **21**, pp. 2148–77.

Lemmer, John [1996]: 'The Causal Markov Condition, Fact or Artifact', *SIGART Bulletin*, **7** (3), pp. 3–16.

McKim, Vaughn R. and Turner, Stephen P. (*eds*) [1997]: *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, Notre Dame: Notre Dame University Press.

Menzies, Peter and Price, Huw [1993]: 'Causation as a Secondary Quality', *British Journal for Philosophy of Science*, **44**, pp. 187–203.

Orcutt, Guy [1952]: 'Actions, Consequences and Causal Relations', *Review of Economics and Statistics*, **34**, pp. 305–13.

Papineau, David [1985]: 'Causal Asymmetry', *British Journal for the Philosophy of Science*, **30**, pp. 273–89.

Papineau, David [1989]: 'Pure, Mixed and Spurious Probabilities and the Significance for a Reductionist Theory of Causation,' in P. Kitcher and W. Salmon (*eds*), *Scientific Explanation, Minnesota Studies in the Philosophy of Science*, **13**, pp. 307–48.

Papineau, David [1991]: 'Correlations and Causes', *British Journal for the Philosophy of Science*, **42**, pp. 397–412.

Pearl, Judea [1995]: 'Causal Diagrams for Empirical Research', *Biometrika*, **82**, pp. 669–88.

Pearl, Judea [1998]: 'Graphs, Causality and Structural Models', *Sociological Methods and Research*, **27**, pp. 226–84.

Price, Huw [1991]: 'Agency and Probabilistic Causality', *British Journal for Philosophy of Science*, **42**, pp. 157–76.

Price, Huw [1992]: 'Agency and Causal Asymmetry', *Mind*, **101**, pp. 501–20.

Price, Huw [1993]: 'The Direction of Causation: Ramsey's Ultimate Contingency', in *PSA 1992*, pp. 253–67.

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

Salmon, Wesley [1985]: *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Simpson, C. [1951]: 'The Interpretation of Interaction in Contingency Tables', *Journal of the Royal Statistical Society*, Series B, **13**, pp. 238–41.

Skyrms, Brian [1984]: 'EPR: Lessons for Metaphysics', *Midwest Studies in Philosophy*, **9**, pp. 245–55.

Sober, Elliott [1987]: 'Discussion: Parsimony, Likelihood, and the Principle of the Common Cause', *Philosophy of Science*, **54**, pp. 465–9.

Sober, Elliott [1988]: 'The Principle of the Common Cause', in J. Fetzer (*ed.*), [1988], pp. 211–28.

Sober, Elliott [1994]: 'Temporally Oriented Laws,' in *From a Biological Point of View*, Cambridge: Cambridge University Press.

Sober, Elliott [1998]: 'Physicalism from a Probabilistic Point of View', typescript.

Spirtes, P. [1995]: 'Directed Cyclic Graphical Representation of Feedback Models', in Philippe Besnard and Steve Hanks (*eds*), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo: Morgan Kaufmann Publishers, Inc.

Spirtes, Peter, Glymour, Clark and Scheines, Richard [1993]: *Causation, Prediction, and Search*, New York: Springer-Verlag.

Teller, Paul [1989]: 'Relativity, Relational Holism, and the Bell Inequalities', in Cushing and McMullin (*eds*), pp. 208–23.

Van Fraassen, Bas [1982]: 'The Charybdis of Realism: Epistemological Implications of Bell's Inequality', *Synthese*, **52**, pp. 25–38.

Von Wright, G. H. [1971]: 'On the Logic and Epistemology of the Causal Relation', in Ernest Sosa and Michael Tooley (*eds*) [1987], *Causation and Conditionals*, Oxford: Oxford University Press, pp. 105–24.

Woodward, James [1993]: 'Capacities and Invariance', in J. Earman, A. Janis, G. Massey and N. Rescher (*eds*), *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolph Grünbaum*, Pittsburgh: University of Pittsburgh Press, pp. 283–328.

Woodward, James [1997]: 'Explanation, Invariance and Intervention', *Philosophy of Science*, **64**, pp. S26–41.

Woodward, James [forthcoming a]: 'Causal Interpretation in Systems of Equations', *Synthese*.

Woodward, James [forthcoming b]: 'Explanation and Invariance in the Special Sciences', *British Journal for the Philosophy of Science*, **51**.

Yule, G. [1903]: 'Notes on the Theory of Association of Attributes in Statistics', *Biometrika*, **2**, pp. 121–34.