CHAPTER 11

AGENCY AND INTERVENTIONIST THEORIES

JAMES F. WOODWARD

1. INTRODUCTION

Agency and interventionist theories of causation take as their point of departure a common-sense idea about the connection between causation and manipulation: causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. Very roughly, if *C* causes *E* then if *C* were to be manipulated in the right way, there would be an associated change in *E*. Conversely, if there would be a change in *E*, were the right sort of manipulation of *C* to occur, then *C* causes *E*. Accounts of causation in this vein have been defended by Collingwood (1940), Gasking (1955), von Wright (1971), Menzies and Price (1993), and Woodward (2003), among others. Similar ideas are defended by many social scientists and by some statisticians and theorists of experimental design. For example, in their influential text, *Quasi-Experimentation*, Cook and Campbell (1979: 36) write, 'The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. Causation implies that by varying one factor I can make another vary' (emphasis in original).

Writers who have developed computational models of causal inference within a Bayes net framework, including Judea Pearl (2000) and Peter Spirtes, Clark Glymour, and Richard Scheines ([1993] 2000), have also stressed the connection between causation and manipulation, although their interest is more in the discovery of causal relationships and the prediction of the effects of manipulations than in appealing to the notion of manipulation to provide a general account of causation.

A manipulationist approach to causation is appealing in part because it appears to provide a natural treatment of the difference between causal and purely correlational claims and why we should care about this difference. As Cartwright (1983) observes, there is a correlation between, on the one hand, whether (*P*) one purchases life insurance from TIAA-CREF (which furnishes life insurance to college teachers) or from some other commercial life insurance company, and, on the other hand, longevity (*L*): purchasers of TIAA-CREF insurance tend to live longer. However, it does not follow from this observation (and it is presumably false) that purchasing TIAA-CREF insurance is a way of manipulating longevity, or that such a purchase is a means or 'effective strategy' for increasing lifespan. A manipulationist account identifies the question of whether *P* causes *L* with the question of whether some appropriate manipulation of *P* will be associated with a change in *L*; we care about whether (1) *P* causes *L* or (2) *P* is merely correlated with *L* at least in part because (1) has very different implications from (2) for whether we can manipulate *L* by manipulating *P*. This is presumably one reason why approaches that stress the connection between manipulation and causation have been popular within experimentally oriented disciplines such as psychology and molecular biology and within disciplines that provide policy recommendations, such as economics.

Despite these appealing features, recent philosophical discussion has been largely unsympathetic to manipulationist theories. In particular, critics have claimed both that such theories are *circular* in an unilluminating way and that they are unduly *anthropocentric*. (See e.g. Hausman 1986; 1998). The charge of circularity arises because, on the face of things, the notion of manipulation looks like a causal notion; to manipulate something is to *cause* it to be in some state. How then can we appeal to the notion of manipulation to elucidate the notion of causation? The charge of anthropocentrism arises because at least on many common understandings of 'manipulation' this notion is tied to activities that human beings can carry out. There is thus a prima facie problem in extending a manipulation-based account to examples involving causal relationships in which there is no possibility of manipulation by human beings. For example, what sense can a manipulationist account give to causal claims about the effects of gravitational attraction between galaxies (in producing clumping and other large-scale structures), given that the manipulation of their causally relevant features (masses and distance from one another) is unlikely ever to be possible for humans?

It is useful to divide manipulation-based accounts into two broad categories. *Agency* theories stress the connection between causation and distinctively human agency (that is, actions and manipulations of a sort that might be carried out by human beings). Some defenders of agency theories (e.g. von Wright, Menzies, and Price) claim that one of the attractions of such theories is that they provide a way of avoiding the charge

of circularity. Their idea is that the concept or experience of human agency gives us independent access to (or purchase on) the notion of causation, because the notion of agency is either not a causal notion at all or at least does not presuppose all the features of a full-blooded notion of causation. However, as we shall see, reliance on a non-causal notion of agency does not seem to yield a normatively acceptable account of causation. Because of this, several more recent accounts (e.g. Pearl 2000; Woodward 2003) that focus on the connection between causation and manipulation have dropped any appeal to distinctively human agency and instead focus on a more abstract notion of manipulation, often called an *intervention*.

2. Background

In the remarks that follow, I focus first on agency theories and then on intervention-based (hereafter interventionist) theories. Before doing so, however, some additional stage setting is in order. I have been speaking very loosely of a general connection between causal claims and claims about what will happen under manipulations. Obviously, there are many different sorts of causal claims: there are claims that one general type or kind of factor causes another ('impacts of rocks cause bottles to break'), so-called token causal claims that the occurrence of some particular event caused another ('the impact of the rock thrown by Suzy at 3 p.m. on 9 Jan. 2005 caused this particular bottle to shatter') and so on. Within a broadly manipulationist framework, we should expect that different sorts of causal claims will be connected in different ways to claims about the outcomes of possible manipulations. Section 7 provides illustrations, but in earlier sections I will abstract away from the differences among different sorts of causal claims when these do not seem relevant to the points I wish to make.

A second point concerns the general form that a manipulationist account should take. It seems uncontroversial that the claim that C causes E can be true even if C is not actually manipulated—any account that suggests otherwise is a non-starter. This observation suggests that manipulationist accounts should be formulated as *counterfactual* claims connecting causal claims to claims about what would happen *if* certain manipulations *were* to be performed. This is what I did in sect. 1 above. On this construal, when we infer to a causal conclusion on the basis of evidence deriving just from passive observation (that is, the evidence is not generated by experimentation) we are attempting to infer what would happen were the appropriate experiment to be performed.

A third and deeper point concerns the notion of manipulation itself. Suppose that both lung cancer L and nicotine-stained fingers N are caused by smoking cigarettes S, but L does not cause N or vice-versa. In other words, the causal relationships are

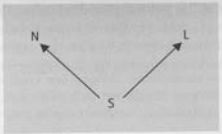

Fig. 11.1

such that they can be represented by the structure at Fig. 11.1, commonly called a directed graph, in which an arrow directed from X into Y means that X 'directly' causes Y and the absence of such an arrow means that X does not directly cause Y (for some additional brief remarks about directed graphs, see sect. 5).

Suppose that we decide to manipulate N by manipulating S (e.g. we force or otherwise induce some subjects to smoke and prevent others from doing so). N and L will be correlated under this manipulation of N. Nonetheless, by hypothesis, N does not cause L, in apparent contradiction to the connection between causation and manipulation on which manipulability theories rest.

This example shows that if the manipulationist approach is to be even remotely plausible, restrictions need to be imposed on what counts as an acceptable manipulation for purposes of the theory. In fact, such restrictions have a natural motivation in the theory of experimental design: everyone agrees that the manipulation just described is a badly designed experiment for the purposes of determining whether N causes L. Intuitively what we need to do to determine whether N causes L is (among other things) to manipulate N in a way that is suitably independent of (some relevant subset of) other possible causes of L such as S. More generally, we want our manipulation of N to be of such a character that if any change in L occurs, it can only occur because of the manipulation of N and not in any other way. One way of achieving this would be by means of a randomized experiment. Suppose that we have a population of both smokers and non-smokers and that depending on the output of some random device, we either assign subjects to a treatment group in which their fingers are caused to be nicotine stained or a control group in which fingers are not so stained. Because the assignment of N is based on the randomizing device, it will be causally and statistically independent of S. We would expect that under this sort of manipulation of N, N and L will no longer be correlated, indicating that N does not cause L.

Manipulations of a target variable X that have the right sort of special characteristics to figure in a well-designed experimental test of whether X causes some second variable Y are called *interventions* in the recent literature (cf. Spirtes, Glymour, and Scheines [1993] 2000; Pearl 2000; Woodward 2003). We will explore below some alternative proposals for how best to characterize this notion but it

should be apparent even at this point that understanding this notion will be central to the development of a plausible version of a manipulationist theory. One of the issues on which we will focus below is whether we can formulate a normatively acceptable notion of intervention just in terms of agency-related notions.

