THE CAUSAL EFFICACY OF MENTAL STATES

Peter Menzies
Department of Philosophy
Macquarie University
Sydney, NSW 2109
AUSTRALIA
Email: pmenzies@laurel.ocs.mq.edu.au

1. INTRODUCTION

You are asked to call out the letters on a chart during an eye-examination: you see and then read out the letters 'U', 'R', and 'X'. Commonsense says that your perceptual experiences causally control your calling out the letters. Or suppose you are playing a game of chess intent on winning: you plan your strategy and move your chess pieces accordingly. Again, commonsense says that your intentions and plans causally control your moving the chess pieces. These causal judgements are as plain and evident as any can be.

However, there is an argument to the effect that non-reductive materialism implies that mental states cannot cause behaviour. Non-reductive materialism is committed to the claim that mental states supervene on physical states and also to the claim that any piece of physical behaviour has a complete physical causal history. These claims can be combined in an argument—the so-called *Exclusion Argument* (Kim, 1998)—that appears to demonstrate the causal inefficacy of mental states. It may appear that your perceptual experiences of seeing the letters on the chart cause you to read them out. But these perceptual experiences supervene on certain physical brain states, which must be presumed to be part of the complete physical cause of your behaviour. If these brain states constitute the complete physical cause of your behaviour, what causal role is left to your distinctively mental states in producing behaviour?

This argument has been endlessly dissected by philosophers. Some have argued that the argument shows the need to give up non-reductive materialism. Others have argued that the argument shows no such thing, as it trades on one or other false assumption. I belong to the second group of philosophers. However, I differ from most of these philosophers in my diagnosis of the argument's error. I

claim that the argument relies on a subtle misunderstanding about the concept of causation.

In essence, the misunderstanding it trades on is that the concept of causation is the concept of a categorical, absolute relation. However, I shall argue that we actually conceptualise causal relations quite differently: we conceptualise them as entities occupying certain functional roles that defined with respect to abstract models. Recognition of this fact opens up the possibility of seeing that there can be different levels of causation. There may be a level at which mental states cause behaviour by way of distinctive psychological pathways; and a different level at which physical brain states cause behaviour by way of distinctive neural pathways. These different levels of causation need not be in competition with each other. This view can be developed, I shall maintain, in a way that's consistent with the fundamental tenets of non-reductive materialism.

Some preliminary remarks first to mark out the general terrain of my discussion. There are, in fact, several different problems of mental causation. One much discussed problem concerns the subjective character of certain mental states, typically non-intentional states: if these mental states have non-physical qualia, how can they be accommodated within the physical picture of the world? Another much discussed problem concerns the extrinsic character of the content of intentional states: if the content of intentional states is constituted, at least in part, by relations to their subjects' physical and linguistic environment, how can these mental states have causal roles distinct from the intrinsic neural properties of their subjects? As interesting as these questions may be, they are not the ones I shall be addressing. To simplify my discussion and to focus in a concentrated fashion on the threat posed to the causal efficacy of the mental by the Exclusion Argument, I shall assume throughout that the mental states being discussed are intentional states that are adequately characterised without reference to their subjects' environments.

2. THE EXCLUSION ARGUMENT AND SOME OF ITS ASSUMPTIONS

There are three main assumptions generally agreed to be necessary to get the Exclusion Argument off the ground. For the sake of the developing the argument, I shall accept all three assumptions, stating them here with only a few brief words of elucidation.

> (1) *The supervenience of the mental on the physical*: Mental states supervene on physical states of the world.

Materialism about the mind requires, at the very least, acceptance of this assumption, though it is a moot issue how the supervenience relation is to be understood in this case. In view of the assumption that the mental states under consideration are intentional states characterised without reference to their subjects' environments, the relevant kind of supervenience is best understood, I think, as 'strong supervenience': necessarily, any two individuals (in the same or different possible worlds) that are indiscernible in their physical properties are indiscernible in their mental properties. Of course, one way in which mental states can supervene on physical brain states is by being identical with them. A *reductive materialist* accepts the supervenience claim by virtue of accepting the stronger claim that mental states are identical with physical states. We will consider this position shortly. For the time being, however, we shall focus on the kind of *non-reductive materialism* that accepts this supervenience assumption without endorsing the identity thesis.

> (2) *The causal closure of the physical world*. Every physical event has a complete physical causal history.

There are a couple of things to explain here. The term 'physical', in its strict application, applies to anything that is the subject matter of physics. This includes physical particulars—such as atoms and quarks—and physical properties—such as mass and velocity. But in the context of the problem of mental causation, the term 'physical' is usually understood more loosely to apply to the particulars and properties of neurophysiology, presumably on the grounds that such entities supervene straightforwardly on the strictly physical. A complete physical causal history of an event consists of a continuous causal chain of physical events leading up to the event in question. It is a causal chain without gaps and without need for completion by non-physical events. In the context of the mental causation problem,

this notion is usually interpreted loosely to mean 'a continuous chain of neural events leading up to the event in question'.

> (3) *The causal relevance of mental properties*. If mental states cause physical effects, they do so in virtue of the mental properties they exemplify.

The question of the relationship between mental states and mental properties is a tricky one. Some believe that mental states are property-like entities; others think that mental states are simply instances of mental properties; others that mental states are very different from properties, belonging to the category of particulars, and that mental properties are just aspects of these particulars. For our purposes, it does not matter what the relationship is between mental states and mental properties. For this reason, I shall use the phrase 'mental states exemplify mental properties' to describe ambiguously the relationship between the two kinds of entity. What matters for the argument is a claim about causation: the claim that mental states have their causal effects in virtue of the mental properties they exemplify.