3. Agency Theories

By far the most detailed recent statement of an agency theory is due to Menzies and Price (1993; see also Price 1991). They propose that:

> An event A is a cause of a distinct event B just in case bringing about the occurrence of A is an effective means by which a free agent could bring about the occurrence of B.

This connection between agency and causation is used to motivate a version of a probabilistic theory of causation according to which 'A causes B' is identified with 'A raises the probability of B', where the probability in question is an 'agent probability':

> agent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized *ab initio*, as a free act of the agent concerned. Thus the agent probability that one should ascribe to B conditional on A . . . is the probability that B would hold were one to choose to realize A. (Menzies and Price 1993: 190)

In other words, the agent probability of B conditional on A is the probability of B conditional on the assumption that A has a particular sort of causal history—that A is realized by a free act. We can see what Menzies and Price intend by reference to the example above: their idea is when whether a subject has nicotine-stained fingers or not (N) is determined by a free act, any correlation between N and L should disappear—N does not raise the probability of L and hence does not cause it. Their claim is thus in effect that free acts function as interventions.

Menzies and Price hold that by appealing to the notion of agency, they can develop a non-circular, reductive analysis of causation. They claim that circularity is avoided because we have a grasp of the experience of agency that is independent of our grasp of the general notion of causation:

> The basic premise is that from an early age, we all have direct experience of acting as agents. That is, we have direct experience not merely of the Humean succession of events in the external world, but of a very special class of such successions: those in which the earlier event is an action of our own, performed in circumstances in which we both desire the later event, and believe that it is more probable given the act in question than it would be otherwise. To put it more simply, we all have direct personal experience of doing one thing and thence achieving another. . . . It is this common and commonplace experience that licenses what amounts to an ostensive definition of the notion of 'bringing about'. In other words, these cases provide direct non-linguistic acquaintance with the concept of bringing about an event; acquaintance which does not depend on prior acquisition of any causal notion. An agency theory thus escapes the threat of circularity. (ibid. 194-5)

Menzies and Price recognize that once the notion of causation has been tied in this way to our 'personal experience of doing one thing and hence achieving another (ibid. 194), a problem arises concerning causes for which there is no practical possibility of human manipulation. To use their own example, what can it mean to say that 'the 1989 San Francisco earthquake was caused by friction between continental plates' (ibid. 195) if no one has (or given the present state of human capabilities could have) the direct personal experience of bringing about an earthquake by manipulating these plates? Their response to this difficulty is complex, but the central idea is captured in these passages:

> we would argue that when an agent can bring about one event as a means to bringing about another, this is true in virtue of certain basic intrinsic features of the situation involved, these features being essentially non-causal though not necessarily physical in character. Accordingly, when we are presented with another situation involving a pair of events which resembles the given situation with respect to its intrinsic features, we infer that the pair of events is causally related even though they may not be manipulable. (ibid. 197)

> Clearly, the agency account, so weakened, allows us to make causal claims about unmanipulable events such as the claim that the 1989 San Francisco earthquake was caused by friction between continental plates. We can make such causal claims because we believe that there is another situation that models the circumstances surrounding the earthquake in the essential respects and does support a means-end relation between an appropriate pair of events. The paradigm example of such a situation would be that created by seismologists in their artificial simulations of the movement of continental plates. (ibid.)

4. Problems with Agency Theories

I will suggest below that these remarks embody an important psychological insight about the role of the subject's own action in causal learning. However, as it stands, the theory doesn't seem to yield a normatively correct treatment of the causal judgements we are justified in making. The crux of the difficulty is that while there is indeed reason to think that the extensions of the concepts 'event due to a free action' and 'event satisfying the conditions for an intervention' overlap to some

significant extent (cf. sect. 11), these concepts are very far from being exactly coextensive—an action may be free, at least in the senses normally recognized by philosophers, and yet fail to satisfy the conditions for an intervention. If, in such cases, we take correlations between X and Y that persist under free acts that realize X to show that X causes Y, we will often reach mistaken causal conclusions. Conversely, an action may not be free (or some process may occur which is not a human action at all) and yet the process may still satisfy the conditions for an intervention. If, when X is produced by such a process, X and Y remain correlated, we may conclude that X causes Y.

To illustrate, suppose that a human experimenter is faced with a common cause structure such as that of Fig. 11.1 and 'freely chooses' to manipulate N in a way that is correlated with the common cause S. (Perhaps the experimenter does this by observing whether individual subjects smoke and then manipulates N accordingly.) Note that there is nothing in the concept of a 'free action' as ordinarily understood, according to which the experimenter's actions are 'unfree' simply because they are influenced by or correlated with S. Then the occurrence of nicotine-stained fingers, when produced by such free actions, will raise the probability of lung cancer, even though the former does not cause the latter. Could we respond to this difficulty by making it part of the characterization of a 'free action' A that produces N that A is not correlated with other causes of N? A moment's thought shows that this additional condition is far too strong. As long as the free action itself has some causes (as it will on a non-libertarian account of free will) or as long as there are causally intermediate events between the free action A and N, these will be 'other causes' of N that are correlated with A. Moreover, the proposed condition is inadequate in other respects as well: for example, if the free act A itself directly causes both N and the other joint effect S (via a route that does not go through N), again the condition will not rule out the mistaken conclusion that N causes L.

More generally, it seems obvious that an adequate statement of the condition that Menzies and Price are looking for will need to make reference not just to N but to the putative effect L of N—as noted above, we need to exclude the possibility that this putative effect is the result of some causal pathway that does not go though N. It is unclear how the agency-related notions invoked by Menzies and Price might be used to make the distinctions needed to do this—instead, the needed distinctions appear to be overtly causal in character, and this undermines the reductionist aspect of Menzies and Price's project.

I turn next to a second set of problems for agency theories, having to do with the fact that our (i.e. human) notion of causation is such that we readily think of causal relationships as holding in contexts in which we have no 'personal experience' of agency. As the passages quoted above make clear, Menzies and Price think of this as primarily raising a problem having to do with causes that humans are unable to manipulate. In fact, however, there is a prior and in some respects more fundamental problem that arises even for readily manipulable causes. Consider the

contrast between two scenarios. In the first, you throw a rock which strikes a second rock, setting it in motion. Here you presumably have direct personal experience of your agency in setting the first rock in motion, but no such experience with respect to the second rock. In the second scenario, you observe one rock, set in motion by some natural cause such as the wind, strike a second rock with the result that it moves. Our concept of cause is such that we think that in both cases the impact of the first rock causes the second rock to move, but why exactly is it, on Menzies and Price's theory, that we are entitled to this judgement? After all, in the second scenario, I presumably have no experience of my own agency at all and even in the first, my experience of agency seems limited to producing the motion of the first rock. Within an agency theory, what justifies grouping the relationship between the movement of my hand and the motion of the first rock in the first scenario and the relationship between the movements of the first and second rocks under the common rubric of 'causation'?