Whatever the intuitive plausibility of this assumption, it is important to notice that it rules out certain positions on the problem of mental causation. For example, token-identity theorists claim that although mental state-types are distinct from physical state-types, each mental state-token is identical with some physical state-token. Because of their token-identities with physical brain states, mental states are causally efficacious in bringing about behaviour. However, on this view, the causal efficacy of a mental state-token does not reside in any way in the mental property it exemplifies: if it resides in any property at all, it is the physical property it exemplifies. So token-identity theories violate this second assumption. In view of the popularity of token-identity theories, more needs to be said, no doubt, in defence of this assumption. But we shall take its truth for granted here.

These three assumptions seem to generate the conclusion that mental states are causally impotent. Take some instance where a mental state M appears to cause a piece of behaviour B. By assumption (1), the mental state M must supervene on a physical

state P. But by assumption (2), there is an entirely physical causal history to behaviour B. If the mental state M appears to cause the behaviour B, it must be because its subvening physical state P is part of this complete physical causal history of B. But in this case it is the physical state P, and the physical property it exemplifies, that are causally responsible for the behaviour B. The mental property exemplified by M is not at all causally implicated in the production of B. So by assumption (3), the mental state does not, despite appearances, really cause behaviour B. Since the argument proceeds completely schematically, the conclusion we seem obliged to draw is that all mental states of the kind under consideration are causally impotent in the production of behaviour. Nonetheless, this conclusion goes completely against the commonsense view that mental states causally control behaviour in a very real way.

It is important to observe that the Exclusion Argument does not apply to mental states uniquely. The argument applies more generally to any kind of state that supervenes on, without being identical to, ontologically more fundamental physical states. Under its more general application, the argument would seem to show that any states that are not identical with physical states must be causally impotent. For example, it is often said that the biological states of organisms supervene on,without being identical to, molecular chemical states. Further, the assumption of the causal closure of the physical world implies that any biological effect has a completely physical causal history. In this case, the same kind of argument could be mounted to show that biological states, and the distinctive biological properties they exemplify, must be causally impotent. Some philosophers accept this consequence of the argument, claiming that causation exists only at the level of fundamental physics.

However, I take this to be *a reductio ad absurdum* of the argument. It flies directly in the face of commonsense and scientific thought to say that the special sciences do not investigate and discover real causal structures. This view makes a mockery of the enormous efforts devoted in the special sciences to formulating experimental and observational methodologies for testing causal hypotheses. It would follow from this position that all these efforts are misdirected because they could not, by definition, reveal anything about the nature of causal processes. There must be something wrong with the

argument if it leads to this highly implausible result. When it comes to diagnosing the flaw of the argument, we need to keep in mind that the diagnosis must reveal the flaw in the argument in all its applications, not just its application to mental causation.

3. SOME MORE ASSUMPTIONS

Many have noted that the Exclusion Argument relies on more assumptions than the three noted above. Some have sought to avoid the argument's conclusion that mental states are causally inefficacious by rejecting one or more of these extra assumptions. In this section I shall examine three additional assumptions that are needed for the argument and consider some attempts to evade the force of the argument by denying one or other of these extra assumptions.

One obvious extra assumption of the argument is this:

> (4) *Non-identity of mental and physical properties*: mental properties (or mental state-types) are not identical with physical properties (or physical state-types).

Identity theorists argue that the Exclusion Argument is circumvented if mental properties are identical with neural properties. For example, if the mental state-type M is identical with its subvening neural state-type P, then instances of M are causally efficacious in producing physical effects because instances of P are undoubtedly efficacious. One of the main attractions of reductive materialism about the mind is that it offers a straightforward solution to the problem of the causal efficacy of the mental.

Nonetheless, the reductive materialist solution faces a number of well-known difficulties, the most important of which is the problem of multiple realisability. The commonly expressed thought is that different species, different people, and even the same person at different times, can be in the same type of mental state although they vary considerably in the way they neurally realise this state. It was this thought, of course, that persuaded philosophers of mind to adopt a functionalist conception of mental states, according to which mental states are understood as the higher-order role-states rather than lower-order realiser-states. On the standard functionalist

conception, the mental state M is not the first-order neural state P which occupies the causal-functional role characteristic of the state, but rather the second-order state of having *some or other* first-order state occupy this causal-functional role.

One standard identity-theorist response (D. Lewis, 1980; F. Jackson, 1996) to the problem of multiple realisability is to relativise the mental-physical identities. To avoid the difficulty posed by the fact that the mental state M is realised by different neural states in different species (persons or person-stages), the identity theorist asserts the relativised type-identities: in species (or person or person-stage) S_1 the mental state M is neural state P_1, in S_2 mental state M is neural state P_2, in S_3 mental state M is neural state P_3 and so on. In response, the functionalist objects—correctly in my view—that these relativised identities miss out significant commonalities between individuals: rather than saying that all individuals behave in way B because they are in mental state M, the identity-theorist must say that S_1 does B because it is in the state P_1 that occupies the M-role in it, S_2 does B because it is in the state P_2 that occupies the M-role in it, and so on. This splintering of mental properties makes it impossible to capture the psychological commonalities among individuals and the psychological laws governing their behaviour. This matters crucially if, as seems very plausible, mental properties are individuated by the role they play in psychological laws.

Another assumption implicit in the Exclusion Argument is the following:

> (5) *The homogeneity of mental and physical causation*. Mental causation and physical causation have the same fundamental character.

This implies that mental causation is not of a different kind from physical causation. The concept of causation is the same notion applied to the physical and the mental alike. Another way of expressing this assumption is to say that the labels 'mental' and 'physical', as applied to causation, are really transferred epithets—what is mental and physical are the relata of causation, not the causation itself. (T. Crane, 1996) One popular strategy for

evading the force of the Exclusion Argument is to reject this assumption.

Frank Jackson and Philip Pettit's (1990) view on program versus process explanations adopts this strategy. They express their view in terms of causal explanation rather than causation, but that difference is not crucial here. They distinguish between explanations of physical effects in terms of mental states—program explanations— from proper causal explanations of physical effects in terms of earlier physical effects—process explanations. They state that the properties cited in a program explanation may be *causally relevant* in some attenuated sense to the physical effect, but they are not truly *causally efficacious* or productive of the effect. Only the physical properties cited in a process explanation are the genuine article—are truly causally efficacious of the physical effect. The term 'program explanation' refers to the fact that a causally relevant property cited in some program explanation programs for, or ensures, the existence of some genuinely efficacious property in a lower-level process explanation. For example, citing a mental state in explanation of a bodily movement programs or ensures that there is some lower-level physical state that realises the mental state and produces the bodily movement. Evidently, this view implies that the kind of causal relevance a mental state has for a physical effect is different in character from the full-blooded kind of causal relation that physical states enjoy with other physical states.

Some have objected to this view as being a disguised form of epiphenomenalism. This objection is not, however, entirely justified. The constraints Jackson and Pettit impose on causal relevance are non-trivial. There is, after all, a big difference between a mental state that is causally relevant to some bodily movement and a mental state that bears no causal relation of any kind to the bodily movement. The real weakness in this position, it seems to me, is not that it smuggles in epiphenomenalism by the back door, but that it does not vindicate in strong enough terms our intuitions about the causal efficacy of mental states. We believe quite evidently that mental states can control behaviour as straightforwardly as physical states can control other physical states. This robust intuition about the full-blooded character of mental causation is not sufficiently respected by

the rather anemic sort of causal relevance accorded to mental states by Jackson and Pettit.

The last point I would make about both these attempts to circumvent the Exclusion Argument—the attempt that embraces type-identities and the attempt that denies causal homogeneity— is that they fail to answer satisfactorily the more general point raised by the argument. I noted above that the argument does not depend on any distinctive thesis about the mental. It is really concerned with non-physical causation. The same argument could be used to demonstrate the causal inefficacy of any non-physical state. For example, one could run the same argument for the claim that biological states or chemical states, in so far as they are not identical to physical states, are really causally impotent in view of the causal closure of the physical world. The universal application of the argument would result in the conclusion that the only genuine causal relations exist at the level of fundamental physics.

In their different ways, identity theorists and deniers of causal homogeneity either fail to answer this point, or end up agreeing with it. The identities hypothesised by reductive materialists to circumvent the argument in the case of mental causation have some degree of credibility. But it is much less plausible to think that there are systematic type-identities between non-physical properties of any kind whatsoever, on the one hand, and, on the other, the properties of fundamental physics. But nothing less than such systematic identities are required by identity theorists to blunt the force of the universal application of the argument to non-physical states. In contrast, deniers of causal homogeneity do not even regard it as necessary to circumvent the universal application of the argument. Jackson and Pettit (1990) accept that their view implies that the only genuine process explanations are those that cite the causally efficacious properties of fundamental physics. However, as explained above, I see no reason for accepting such an extreme conception of causation that is so obviously out of kilter with the way the concept is used in actual scientific practice.

If the five assumptions we have examined so far are innocent of error, where does the Exclusion Argument go awry? It turns out that there is an additional assumption required for the argument and

this assumption is false, or so I shall argue. I model the formulation of this additional assumption on a principle endorsed by Kim (1989).

> (6) *The exclusion assumption*. With the exception of cases of overdetermination, no event has more than one complete causal history.

Kim argues for this assumption by noting that its apparent counterexamples all involve two or more *incomplete* causes. He observes that we are sometimes called on to select a causal factor that, for various epistemic or pragmatic reasons, is the most appropriate to the situation. A stock example goes like this: we may cite as the cause of the car accident the icy road, the faulty brakes, the driver's inexperience, depending on the explanatory context, even though each of these conditions played an essential role in causing the accident. Kim argues that such cases do not really involve two or more *complete* causes —complete in the sense of being sufficient, without need of supplementation, for the effect.

Kim concedes that examples of overdetermination, if they exist, count as exceptions to the assumption. In such examples, it appears that two causes, each sufficient for the effect, lead by independent causal chains to the effect. Thus, a man dies because he is shot by two assassins whose bullets hit him at the same time; or a building catches fire because of a short circuit in the faulty wiring and a bolt of lightning that hits the building at the same instant. Kim denies, however, that examples of mental causation can be assimilated to cases of overdetermination. Examples of mental causation are importantly different in that they demonstrate a dependence between the putative causes that does not exist in cases of overdetermination. Our earlier schematic example of mental causation involved a piece of behaviour B with two apparent causes, a mental state M and a physical state P, the first supervening on the second. Such supervenience relations between putative causes do not exist in examples of overdetermination. It is the fact of this difference that makes the existence of multiple causal pathways seem coincidental in cases of overdetermination but unnecessarily duplicative in cases of mental causation.

This is not Kim's only argument in favour of the exclusion assumption. He believes that the special character of the

supervenience relation between the mental and the physical actually entails the exclusion assumption in the light of a further principle he endorses. In his view, the best explanation of a mental-physical supervenience relation is that the mental state is a second-order functional state that is realised by a first-order physical state. The realisation relation is an especially clear instance of the supervenience relation. But the following principle, he argues, is very plausibly true of states that are related by the realisation relation:

> *The causal inheritance principle*: if a second-order state $\underline{S_1}$ is realised by a first-order state $\underline{S_2}$, then the causal powers of $\underline{S_1}$ are identical with those of $\underline{S_2}$.

It follows from this principle that the causal power of a second-order mental state M cannot be different from the causal power of the first-order physical state P that realises it. Consequently, the causal chain from M to some piece of behaviour B cannot be different from the causal chain from P to B: there is at most one complete causal history for the behaviour B. In this way, the application of the exclusion assumption to mental causation is vindicated in the clearest possible terms.