It might seem that one possible answer (call this the 'projection hypothesis) is that in thinking of the movement of the first rock as a cause of the movement of the second, subjects consciously or unconsciously transfer their own experience of agency to the first rock—that is, they think of the first rock as (in some way) an agent which 'acts' on the second rock.[1] However, as it stands, this suggestion is not very satisfying. To begin with, we need more details about how this projection process works and why people engage in it, especially since the attributions in question, if taken literally, are so obviously mistaken—rocks are not really agents and so on. If the idea is that agents somehow find it useful to think about causal interactions involving rocks 'as if' they involved agents, even though such reasoning is not literally correct, we need some story about *why* this is useful and how this (mistaken) reasoning allows us capture the contrast between (what we normally think of) as true and false causal claims involving inanimate objects.[2] We also need to ask how the projection hypothesis might be tested and what evidence supports it. On one natural construal, the hypothesis predicts that the brain areas/psychological processes involved in the agent's own sense of agency and attribution of mental states to others are also centrally involved when agents attribute causal influence to inanimate objects. This prediction appears to be false.[3]

[1] Some of Menzies and Price's language suggests such a view and it is also advanced by other writers. For example Hausman (1998) suggests that this is the origin of our idea of causal necessity.

[2] That is, on the projection hypothesis there is an obvious sense in which all causal claims involving inanimate objects are in error. We thus need an account of why it is justifiable to think of some of these claims as 'true' while others are false.

[3] There is a great deal of evidence that specialized neural systems are involved in the sense of agency or ownership of action and in the attribution of mental states to others and that these are largely distinct from the systems involved in understanding the behaviour of inanimate objects. On one natural construal of the projection hypothesis, it predicts, on the contrary, that the same areas are active both in theory of mind/attribution of agency and in the attribution of causal influence to inanimate objects. My reading of the available evidence is that there is relatively little overlap. Most

Notice that this is *not* an issue about unmanipulable causes—it may be easy for me to manipulate the rocks in the second scenario if I choose to do so. The problem is rather that within Menzies and Price's framework we need some empirically grounded account of the processes of inference, analogical reasoning, imaginative extension, and so on (the processes that underlie projection) that lead from the fundamentally first-person experience of agency to a concept of cause that does not seem to require this experience for its application. It is interesting to note (cf. sect. 11) in this connection that several authors have suggested that many non-human animals, including other primates, have a grasp of an egocentric cause-like notion in the sense that they are capable of learning relationships between their own actions and the outcomes those actions produce, but that they fail to grasp that the very same relationships can hold between objects and events in the world, independently of their (or anyone's) actions or experience of agency, and that this has important consequences for the causal learning and understanding they are capable of. If so, such animals may possess (or at least behave as though they are guided by) an agency-based cause-like notion resembling the notion described by Menzies and Price, but not the concept of cause possessed by adult humans.

Quite apart from the projection problem just described there is also, as Menzies and Price recognize, a distinct problem having to do with the extension of their theory to causes for which there is no possibility of human manipulation. Menzies and Price attempt to resolve this problem by appealing to the idea that cases involving non-manipulable causes 'resemble' or can be modelled by cases involving manipulable causes and that, in virtue of this resemblance relationship, we can use our grasp of the latter to understand the former. It is of course crucial to this strategy that (as Menzies and Price claim) the resemblance in question be specified in non-causal terms. If, in specifying what it means for the movements of the continental plates to 'resemble' the artificial models that the seismologists are able to manipulate, we had to appeal to *causal* similarities between the two structures, we would no longer be explaining the content of claims about unmanipulable causes in terms of claims about manipulable causes. Instead, we would be relying on an unexplained notion of causal similarity between manipulable causes understood on the basis of agency and unmanipulable causes that must be understood in some other way. However, Menzies and Price provide no reason to believe that the needed resemblance relation can be specified non-causally and there is good reason

studies show that attribution of agency/mental states involves the insula, anterior cingulate cortex, superior temporal sulcus (STS), temporal poles, ventromedial prefrontal cortex, perhaps amygdala to some extent. Causal attribution to inanimate objects involves V5/MT, STS, and some parietal areas in the case of Michottean, launching-style causal interactions, and dorsolateral PFC in the case of more abstract causal learning. The projection hypothesis also seems to predict that subjects (e.g. high-functioning autistics) who have deficits in mental state attribution would also have difficulties with causal attribution involving inanimate objects. The available evidence seems to contradict this prediction.

to be sceptical of this claim. It is well known that small-scale models and simulations of naturally occurring phenomena that superficially resemble or mimic those phenomena may nonetheless fail to capture their causally relevant features because, for example, the models fail to 'scale up'—because causal processes that are not represented in the model become quite important at the length scales that characterize the naturally occurring phenomena.

5. Interventions

Our discussion so far has shown that if we wish to formulate a satisfactory statement of the connection between causal claims and the outcomes of ideal manipulations ('interventions'), we need to be precise about what constitutes an intervention. There have been a number of attempts to do this in the recent literature, including Spirtes, Glymour, and Scheines (1993; 2000), Hausman (1998), Pearl (2000), Woodward and Hitchcock (2003), Woodward (2003). All these writers focus on what is broadly the same idea but offer formulations that differ somewhat in detail, in part because they are animated by somewhat different theoretical purposes. I begin with Pearl (2000) who provides one of the most detailed recent attempts to think systematically about interventions and their significance for understanding causation.[4] Pearl follows a standard tradition in the econometrics and the causal modelling literature of using systems of equations to represent causal relationships. He also employs directed graphs for the same purpose. His work provides a striking illustration of the heuristic usefulness of a manipulationist framework in giving a causal interpretation for such representations. Since the notion of an intervention in Pearl's work is characterized in terms of equations and graphs, I begin with some brief remarks about these. For Pearl, a functional causal model is a system of equations $X_i = F(Pa_i, U_i)$ where X_i, Pa_i, and U_i are all variables (See sect. 6 below). Pa_i represents the direct causes or, as they are sometimes called, the 'parents' of X_i that are explicitly included in the model and U_i represents an error variable that summarizes the combined impact of all other variables that are causes of X_i. Pearl takes each equation to represent a distinct 'causal mechanism' which is understood to be 'autonomous' in the sense in which that notion is used in econometrics; this means roughly that it is possible to interfere with or disrupt each mechanism (and the corresponding equation)

[4] A broadly similar framework is developed in Spirtes, Glymour, and Scheines ([1993] 2000). My understanding is that these latter authors were the first to introduce the 'arrow breaking' or 'equation wipe out' conception of interventions into current discussion.

without disrupting any of the others. At least for the purposes of defining the notion of an intervention the notion of a causal mechanism or direct cause is taken as primitive and the notion of an intervention is defined in terms of it.

The simplest sort of intervention in which some variable X_i is set to some particular value x_i amounts, in Pearl's words (ibid. 70), to 'lifting X_i from the influence of the old functional mechanism $X_i = F_i(Pa_i, U_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms undisturbed' (I have altered the notation slightly). In other words, the intervention disrupts completely the relationship between X_i and its parents so that the value of X_i is determined entirely by the intervention. Furthermore, the intervention is 'surgical' in the sense that no other causal relationships in the system are changed. (This is sometimes described as the 'equation wipe out' conception of interventions.) Formally, this amounts to replacing the equation governing X_i with a new equation $X_i = x_i$, substituting for this new value of X_i in all the equations in which X_i occurs but leaving the other equations unaltered. It is assumed that the other variables in the system that change in value under this intervention will do so only if they are effects of X_i.

As an illustration, consider again the common cause structure from sect. 2, which may be represented by the equations:

(5.1) $N = F_1(S, U_1)$
(5.2) $L = F_2(S, U_2)$

The effect of an intervention on the variable N is represented by replacing equation (5.1) with a new equation (5.3) $N = n_i$, indicating that N has been set by the intervention to the value n_i and is no longer causally influenced by the variable S that was previously its direct cause. The other equation in the system is undisturbed by this alteration, so that the structure of the new system in which the intervention has occurred is represented by (5.3) and (5.2). Within a framework like Pearl's, causal relationships may also be represented by directed graphs. An arrow from a variable X_i *into* a second variable X_j (that is, an arrow with X_i at the tail and X_j at the head) means that X_i is a direct cause of X_j. Thus, as we have already noted, the common cause structure (5.1)—(5.2) can also be represented by Fig. 11.1.

Interventions also have a simple graphical representation: an intervention I on variable X_j 'breaks' or removes all other arrows directed into X_j and replaces these with a structure in which the only arrow into X_j is an arrow from I. All other arrows in the graph are left undisturbed. Thus the effect of an intervention on N in Fig. 11.1 is to replace this structure with that shown in Fig. 11.2, again representing that N is entirely under the control of the intervention variable and that other causal influences on N have been broken.

Pearl's talk of 'lifting' the variable intervened on from the influence of its (previous) direct causes may seem puzzling to philosophers who are accustomed

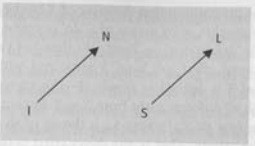

Fig. 11.2

to associate causal relationships with the instantiation of (presumably inviolable) 'laws of nature' but in fact the underlying idea is quite intuitive.[5] An intervention replaces a situation in which the variable X intervened on is sensitive to changes in the values of certain variables with a new situation in which the value of X is no longer sensitive to such changes but instead depends only on the value assigned to it by the intervention.[6] Many real-life experiments aim at (and succeed) in accomplishing this. For example, in an experiment to test the impact of a drug on recovery in which subjects are randomly assigned to a treatment group that receives the drug and a control group from which the drug is withheld, the ideal at which one aims is that who receives the drug should be determined entirely by the random assignment (the intervention), and not, as it presumably was previously, by such other factors as the subject's own decisions. As a matter of methodology, the reasons for employing experiments with this feature (when this is possible) is straightforward: by severing the relationship between the variable intervened on and its previous (or 'endogenous') direct causes, we eliminate some (although not all) possible sources of 'confounding'—for example, we ensure that those previous causes are not common causes of both the variable intervened on and the putative effect. We also ensure that the variable intervened on has whatever value the experimenter intends to give it, since this value is not affected by anything other than the intervention process.

[5] In addition to the considerations described below, the idea (and indeed the whole notion of an intervention) is broadly suggestive of Lewis's (1973) account of causation in terms of counterfactuals, the antecedents of which are made true by miracles. Readers interested in a more systematic exploration of these similarities (and some differences) are referred to Woodward (2003).

[6] As Woodward and Hitchcock (2003) observe one natural way of representing this is to think of the intervention variable as a 'switch' variable. For some range of values of the intervention variable I (values for which I takes an 'off' value) the variable intervened on, X, is a function of its parents and the value of I. For other values of I (values for which I is 'on'), the value of X is a function of the value of I alone.

Pearl represents the proposition that the value of X has been set by an intervention to some particular value, x, by means of a 'do' operator *do* $(X = x)$ or *do* (x). This allows for simple 'definitions' (as Pearl calls them) of various causal notions. For example, the 'causal effect' of X on Y associated with the 'realization' of a particular value x of X is defined as $P(y/do\ x)$—this represents the 'total effect' of $X = x$ on Y through all different paths from X to Y. By contrast, the 'direct effect' of $X = x$ on Y is $P(y/do\ x,\ do\ S_{xy})$ where S_{xy} is the set of all endogenous variables except X and Y in the system. That is, the direct effect is the distribution that Y takes under an intervention that sets $X = x$ and fixes by interventions the values of all other variables in the system—according to Pearl (ibid. 126), this represents the sensitivity of Y to changes in X alone.

We can further clarify the notion of an intervention by contrasting it with the more familiar notion of 'conditioning' (on a passively observed value of a variable) (cf. Meek and Glymour 1994). In the structure from sect. 2 in which S is a common cause of L and N, L and N are unconditionally dependent. That is, the probability of L conditional on N is different from the unconditional probability of L: $P(L/N) \neq P(L)$. However, under an intervention on N, which we may represent by conditioning on *do* N, L and *do* N will be independent: $P(L/do\ N) = P(L)$. Conditioning on an observed value of N is thus a fundamentally different operation from intervening on N. This is because when we condition on the observed values of N, we assume that whatever causal structure generates those values remains intact, while intervening on N alters the causal structure of this system. Thus if we observe the values of N and know that the value of L is generated by Fig. 11.1, this provides information about L; not so if the value of N is set by an intervention. Causal claims have to do with what will happen under interventions, although, given plausible assumptions, they also will have certain systematic connections to conditioning relationships.[7]

Pearl's characterization of the notion of an intervention seems ideally suited for the purposes for which he uses it. Basically these purposes are calculational or predictive rather than the more foundational ones that motivate philosophical accounts of causation. In particular, much of the focus of Pearl's discussion is on showing how to calculate the quantitative value of (to 'identify') causal effects and to predict the effects of interventions when we have qualitative information about causal structure and information about the probability distribution of the variables in the system of interest.

Arguably, however, Pearl's characterization is less well suited to the task of using the notion of an intervention to characterize what it is for a relationship to be causal.

[7] For a discussion of such connections, see Woodward (2003: ch. 7). One natural formulation is: If $P(Y/do\ X) = P(Y)$, then $P(X/parents(X)) = P(X/Parents\ (X).Y)$. This is one way of stating the so-called Causal Markov Condition.

One reason[8] for thinking this derives from Pearl's requirement that an intervention on X_i leave intact all other mechanisms besides the mechanism that previously determined the value of X_i. If, as Pearl apparently intends, we understand this to include the requirement that an intervention on X_i must leave intact the causal mechanism if any, that connects X_i to its possible effects Y, then an obvious worry about circularity arises, if we want to use the notion of an intervention to characterize what it is for X_i to cause Y. In part for this reason, Woodward and Hitchcock (2003) and Woodward (2003) explore a different way of characterizing the notion of an intervention that does not make reference to the relationship between the variable intervened on and its effects. For Woodward and Hitchcock (hereafter WH), in contrast to Pearl, an intervention I on a variable X is always defined with respect to a second variable Y (the intent being to use the notion of an intervention on X with respect to Y to characterize what it is for X to cause Y). An intervention I must meet the following requirements to count as an (WH) intervention:

(M1) I must be the only cause of X—that is, as with Pearl, the intervention must completely disrupt the causal relationship between X and its previous causes so that the value of X is set entirely by I,

(M2) I must not directly cause Y via a route that does not go through X.

(M3) I should not itself be caused by any cause that affects Y via a route that does not go through X,

(M4) I must be probabilistically independent of any cause of Y that does not lie on the causal route connecting X to Y.

Before considering how the WH notion of an intervention might be used to characterize various causal notions, let us note how both it and Pearl's notion differ from the agency-related notions to which Menzies and Price appeal. Neither Pearl's nor WH's notion involves human agency or activity—instead both define interventions in terms of causal and (in the case of WH) correlational relationships. A purely natural process, not involving human activity at any point, will count as an intervention as long as it has the right causal and correlational characteristics. This allows such intervention-based accounts to avoid at least some versions of the charge of anthropomorphism, although as we shall see (sect. 10) there still remain questions about their range of application. However, since the characterization of an intervention is overtly causal, it may seem that worries about 'circularity' in interventionist accounts become even more pressing—at the very least it is clear that we cannot appeal to Pearl's or WH's notion of an intervention to give a reductive account of what it is for a relationship to be causal. I will address this worry about circularity below, but I want first to explore in more detail how the notion of an intervention can be used in the characterization of causal relationships.

[8] Some additional reasons are described in Woodward (2003: ch. 3).