Nonetheless, in the following sections I am going to argue, *contra* Kim, that the exclusion assumption and the causal inheritance principle are false. Their plausibility trades on a subtle misunderstanding of the concept of causation. Once the falsity of these principles is highlighted, the flaw in the Exclusion Argument will become evident.

4. CAUSES AS DIFFERENCE-MAKERS

The customary procedure of philosophers when discussing mental causation is to work in terms of particular causal intuitions, rarely referring to substantive theories of causation to support their claims. This way of proceeding is problematic, I think, because untutored intuitions can be deceptive. To give them their proper weight we need to bring them into reflective equilibrium with more general theories. In this section I shall outline a theory of causation which I have been developing in a number of publications (Menzies 1996; 1998; 1999). I shall not describe every part of the theory, only those

parts that bear on the issue at hand. Nor shall I try to justify the theory in detail, beyond showing how it deals with some standard difficulties. My aim is simply to show how the theory can serve as a model for understanding the way in which causation violates the exclusion assumption.

The theory is a theory of the *concept* of causation. I claim that this concept has several different strands, but one important strand is embodied in the idea that a cause is a condition that makes a difference to its effect. (The talk of conditions is meant to refer ambiguously to events, states, absences, and omissions.) This idea has been taken to mean different things and so has been used to motivate quite different approaches to causation: regularity, counterfactual, as well as probabilistic approaches to causation. Without going into all the different ways of spelling out this intuitive idea, I simply note that this idea cannot adequately be spelled out in terms of actual changes in conditions. It may be just an accident that a certain condition is present when the effect is present and is absent when the effect is absent. The idea must be rendered as having modal force in order to rule out such accidental covariations in cause and effect.

How should the modal force of this idea be rendered? It will be useful to start with a purely schematic account of its modal force. In explaining the relevant modal concepts, I shall adopt a modified possible worlds framework—modified with respect to the way possible worlds are conceived. The possible worlds I shall employ are to be understood as mini-worlds rather than alternative large-scale universes. They are alternative courses of development of typically small-scale systems: they are more like trajectories—sequences of states—in the state space of a scientific theory. So while I use the traditional term ‘possible world’, it should always be kept in mind that I understand it unconventionally in this ‘mini-world’ sense.

The following schematic account represents one intuitively plausible way of capturing the required modal force of the idea that a cause makes a difference to its effect:

> *Definition 1*: C *makes a difference* to E in the actual world if and only if in all the possible worlds relevantly similar to the actual world C is always accompanied by E and in these worlds ~C is always accompanied by ~E.

This formulation captures the idea that a cause is a condition the presence / absence of which *necessitates*, at least with respect to a restricted set of possible worlds, the presence / absence of the effect. Consequently, a condition that just happens to covary with the effect cannot count as a cause. (Here I am restricting attention, to be sure, to deterministic settings in which it makes sense to think a cause is sufficient and necessary for the effect. The account can be generalised readily enough to indeterministic settings. To simplify discussion, however, I shall ignore indeterministic settings in the remainder of the discussion.)

Of course, the all important question to be answered here is: which worlds count as relevantly similar to the actual world? To answer this question,I shall argue, we need to recognise that there is not a unique kind of similarity that is involved in each kind of causal claims. Rather a causal claim must be understood as relative to a certain contextual parameter and, depending on the way in which this contextual parameter is set, an appropriate notion of similarity is determined for the given causal claim. Different kinds of causal claims require different similarity relations. Let me offer a brief justification for these remarks.

Any given concrete situation is exceedingly complex in its causal structure. Consider, for example, the simple situation consisting in an individual's performing a simple intentional action—say, the action of raising an arm. There are many different levels of processes going on: neurophysiological processes, biochemical processes, quantum mechanical processes, and so on. There are also countless specific, sometimes distinctive and idiosyncratic, facts about this particular action that do not apply more generally to other actions: perhaps the agent is suffering from a severe disease, has been exposed to radiation, is suffering from great stress, and so on. Our finite mental capacities make it impossible for us to take in at once all these different kinds of processes and all the messy specificity of such a concrete situation. In order to understand the causal structure of

such a concrete situation, we focus on some aspects of what is going on and ignore others. In other words, our causal thinking is steeped in *abstraction*: the causal schemas by which we interpret the world are irremediably permeated by the kind of abstraction that allows us to selectively attend to some aspects of the world while backgrounding others. There seem to be two essential elements to the kind of abstraction that underlies our causal thinking.

The first element is *generalisation*. We simplify the complexities of a concrete system by seeing it as instance of a certain kind of system and furthermore of certain kind of system governed by certain laws. That is to say we implicitly generalise: we think that what holds for the particular system under consideration should hold for any system of the same kind. Hume was right to the extent that he thought that singular causal relations depend on general causal claims and laws. To take our example of the simple intentional action of an individual's raising an arm: one way to understand this concrete situation is to see the individual as a purely neurophysiological system conforming to the laws of neurophysiology; another way of understanding it is to see the individual as a rational agent conforming to the laws of rational intentional action. Both ways of understanding the situation involve abstracting from the complexities of the situation by selectively focusing on one distinctive set of features of the world.

The second element in the abstraction that permeates our causal thinking is *idealisation*. The laws that govern the functioning of the systems we consider typically hold only under *ceteris paribus* conditions. Such *ceteris paribus* laws imply that the systems in question evolve along certain trajectories *provided nothing interferes*. This is, of course, an idealisation because actual systems are often enough subject to interfering factors, as well as the causal influences and forces covered by the laws. Recent work by Geoffrey Jopseph (1980), Nancy Cartwright (1983; 1999), Ronald Giere (1988), and Fred Suppe (1989) has done much to support the idea that scientific laws, both in the physical and special sciences, are invariably *ceteris paribus* laws. The simple observation that human intervention is possible in the operation of practically all natural systems makes it implausible to think that exceptionless laws, unqualified by *ceteris paribus* clauses, can be formulated to cover every contingency to do

with human intervation. The laws of neurophysiology, as well as the laws of rational intentional psychology, hold only under the condition that there are no external interferences, including interventions by other agents.