6. Causation and Interventions

Within a manipulationist framework, causes and effects must be manipulable, at least 'in principle'. This in turn suggests that we should think of causal relata as capable of varying or of being in a range of different possible states or conditions. It is thus natural to think of causal claims as having to do with relationships between *variables*, where the mark of variable is that it is capable of taking more than one value. I have already employed this convention in connection with many of the examples discussed above and the use of variables in the representation of causal relationships is standard practice in many areas of science. Our initial focus will be on type causation and on capturing a broad notion of causal relevance that corresponds to the idea of one factor being positively, negatively, or of mixed causal significance for another. The usual assumption in the philosophical literature that causation is a relationship between events or event types can be readily captured within this variable-based framework in terms of 'indicator' or two valued variables corresponding to the occurrence or non-occurrence of the events of interest. Thus we may express the causal claim that short circuits cause fires in terms of a relationship between two variables S and F, with S taking two possible values corresponding to the occurrence or non-occurrence of a short circuit, and F taking two possible values corresponding to the occurrence or non-occurrence of the fire.

Consider now the following proposals that give candidates for necessary and sufficient conditions for 'X causes Y' where X and Y are variables and 'causes' means (as explained above) 'is causally relevant to':

(SC) If (i) there are possible interventions that change the value of X such that (ii) under such interventions (and no others) X and Y are correlated, then X causes Y.

(NC) If X causes Y then (i) there are possible interventions that change the value of X such that (ii) under such interventions (and no other interventions) X and Y are correlated.

The causal notion captured by NC and SC is relatively weak and uninformative. It corresponds to the question 'is X causally relevant to Y at all', where this is interpreted as the question of whether there is there *some* change in the value of X which will change the value of Y or the probability distribution of Y. We are, of course, also interested in the elucidation of more precise causal claims having to do with the exact way in which X is causally relevant to Y—which from a manipulationist perspective has to do with exactly which changes in X will be associated with which changes in Y and under what conditions. As we shall see, the content of such claims may be captured within a manipulationist framework by extending the characterizations below in obvious ways.

SC says, in effect, that if it is possible to manipulate Y by intervening on X, then we may conclude that X causes Y, regardless of whether the relationship between X and Y lacks various other features that are sometimes regarded as necessary for causation. This is a highly non-trivial claim. It implies, for example, that 'double prevention' (Hall 2000) or 'causation by disconnection' (Schaffer 2000) involves genuine causal relationships (because these are relationships that support manipulation) even though the cause is not connected to its effect via a spatio-temporally continuous process and even though there is no transfer of energy and momentum from cause to effect. Similarly, if an 'action at a distance' version of Newtonian gravitational theory had turned out to be correct, this would be a theory that described genuine causal relationships, on an interventionist account of causation. This illustrates one respect in which an interventionist theory will reach very different conclusions about which relationships are causal from other competing theories.

What about NC? Consider the causal structure represented by means of the equations

$$(6.1)\quad Y = aX + cZ$$

$$(6.2)\quad Z = bX$$

and by the associated directed graph shown in Fig. 11.3.

If $a = -bc$, the direct causal influence of X on Y will be exactly cancelled out by the indirect influence of X on Y that is mediated through Z. If it is nonetheless correct to think that X (in some relevant sense) causes Y, then NC will be false, since there are no interventions on X alone that will change Y.[9]

This example shows that we need to distinguish between two notions of 'cause'.[10] Let us say that X is a *total cause* of Y if and only if it has a non-null total effect on

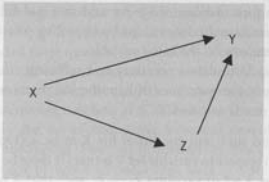

Fig. 11.3

[9] Spirtes, Glymour, and Scheines (1993) call this a failure of 'faithfulness'.

[10] This distinction is drawn in Pearl (2000)and Hitchcock (2001*b*), as well as Woodward (2003).

Y—that is, if and only if there is some intervention on *X* alone (and no other variables) such that for some value of other variables besides *X*, this intervention on *X* will change the value of *Y*. The notion of a total cause contrasts with the notion of a *contributing cause* which is intended to capture the intuitive idea of *X* influencing *Y* along some route or directed path even if, because of cancellation, *X* has no total effect on *Y*. While both SC and NC are plausible if 'cause' is interpreted as 'total cause' (where, it should be recalled the causal notion we are trying to capture is a broad notion of a causal relevance), NC is not plausible if 'cause' is interpreted as 'contributing cause' although SC remains plausible under this interpretation.

Can we capture the notion of a contributing cause within a interventionist framework? The strategy followed in Woodward (2003) is to first formulate a necessary and sufficient condition for *X* to be a *direct cause* of *Y* and then to use this formulation to arrive at a necessary and sufficient condition for *X* to be a contributing cause of *Y*. Woodward characterizes direct causation thus:

(DC) A necessary and sufficient condition for *X* to be a direct cause of *Y* with respect to some variable set V is that there be a possible intervention *I* on *X* that will change *Y* (or the probability distribution of *Y*) when all other variables in V besides *X* and *Y* are held fixed at some value by additional interventions that are independent of *I*.

Note that this characterization appeals to what will happen to *Y* under *combinations* of interventions, on both *X* and on other variables besides *X*. For example, *X* is a direct cause of *Y* in Figure 11.3 because, if we intervene to fix the value of *Z* and then intervene to change the value of *X* in a way that is independent of the intervention on *Z*, the value of *Y* will change. This contrasts with the characterization of total cause which appeals just to what will happen to *Y* under an intervention on *X* alone and no other variables. Note also that at this point we have moved well beyond earlier formulations of agency and manipulability theories that attempt to characterize causal relationships by appealing just to what will happen under *single* interventions on the cause variable.

Using DC we may formulate a necessary and sufficient condition, expressed in terms of claims about the outcomes of hypothetical interventions, for *X* to be a contributing, (type-level) cause of *Y*:

(M) A necessary and sufficient condition for *X* to be a (type-level) *contributing cause* of *Y* with respect to variable set V is that (i) there be a directed path from *X* to *Y*—that is, a set of variables $Z_1 \ldots Z_n$ such that *X* is a direct cause of Z_1 which is in turn a direct cause of Z_2 which is a direct cause of $\ldots Z_n$ which is a direct cause of *Y* and that (ii) there be some intervention on *X* that will change *Y* when all other variables in V that are not on this path are fixed at some value. If there is only one path *P* from *X* to *Y* or if the only alternative path from *X* to *Y* besides *P* contains no intermediate variables (i.e. is direct) then *X* is a contributing cause of *Y* along *P* as long as there is some intervention on *X* that will change the value of *Y*, for some values of the other variables in V.

(Motivation for this definition as well as illustrative examples are given in Woodward 2003.)

7. Other Causal Notions

We noted above that the causal claims characterized by NC, SC, and M are in one sense very weak—they refer only to there being some correlation between *X* and *Y* under some interventions on *X*. There are a variety of ways in both science and common sense that more detailed and specific causal information may be conveyed and these also have a natural interpretation within an interventionist framework. These include the formulation of quantitative relationships, represented by functions. For example, the relationship between the extension *X* and restoring force *F* exerted by a particular type of spring $F = -kX$ tells us exactly how for some range of interventions that alter *X*, *F* will change. Various qualitative locutions may be used for a similar purpose. For example, we often use 'causes' to express the idea that one factor *X* is a positive or promoting causal factor for another factor *Y* (rather than just being causally relevant to *Y*), as when we say that smoking causes lung cancer. Depending on the details of the case, such locutions may be interpreted within an interventionist framework as claims about various qualitative features of the functional form linking cause and effect—for example, that for variables representing smoking *S* and lung cancer *L* that take two possible values {present, absent}, an intervention that changes the value of *S* from absent to present increases the probability that the value of *L* will be present rather than absent. Alternatively, if *S* and *L* are understood as more quantitative variables—for example, if *S* is measured by average number of cigarettes consumed per week and *L* by probability of lung cancer, what might be intended by the claim that *S* is a promoting cause of *L* is that *L* is a monotonically increasing function of *S*, at least over much of its domain.

In still other cases, the use of *contrastive focus* will provide a natural way of conveying information about how manipulation of the cause will alter the effect. Suppose that the presence of potassium salts in a warehouse fire causes the flames to be purple. This claim might be interpreted as meaning something like: an intervention that changes whether potassium salts are present (when a fire occurs) will be associated with a change of colour in the flames. However (barring special circumstances) such an intervention will not change whether a fire occurs (rather than not occurring)—that is (as we might say),

the presence of the salts causes the fire to be purple rather than some other colour, but does not cause the fire to occur.

In general, then, as these examples illustrate, within an interventionist framework spelling out the content of detailed and specific causal claims will be a matter of specifying exactly which interventions on the cause variable will be associated with which changes in the effect variable and under which background circumstances. In this respect, an interventionist approach is more general than many other accounts in the philosophical literature that are predicated on the assumption that causal claims must assume some more specific canonical form—for example, accounts that assume that all causal claims must relate binary 'events' or must relate random variables with a well-defined joint probability distribution, as so-called probabilistic theories of causation do.

So far our focus has been on type causal claims. There are also several proposals in the literature that provide interventionist treatments of token or actual cause claims. For reasons of space, I will not attempt to describe these proposals in detail but will merely gesture at the basic idea, which is to appeal to what will happen to the effect under combinations of interventions that both affect the cause and fix certain other variables to specific values. It is worth noting that accounts taking this form are able to deal in a reasonably intuitive fashion with many of the standard counterexamples to certain other treatments of token causation. Suppose gunman 1 shoots (s_1) victim causing his death (d), where gunman 2 does not shoot but would have shot (s_2) also causing d, if s_1 had not occurred. If we fix the behaviour of gunman 2 at its actual value (he does not shoot), then an intervention that alters whether gunman 1 shoots will alter whether victim dies, thus identifying s_1 as the actual cause of d, despite the absence of counterfactual dependence (of the usual sort) between d and s_1.[11]

Although this appeal to combinations of interventions may strike some as artificial, in fact it maps onto standard experimental procedures in a natural way. Consider a case of genetic redundancy—gene complex G_1 is involved in causing phenotypic trait P but if G_1 is inactivated another gene complex G_2 (which is inactive when G_1 is active) will become active and will cause P. The geneticist may test for this possibility by first interfering with G_2 so that it is rendered permanently inactive and then intervening to vary G_1 and observing whether there is a corresponding change in P, and, second, intervening to render G_1 inactive and then, independently of this, turning G_2 on and off and observing whether there is a change in P. As this example illustrates, we may think of different complex causal structures in which there are multiple pathways, redundancy, cancellation, and so

[11] Proposals along these lines are given in Halpern and Pearl (2001), Hitchcock (2001*a*), Woodward (2003). For an improved proposal that addresses some shortcomings in these previous accounts, see Hitchcock (2007*b*).

on, as encoding different sets of claims about what will happen under various possible combinations of interventions.

8. The Problem of Circularity

Suppose, as argued above, that plausible versions of the manipulationist approach must appeal to a notion of intervention that is itself causal in character. How damaging is this to such accounts? The answer will depend in part on what we take the legitimate goals and aspirations of an account of causation to be. Many philosophers have supposed that an acceptable theory of causation must be reductionist—that is, it should explain causal notions in terms of concepts that are not themselves causal and that meet certain agreed-upon criteria for intelligibility and testability. Typically these criteria are broadly empiricist—thus it is assumed that the reduction will involve such non-causal concepts as regularity, spatio-temporal contiguity, and the like. It is obvious that an intervention-based account of causation will not be reductive in this sense. On the other hand, there is no generally accepted reductive account of causation and a number of writers (e.g. Cartwright 1983) have argued that there are good reasons for supposing that no such account is possible. In addition, there are many examples from both science and common sense of interrelated families of concepts that do not seem reducible to concepts that lie outside such families and yet seem nonetheless to satisfy reasonable standards of intelligibility and testability—'probability' is a standard illustration. This suggests that we can often make real progress in elucidating some concept of interest by showing how it connects up with other concepts and how claims involving it can be tested even if we cannot provide a non-circular reduction. Advocates of interventionist theories can claim that a similar point holds for 'cause'—even if we cannot reduce the various versions of this concept to something else, we can elucidate its content by showing how it connects up with other causally based concepts such as intervention, and how claims involving it can be tested both experimentally and otherwise.

It is thus worth asking those who require that an acceptable account of causation must be reductionist just what the motivation or rationale for this requirement is. If we think of various theoretical concepts (such as 'electron') that figure in scientific theories, the idea that they must be definable in terms of or reducible to some other set of supposedly more empirically legitimate concepts (e.g. 'observable concepts') was abandoned a long time ago as indefensible 'concept empiricism'. On the face of things, those who contend that any acceptable account of causation must be reductionist are urging the analogue of concept empiricism for

'causation'. They need to explain more clearly than they have hitherto why this demand is in order in the case of 'causation' even though it has been given up in other cases.[12]

A related point is that it is simply a mistake to suppose that because manipulability approaches are non-reductive, they are trivial, tautological, or lacking in interesting content. For one thing, as the discussion in sect. 6 shows, manipulability accounts can *conflict* with other accounts of causation, leading to different causal judgements in particular cases (e.g. in cases in which there is action at a distance, double prevention, etc.). In addition, the issue of how best to characterize the notion of an intervention and how to connect it to causal claims in such a way as to avoid obvious counterexamples (such as those discussed in sect. 2) is a highly non-trivial matter. Moreover, as we have noted, there are a number of different causal concepts—total causation, direct causation, token causation, and so on. Even within a broadly interventionist framework, we face a number of non-trivial choices about how such concepts connect to each other, and to the notion of intervention. An interventionist framework can thus be very useful in exhibiting the differences and interconnections among different causal concepts even if it fails to be reductive. Note also that although defenders of an interventionist account are committed to the idea that such an account can be worthwhile and illuminating without being reductionist, there is nothing in the interventionist approach per se that excludes the possibility that such an account might be supplemented or complemented by some other approach that does offer a reduction of key interventionist concepts such as 'intervention'. Interventionists may be sceptical that such an account will ever be forthcoming, but they need not reject its possibility a priori.

There is yet another observation that bears on the issue of circularity. Note that although the WH characterization of an intervention I on X with respect to Y does make use of causal information, this is *not* information about the existence or non-existence of a causal relationship *between X and Y*. Instead the information concerns the causal relationship between I and X, between I and other causes of Y besides X, and so on. In other words, the WH characterization connects information about *other* causal relationships besides the $X \rightarrow Y$ relationship and

[12] Another related argument concerning reduction is raised by the common claim that counterfactuals (including interventionist counterfactuals) as well as causal claims cannot be 'barely true' but instead require non-modal truthmakers of some kind—laws of nature being standard candidates for this role. If so, the argument continues, we should we appeal directly to this non-modal notion of law rather than to interventionist counterfactuals to explicate causal claims. Space precludes detailed discussion but note that this argument assumes that it is possible to explicate the notion of a law without appeal to interventionist counterfactuals and then use the former to ground the latter. Woodward (2003) denies this, arguing that invariance is the key feature of laws, where invariance is a counterfactual notion that has to do with stability under possible changes and is not reducible to non-modal notions.

correlational information to a claim about what must be true for X to cause Y. The characterization of an intervention on X with respect to Y is thus not viciously circular in the sense that it presupposes the very thing we are trying to elucidate—whether there is a causal relationship from X to Y. Regardless of whether the WH characterization or other characterizations found in the literature are fully adequate, there is a very compelling reason for thinking that *some* non-viciously circular characterization must be possible: we do after all learn about causal relationships by performing relatively black-box experiments, and it is not easy to see how this is possible unless we can sometimes recognize whether there has been an intervention on X with respect to Y without presupposing an answer to the question of whether X causes Y.