To capture the fact that our causal thinking about a concrete situation is permeated by these two kinds of abstraction, I shall say our causal claims must be understood as relative to a causal model of the concrete situation. I represent a *causal model* as an ordered pair <S, L>, where first element S is the kind of system in terms of which the concrete situation is conceptualised, and the second element L is the set of laws, typically *ceteris paribus* laws, governing the operation of that kind of system.

How exactly are causal claims about a concrete situation relative to a causal model? I want to argue that the relativity of causal claims to a model consists in the fact that the model determines the respects of similarity used in evaluating whether a putative cause makes a difference to an effect. The following definition makes this idea more precise.

> *Definition 2*: A model <S, L> *generates a set of most similar worlds* to the actual world (for a putative causal condition C) consisting of the following worlds:
> (i) the worlds contain a system of kind S with the same history as the actual system of kind S up until shortly before the time of C;
> (ii) the worlds conform to the laws L governing the systems of kind S;
> (iii) the worlds are ones in which the *ceteris paribus* conditions of the laws L, if any, hold (ie systems of kind S evolve lawfully in the absence of interferers).

An example may help to explain the various features of this definition. Suppose we are considering whether a pattern of neuronal firings in an individual's motor cortex made a difference to the individual's arm rising on a particular occasion. To answer this question we have to consider all the most similar possible worlds generated by the neurophysiological causal model implicitly under consideration. In the first place, the most similar worlds will be like

the actual world in containing a human body with the appropriate nervous system. These counterpart systems do not have to be atom-for-atom identical with the actual human body, merely similar in all the respects that count from a neurophysiological point of view. Further, the counterpart human bodies should have the same history as the actual one up until shortly before the time at which the putatively causal neuronal firings took place. We need to preserve the history of the actual body in the most similar worlds so as to hold fixed every causal factor, at the level of neurophysiology, that might have influenced the arm's rising. Secondly, the most similar worlds will be like the actual world in conforming to the laws of neurophysiology that govern the functioning of human nervous system. Perhaps, laws of biochemistry and anatomy required for the operation of neurophysiological laws are to be held fixed in the these worlds. However, it would be unfaithful to the way we actually conceptualise these matters to require that the most similar worlds hold fixed extraneous laws such as those of quantum mechanics and special relativity. Thirdly, the most similar worlds will preserve the *ceteris paribus* conditions of the relevant laws. As explained above, this reflects the idealisation by which we conceptually abstract from the complexities of a concrete situation the salient features relevant to establishing a difference claim. This feature will assume greater importance in the next section and will be discussed at further length there.

In terms of the terminology just introduced, we can explain the idea that a cause makes a difference to its effect in a way that acknowledges the relativity to models.

> *Definition 3*: $\underline{C}$ makes a difference to $\underline{E}$ relative to the model $\underline{X} = \langle \underline{S}, \underline{L} \rangle$ if and only if every most similar $\underline{C}$-world generated by the model is a $\underline{E}$-world and every most similar $\sim\underline{C}$-world generated by the model is an $\sim\underline{E}$-world.

The figure below represents the situation in which $\underline{C}$ makes a difference to $\underline{E}$ relative to the model $\underline{X}$. @ represents the actual world and $\underline{W}_X$ represents the set of most similar worlds to the actual world generated by the model $\underline{X}$.

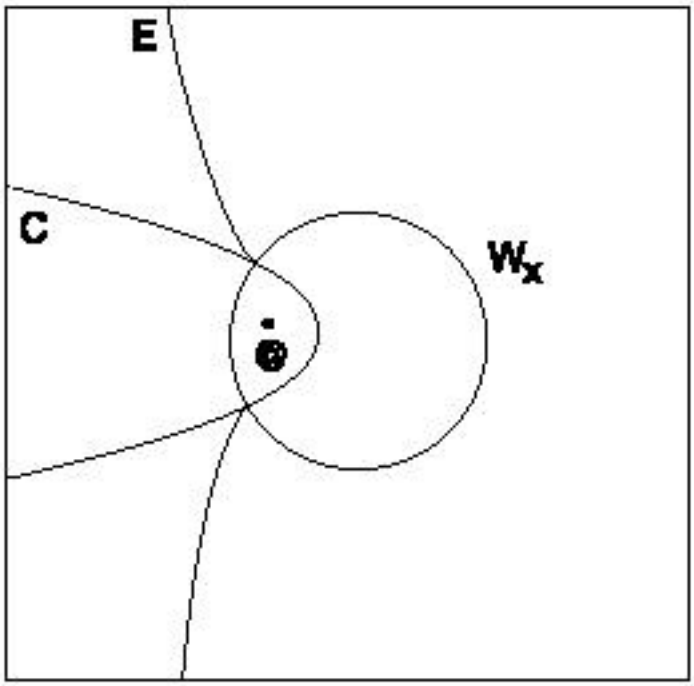

Figure 1

This relativised truth-condition can be conjoined with the usual truth-conditions for counterfactuals to yield the following condition:

> *Definition 4*: C *makes a difference* to E relative to the model X = <S, L> if and only if C œ->$_X$E and ~C œ->$_X$~E.

Here the subscript X on the counterfactual operator signifies that the operator is defined over the set of most similar worlds generated by the model X.

Of course, this counterfactual construction is very similar to the notion of *counterfactual dependence* that plays the central role in Lewis's counterfactual theory of causation. (Lewis, 1973) It will be useful to be able to take over this terminology. But the way I will use the term is different from the way Lewis uses it in two respects. First, the notion of counterfactual dependence, as I will use it, inherits the relativity to a model of the counterfactuals that define it. The truth-conditions of counterfactuals in my theory are defined over the most similar situations *generated by a model*. Lewis's notion involves no such relativity, mostly because he assumes that the laws which govern the evolution of his large-scale possible worlds are exceptionless laws, unqualified by *ceteris paribus* conditions.