9. In What Sense Must Interventions be Possible?

SC, NC, and M refer to 'possible interventions'. There is a range of ways this phrase might be interpreted, corresponding to more or less strict notions of possibility. Note first that because the notion of an intervention has been given a non-anthropomorphic characterization, there is nothing in the versions of interventionist theory formulated above that motivates restriction of 'possible interventions' to interventions that are within the present practical or technological powers of human beings. As long as we can sensibly entertain counterfactuals about what would happen to Y if some natural process meeting the conditions for an intervention were to occur on X, we can apply the interventionist theory. This will certainly include some large range of cases in which the interventions in question are of such a character that they cannot at present be carried out by humans. Matters become stickier, however, when we consider circumstances in which the relevant notion of an intervention is physically or nomologically impossible. Consider (cf. Woodward 2003) the claim that the gravitational attraction F_m of the moon causes the behaviour of the tides. It is arguable that not only is it not technologically possible for humans to change the value of F_m (e.g. by changing the position of the moon) but that any physically possible process that might accomplish this would violate the conditions for an intervention, roughly because the process would not be sufficiently 'surgical'. For example, if nature were to change the position of the moon by introducing a new gravitating body in its vicinity, this body would exert an independent gravitational influence of the tides in violation of the requirement in the characterization of an intervention. Woodward (2003) responds by suggesting that in at least some cases of this sort (including the one under discussion) we

are in possession of a well-confirmed theory (Newtonian gravitational theory) that tells us what would happen in the (arguably) contra-nomic circumstances in which the moon occupies a different position as a result of an intervention and this is sufficient for the evaluation of the appropriate counterfactuals. Others may think, however, that at this point we have moved beyond the most natural range of application of the manipulationist theory. I will return to this issue in sect. 11.

Finally, consider cases in which interventions may be impossible or ill-defined for conceptual or metaphysical reasons. For example, some hold that there is no well-defined process of changing an animal of one biological species into a member of some other biological species—if so, claims like 'N being a tiger causes N to run fast' will lack an interventionist interpretation. Several prominent statisticians who favour manipulationist accounts of causation (e.g. Rubin 1986; Holland 1986) have argued on similar grounds that claims attributing causal efficacy to race and gender are not meaningful. Others (e.g. Glymour 2004) agree that such candidates for causes are unmanipulable in principle but are also antecedently convinced that causal claims involving them are meaningful and hence take such examples to reflect an important limitation on the scope of manipulationist accounts. However, as Woodward (2003) argues, causal claims involving unmanipulable causes often *are* unclear in meaning or ambiguous and that their meaning can often be clarified by replacing them with similar but related claims involving manipulable causes. This is just what one would expect if a manipulationist account of causation is correct.

10. Scope of Interventionist Accounts

In sect. 9, we observed that although it is natural to formulate an interventionist account in terms of counterfactuals about what would happen under possible interventions, it is arguable that as we make the relevant notion of 'possible intervention' more and more permissive, so that it includes contra-nomic possibilities and so on, we reach a point at which this notion and the counterfactuals in which it figures become so unclear that we can no longer use them to illuminate or provide any independent purchase on causal claims. It is an interesting and unresolved question whether the point at which this happens is also the point at which the associated causal claims no longer strike us as clear or useful, which is what one would expect if interventionism is a complete account of causation.

This issue arises in a particularly forceful way when we attempt to apply such accounts to fundamental physical theories understood as applying to the whole universe. Consider this claim:

(10.1) The state S_t of the entire universe at time t causes the state S_{t+d} of the entire universe at time $t + d$.

where S_t and S_{t+d} are specifications in terms of some fundamental physical theory.

On an interventionist construal, (10.1) is unpacked as a claim to the effect that under some possible intervention that changes S_t, there would be an associated change in S_{t+d}. The obvious worry is that it is unclear what would be involved in such an intervention and unclear how to assess what would happen if it were to occur, given the stipulation that S_t is a specification of the entire state of the universe. How, for example, might such an intervention be realized, given that there is nothing left over in addition to S_t to realize it with?

Commenting on an example like this, Pearl (2000: 350) writes, 'If you wish to include the whole universe in the model, causality disappears because interventions disappear—the manipulator and the manipulated lose their distinction.' Whether or not Pearl is right about this, it seems uncontroversial that it is far from straightforward how to interpret the interventionist counterfactual associated with (10.1). The interventionist account seems to apply most naturally and straightforwardly to what Pearl calls 'small worlds'—cases in which the system of causal relationships in which we are interested is located in a larger environment which serves as a potential source of outside or 'exogenous' interventions. The systems of causal relationships that figure in common-sense causal reasoning and in the biological, psychological, and social sciences all have this character but fundamental physical theories do not, at least when their domain is taken to be the entire universe.

There are several possible reactions to these observations. One is that causal claims in fundamental physics such as (10.1) are literally true and that it is an important limitation in interventionist theories that they have difficulty elucidating such claims. A second, diametrically opposed reaction, which I take to be Pearl's, is that causal concepts do not apply, at least in any straightforward way, to some or many fundamental physics contexts and that is a virtue of the interventionist account that it helps us to understand why this is so. This second suggestion may seem deeply shocking and counterintuitive to philosophers who believe that all causal claims must be 'grounded' in ('made true by') fundamental physical laws. In fact, however, the view that fundamental physics is not a hospitable context for causation and that attempts to interpret fundamental physical theories in causal terms are unmotivated, misguided, and likely to breed confusion is probably the dominant, although by no means universal, view among contemporary philosophers of physics.[13] According to some writers (Hitchcock 2007*a*; Woodward 2007*a*),

[13] Norton (2007) provides one statement of the dominant view. In contrast, Frisch (2002) argues for an interpretation of classical electrodynamics that relies on 'rich causal assumptions', understood in explicitly interventionist terms.

we should take seriously the possibility that causal reasoning and understanding apply most naturally to small world systems of medium-sized physical objects of the sort studied in the various special sciences and look for an account of causation, such as the interventionist account, that explains this fact. The question of the scope of interventionist theories and causal claims in general is thus an important and at present unresolved issue.[14]

11. Agency and Interventionist Accounts in Psychological Perspective

So far our focus has been on the evaluation of agency and interventionist accounts as normative theories of causal inference and judgement. However, both theories also can be interpreted as suggesting various descriptive claims about the empirical psychology of causal learning and judgement among both humans and non-humans. For example, Menzies and Price's version of the agency theory is, inter alia, a theory about the origins of causal concepts in humans. There are very rich and rapidly growing literatures within (human) cognitive and developmental psychology, primatology, and animal learning that bear on these empirical claims. Because of space constraints, I can only gesture at a few themes within this literature.[15] First, a number of writers have noted the close similarity between instrumental or operant (as opposed to classical) conditioning and causal learning when viewed from a manipulationist perspective. In instrumental conditioning, what is learned is an association between some behaviour produced by the subject and an outcome, as when rats learn an association between pressing a lever and the provision of a food pellet. There are striking, if incomplete, parallels between instrumental conditioning in non-human animals and causal learning and judgement in humans—for example, human judgements of causal strength are subject to discounting effects when alternative causes are present, and exhibit backward blocking, just as instrumental conditioning does. In general, humans behave as though estimates of the instrumental efficacy of their action tracks causal efficacy, which is what one would expect on a manipulationist theory of causation.

[14] Yet another possible position would be to hold that causal claims play a central role in fundamental physics but that for the reasons described above, interventionist accounts fail to capture this role. However interventionist accounts are successful at elucidating causal claims in the special sciences. On this view, causal claims in fundamental physics would need to be given some other, non-interventionist elucidation.