Secondly, my notion of counterfactual dependence differs from Lewis's in that I reject the Centering Assumption that he imposes on counterfactuals. On his theory of counterfactuals, where an antecedent C is true, there is only one closest-antecedent world to consider—namely, the actual world. This means that the counterfactual C œ->E is true if E is true in the actual world as well as C. I believe, on the other hand, that even where C is true, the truth of C œ->$_X$E requires E be true, not just in the actual world, but in all the most similar worlds generated by the relevant model X for the condition C.

Imposing this requirement enables us to maintain a useful distinction that is blurred in Lewis's theory. According to this theory, if a determinable condition C (say, my waving my arm) and a determinate condition C* (say, my waving my arm vigorously) both hold in the actual world, the fact that some condition E (say, my attracting the attention of a taxi driver) counterfactually depends on a determinate condition C* implies that E counterfactually depends on the determinable C. But a theory of causation should be able to say that a condition E (my attracting the taxi driver's attention) was

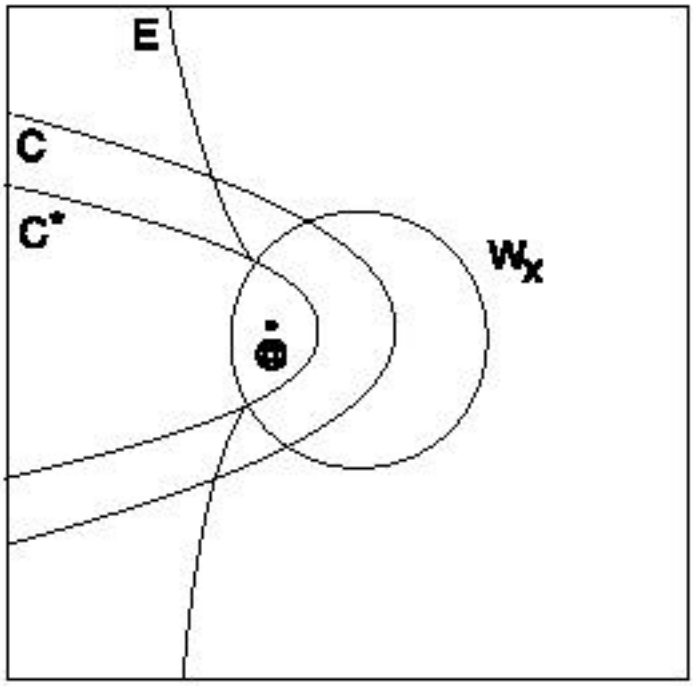

Figure 2

caused by the determinate condition $\underline{C}^*$ (my waving my arm vigorously) and not by the determinable condition $\underline{C}$ (my waving my arm *simpliciter*). By requiring that the counterfactual $\underline{C}$ œ->$_X$ $\underline{E}$ holds non-trivially as well as ~$\underline{C}$ œ->$_X$ ~$\underline{E}$, we can draw this distinction. The diagram above shows the case where Lewis's theory implies that $\underline{E}$ counterfactually depends on $\underline{C}$ because it depends on $\underline{C}^*$, but my theory implies that $\underline{E}$ counterfactually depends on $\underline{C}^*$, but not on $\underline{C}$.

This account of making a difference allows that two distinct kinds of conditions can make a difference to a piece of behaviour relative to different models. Relative to a neurophysiological causal model, for example, it may be certain neuron's firing in the motor cortex that make a difference to the agent's raising of the arm. While relative to an intentional causal model, it may be the agent's reasons, consisting in a complex of a belief and desire—say, the desire to catch a taxi-driver's attention and the belief that the best way of doing so is to raise an arm— that cause the agent's raising of the arm.

A diagrammatic representation will help to make the point. The following diagram represents a logical space of possible worlds fine-grained enough to model the lawful evolution of a single individual, considered both as a rational agent and as a neurophysiological system. In this diagram $\underline{W}_N$ represents the set of most similar worlds generated by a neurophysiological causal model $\underline{N}$; and $\underline{W}_I$ represents the set of most similar worlds generated by an intentional causal model $\underline{I}$. These are different sets of possible worlds, as the things that have to be held fixed in considering the evolution of an individual as rational agent are different from the things that have to be held in considering the evolution of the individual as a neurophysiological system. (For a start, different *ceteris paribus* laws have to be held fixed, as do different kinds of history of the individual.) The actual world @ is included in the overlap of these sets of possible worlds, as the particular individual is both a rational agent and a neurophysiological system. In the diagram, $\underline{A}$ is the action of raising an arm, $\underline{R}$ is the state of having reasons for performing $\underline{A}$, and $\underline{NF}$ is the state of having certain neurons in the motor cortex fire. The diagram shows that the following counterfactuals are true.

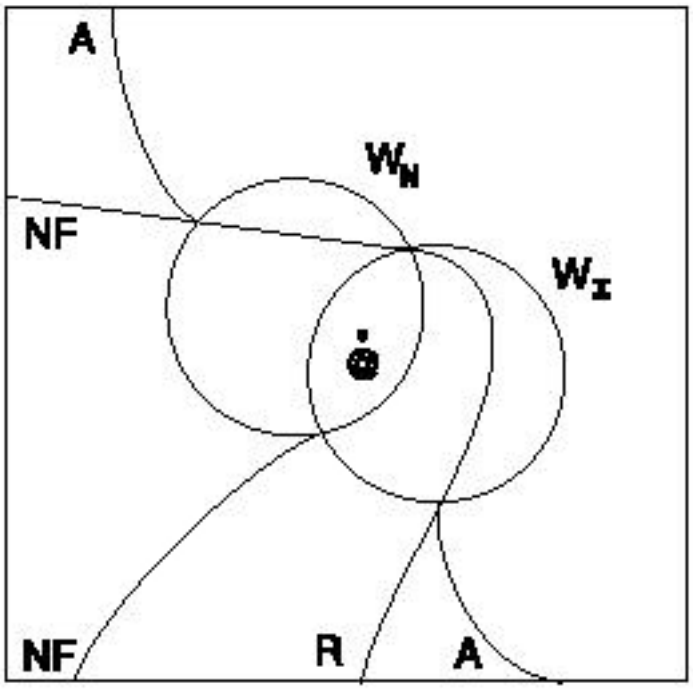

Figure 3

$\underline{NF}$ œ->$_N$ $\underline{A}$ and ~$\underline{NF}$ œ->$_N$ ~$\underline{A}$
$\underline{R}$ œ->$_I$ $\underline{A}$ and ~$\underline{R}$ œ->$_I$ ~$\underline{A}$

Relative to the intentional model $\underline{I}$, the action $\underline{A}$ counterfactually depends on reasons $\underline{R}$; but relative to the neurophysiological model $\underline{N}$, it depends on neural firing $\underline{NF}$.

5. CAUSATION FUNCTIONALISED

If causation simply consisted in making a difference, as explicated in terms of counterfactual dependence, the falsity of the exclusion assumption could be demonstrated quite straightforwardly with an example such as the one above. But we cannot proceed so quickly. The notion of counterfactual dependence does not constitute the whole of the concept of causation. Indeed, in my view, the existence of a counterfactual dependence is merely the surface marker of a deeper phenomenon that really counts as causation. I have tried to show elsewhere (1989;1996) that examples involving actual or

potential pre-emption demonstrate the impossibility of analysing causation in terms of counterfactual dependence.

To illustrate this point, consider Harry Frankfurt's (1969) well-known pre-emption example, which I shall call the Stand-by Intervener. A neuroscientist wants an individual Jones to raise his arm. She hopes that Jones, left to himself, will do so. By reading his brain, she can predict what he will do if left to himself. If she reads that he will do what she wants, she does nothing further. But if instead she reads that he will not do as she wants, she takes control, manipulating his brain and nervous system to produce the desired behaviour. As it happens, Jones' brain is in the state that would normally lead to his raising his arm; and indeed he does raise his arm and the neuroscientist remains idle. In this case, the state of Jones' brain state is a pre-empting cause of his behaviour and the idle neuroscientist is a pre-empted backup.

Such examples pose special problems for purely counterfactual theories of causation. Suppose we tried to analyse causation in terms of counterfactual dependence in the following terms: condition C is a cause of a distinct condition E relative to the model X if and only if C makes a difference to E (or, alternatively, E counterfactually depends on C) relative to this model. Such an analysis would be refuted by the Stand-by Intervener, if the model in question identified the relevant system as Jones' brain conjoined with the stand-by neuroscientist. For in this case no counterfactual dependence would hold between Jones' brain state and his behaviour with respect to the most similiar worlds generated by this model. It would be true that, given Jones was in the brain state, he had to raise his arm. But it would not be true that if he had not been in this brain state, he would not have raised his arm: for if he had not been in that brain state, the neuroscientist would have intervened to ensure that he behaved in the desired way.

One way of overcoming this particular kind of difficulty has been explored by Lewis (1973;1999). It amounts to analysing causation in terms of chains of stepwise counterfactual dependences. Adapted to the present framework, the suggestion is this: C is a cause of a distinct condition E relative to the model X if and only if running from C to E is a chain of stepwise counterfactual

dependences that hold relative to this model. Lewis has pointed out that a careful examination of this example reveals an event that constitutes an intermediate step in a chain of counterfactual dependences between Jones' initial brain state and his behaviour. Consider a time after the neuroscientist has read Jones' brain but before she would intervened if the reading had been different. Occurring at this time is a complex event consisting of Jones' brain state at that time combined with the neuroscientist's decision not to intervene. This complex event is counterfactually dependent on Jones' initial brain state; and Jones' behaviour is, in turn, counterfactually dependent on it. Thus, we have a chain of stepwise counterfactual dependences, and so causation, as required.

This ingenious idea does not, however, deal satisfactorily with all the problem cases. Consider a modification of the Stand-by Intervener Case, which I call the Active Intervener. The neuroscientist is still concerned to ensure that Jones raises his arm. She can still read his brain to determine whether, left to himself, he will raise his arm. However, if she reads that his brain is in the state that would normally lead to his raising his arm, she seizes control of his brain; she intervenes to prevent the brain state from having its immediate effect and then, further along the usual causal pathway, to ensure that he does raise his arm after all. On the other hand, if she reads that Jones is in some other brain state, she does nothing at all, thus ensuring that he does not raise his arm. As it happens, Jones' brain is in the state that would normally lead to his raising his arm. The neuroscientist intervenes to prevent this brain state from having its normal effects and then causes him to raise his arm by a non-standard causal route.

The Active Intervener is a counterexample to the sufficiency of the hypothetical analysis, whereas the Stand-by Intervener is a counterexample to its necessity. On the assumption that the model under consideration identifies the relevant system as including the neuroscientist's presence, which is accordingly held fixed in the most similar worlds genertaed by the model, there is a counterfactual dependence between Jones' brain state and his raising his arm. For, given that Jones was in the brain state, the neuroscientist's interventions guaranteed that he would raise his arm; and if he had not been in this brain state, she would not have intervened, ensuring

that he would not raise his arm. So there is a counterfactual dependence and so, *a fortiori*, a chain of counterfactual dependences between the two events. The problem for the hypothetical theory is that the counterfactual dependence is not matched by a corresponding causal connection. Jones's brain state did not cause him to raise his arm; rather it was the neuroscientist's interventions that caused him to do so. (The fact that Jones' initial brain state caused the interventions does not entail the required causal conclusion, as the transitivity of causation cannot be taken for granted here.)