[15] For more detailed discussion, see Woodward (2007*b*).

A second theme concerns the role of a subject's own actions in facilitating causal learning. There is a great deal of evidence that the ability to intervene or manipulate facilitates causal learning in both adults and small children in comparison with passive observation. Both groups are able to reason to normatively correct causal conclusions in cases involving both a single intervention and combinations of interventions, and to distinguish between intervening and conditioning in normatively appropriate ways. These observations suggest that interventionist accounts capture something that is 'psychologically real' in human causal judgement. Although, for reasons explained above, it is dubious that the human concept of cause is derived just from the experience of agency, it is a natural interpretation of the experimental evidence that this experience plays an important role in learning particular causal relationships. Woodward (2007*b*) suggests that humans including infants have (1) a default tendency to behave or reason as though they take their own voluntary actions to have the characteristics of interventions and (2) associated with this a strong tendency to take changes that temporally follow those interventions with a relatively short delay as caused by them. We see evidence for this tendency in the existence of well-known causal illusions in which we experience salient changes that follow our voluntary actions as caused by them. Of course, as noted above, by no means all voluntary actions qualify as interventions, but nonetheless it may be that people have a defeasible tendency to assume this and that this tendency facilitates causal learning, especially in young children. Thus, even if agency-based accounts do not yield a normatively adequate account of causation, they may have a great deal of value as accounts of the acquisition of causal knowledge.

A third issue concerns the relationship between the sorts of capacities/causal understanding that are manifested in the ability to intervene and manipulate and other capacities that are often associated with causal understanding or possession of a concept of causation. For example, a number of psychologists (e.g. Leslie and Keeble 1987) and philosophers (Prinz 2002) claim that the visual responses (as measured by looking time) of infants to so-called launching phenomena (mechanical collisions of the sort studied by Michotte involving the perception of causation) and to object permanence tasks show that even very young children possess a concept of causation and are capable of causal reasoning. Although the evidence is controversial, it is widely believed that there is a dissociation between these abilities and success in related manipulation tasks—in human infants, sensitivity to launching and object permanence emerges before the ability to use such information to manipulate and a similar dissociation appears to be present even in non-human adult primates. From an interventionist perspective, this raises the very interesting question of the relationship between these two sets of abilities—is it appropriate to think of the abilities associated with sensitivity to launching and to object recognition as manifesting causal understanding at all, if these do not transfer to capacities for manipulation and control? How good is the evidence for dissociation/non-

transference? What are the processes by which, at least in older children, such transfer is achieved, and what implications does this have for how we should think about what it means to possess a concept of causation?

Finally, it is a striking fact that other primates, including chimps, are greatly inferior to humans, including small children, in causal understanding, particularly in connection with tool use and object manipulation. The source and character of these deficits is a matter of ongoing controversy but one possibility, suggested by the primatologists Call and Tomasello (1997), and by Woodward (2007*b*), is that non-human primates possess only an egocentric kind of causal (or cause-like) understanding—they readily learn about the instrumental consequences of their own actions but fail to appreciate that the very same causal relationships can be present both between their own actions and their effects, between the actions of conspecifics and the outcomes of their actions and between events occurring in nature that do not involve the actions of other creatures at all. One indication of this limitation is the apparent difficulty that non-human primates have in transferring information across these different contexts: for example, they don't seem very good at learning what the consequences would be if they were to perform various actions by observing the consequences of the actions of others or by passive observation of causal relationships as they occur in nature. (This in turn is connected to the well-known limitations of non-human primates in tasks involving imitation.) By contrast, even very small human children are adept at such learning. These observations fit naturally with the picture of the relationship between causal understanding and the outcome of interventions suggested in sect. 4—the human notion of causation transcends the actor's own experience of agency and is rather the notion of a relationship that has to do with what would happen under an abstract notion of intervention that can be realized by other actors or by nature.

Further Reading

Menzies and Price (1993) is the most philosophically sophisticated recent defence of an agency theory. Hausman (1998) is a very detailed and systematic exploration of the interrelations between manipulation, agency, and causal asymmetries. Pearl (2000) is a lucid presentation of a broadly manipulationist approach to causation within a Bayes net framework, with emphasis on the formal characterization of various causal notions, and their representation in terms of directed graphs. Both Pearl and Spirtes, Glymour, and Scheines (1993) discuss and motivate the arrow-breaking conception of intervention and the prediction of the effects of interventions. The latter particularly focuses on issues of causal inference but both books are philosophically very rich, as well as important contributions to the allied literature in statistics and artificial intelligence. Finally, Woodward (2003) develops an interventionist approach to causation and explanation.

References

Call, J., and Tomasello, M. (1997). *Primate Cognition.* New York: Oxford University Press.

Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford: Clarendon.

Collingwood, R. (1940). *An Essay on Metaphysics.* Oxford: Clarendon.

Cook, T., and Frisch, M. (2002). 'Non-Locality in Classical Electrodynamics', *British Journal for the Philosophy of Science* 53: 1–19.

Gasking, D. (1955). 'Causation and Recipes', *Mind* 64: 479–87.

Glymour, C. (2004). 'Review of James Woodward, *Making Things Happen: A Theory of Causal Explanation*', *British Journal for Philosophy of Science* 55: 779–90.

Hall, Ned (2000). 'Causation and the Price of Transitivity', *Journal of Philosophy* 97: 198–222.

Halpern, J., and Pearl, J. (2001). 'Causes and Explanations: A Structural-model Approach—Part I: Causes', *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence.* San Francisco: Morgan Kaufmann, 194–202.

Hausman, D. (1986). 'Causation and Experimentation', *American Philosophical Quarterly* 23: 143–54.

—— (1998). *Causal Asymmetries.* Cambridge: Cambridge University Press.

Hitchcock, C. (2001*a*). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

—— (2001*b*). 'A Tale of Two Effects', *Philosophical Review* 110: 361–96.

—— (2007*a*). 'What Russell Got Right', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited.* Oxford: Oxford University Press.

—— (2007*b*). 'Prevention, Preemption, and the Principle of Sufficient Reason', *Philosophical Review* 116: 495–532.

Holland, P. (1986). 'Statistics and Causal Inference', *Journal of the American Statistical Association* 81: 945–60.

Leslie, A., and Keeble, S. (1987). 'Do Six-Month-Old Infants Perceive Causality?' *Cognition* 25: 265–88.

Lewis, D. (1973). 'Causation', *Journal of Philosophy* 70: 556–67.

Meek, C., and Glymour, C. (1994). 'Conditioning and Intervening', *British Journal for the Philosophy of Science* 45: 1001–21.

Menzies, P., and Price, H. (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44: 187–203.

Norton, J. (2007). 'Causation as Folk Science', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited.* Oxford: Oxford University Press.

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

Price, H. (1991). 'Agency and Probabilistic Causality', *British Journal for the Philosophy of Science* 42: 157–76.

Prinz, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, Mass.: MIT.

Rubin, D. (1986). 'Comment: Which Ifs Have Causal Answers?' *Journal of the American Statistical Association* 81: 961–2.

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

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search*. New York: Springer. 2nd edn. 2000. Cambridge, Mass.: MIT.

von Wright, G. (1971). *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

—— (2007*a*). 'Causation with a Human Face', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

—— (2007*b*). 'Interventionist Theories of Causation in Psychological Perspective', in A. Gopnik and L. Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation*. New York: Oxford University Press.

—— and Hitchcock, C. (2003). 'Explanatory Generalizations, Part 1: A Counterfactual Account', *Noûs* 37: 1–24.

PART III

ALTERNATIVE APPROACHES TO CAUSATION