How are we to understand such examples? The key to understanding them lies, I think, in realising that the examples involve the presence of an actual or potential intervener—the neuroscientist. It is very natural to think of her presence as an interfering factor of the kind that should be excluded by a proper identification of the relevant system. If we model the situations in this way to exclude the neuroscientist from the relevant systems, treating her acts as interfering factors, the resulting counterfactual dependences are particularly well suited to establishing causal conclusions. For example, if the most similar possible worlds generated by a model of the Stand-by Intervener example hold fixed the absence of the neuroscientist, then there will, after all, be a counterfactual dependence between Jones' initial brain state and his behaviour. This counterfactual dependence reflects the causal relation that exists between these events.

But how does the fact that this counterfactual dependence exists in a *hypothetical* situation that abstracts away from the presence of the neuroscientist help to identify the causal relations in the *actual* situation in which she is very much present? The answer is that there is a type of process picked out by the counterfactual dependence defined for the system of brain-*minus*-intervener and a process of this type can exist in the system of brain-*plus*-intervener. The process in question is a temporally ordered sequence of events taking place in Jones' brain and nervous system. It might consist of a sequence of events such as this: certain neurons fire in Jones' motor cortex, spikes run down his spinal cord, certain motor neurons fire in his spinal cord, a spike runs along these neurons to muscles in his arm, these muscles contract and so on. This very process may hold

between Jones' initial brain state and his behaviour in the actual circumstances of the Stand-by Intervener example, even though the presence of the neuroscientist frustrates a corresponding counterfactual dependence between them. The existence of this process represents a very good reason for thinking that the two events in question are causally related in the actual circumstances.

There are several ideas used in this informal argument that require clarification. An important one is the idea of a process picked out by a counterfactual dependence.

> *Definition 4*: The counterfactual dependence of E on C relative to the model X *picks out a process* (a temporarily ordered sequence of events) if and only if the process is present in all the most similar C-worlds generated by the model that are E-worlds and is absent in all the most similar ~C-worlds generated by the model that are ~E-worlds.

With this notion in hand, we can capture the above informal reasoning about causation in the following definition.

> *Definition 5*: The condition C is *a cause* of the distinct condition E relative to the model X of an actual situation if and only if (i) E counterfactually depends on C relative to the model; (ii) this counterfactual dependence picks out a process; and (iii) this process connects C and E in the actual situation.

The intuitive idea behind this analysis is easy enough to understand when it is applied to particular examples. Here are the steps we should go through in determining whether Jones's brain state is a cause of his behaviour in the Stand-by Intervener and Active Intervener examples. First, let us assume that the relevant system we are considering consists of Jones' nervous system, minus the presence of the neuroscientist. It is natural to generalise over systems of this kind and to regard the neuroscientist as an actual or potential intefering factor. Now we ask about this system of Jones's brain-minus-intervener whether there would be a counterfactual dependence between his brain state and his behaviour, relative to the model under consideration. If the answer is 'No', we can conclude immediately that there's no causal relation. If the answer is

‘Yes’, as we find to be the case with both the examples, we proceed to ask the following question: does the counterfactual dependence pick out a process? If the answer is ‘No’, we can conclude that there’s no causal relation. If, on the other hand, the answer is ‘Yes’, as is the case with both these examples, we proceed to ask the final question: does this process exist in the actual situation? If the answer is ‘No’, there’s no causal relation between Jones' brain state and his raising his arm. If, on the other hand, the answer is ‘yes’, we can conclude that there is a causal relation. In the particular examples in question, there is a process picked out by the counterfactual dependence in the system of brain-minus-intervener. It is the sequence of events occurring in Jones' nervous system that was described earlier. A process of this same type exists in the actual situation of the Stand-by Intervener. So we can reason that there is a causal relation between his brain state and his behaviour. However, no process of this type exists in the actual situation of the Active Intervener. Accordingly, we can reason that there is no causal relation between the two events in this case. These conclusions are the intuitively correct ones.

At this point I should try to answer an obvious objection to this account of causation. The objection is that the account is vitiated as a reductive analysis by its implicit appeal to the causal notion of an interferer in a system. The notion of an interferer enters the account in several places. For instance, to work out whether a causal relation exists in an actual situation, the account says we have to identify the situation as involving a certain kind of system and consider how the system would evolve in conformity with its laws in the absence of any interfering factors. But what does it mean for some factor to be an interfering factor? Without being overly precise about the matter, one can say that an interfering factor in a system is any factor having an independent causal history which would, if present, introduce new causal processes into the system. (On this understanding, human interventions into a system are paradigm interferers.) Evidently, the notion of an interferer is a causal notion. So the use of this causal notion seems to introduce a vicious circularity in the analysis.

My reply to this objection is to concede that the notion of an interferer is indeed a causal notion, but to argue that this fact does

not vitiate the account. An enlightening way to understand the account is to see it as analysing kinds of causal relations in terms of their functional roles. Thus, an alternative formulation of the analysis runs like this: the causal relation between the condition C and the distinct condition E relative to the model X is the process that is picked out by the counterfactual dependence holding between C and E relative to the model. Roughly speaking, the analysis identifies a causal relation as the unique process that occupies a certain counterfactually defined role.

It is useful to compare this analysis with functional-role analyses of mental states of the kind advanced by early identity theorists like Armstrong and Lewis. According to such analyses, the mental state of belief is to be understood as the state that occupies the characteristic functional-role associated with belief. (An even better comparison is with the relativised functional-role analyses that Lewis (1980) offered in response to the multiple realisation problem. On the relativised analysis, a belief for kind K is the state that occupies the belief functional role in kind K.) Of course, my intention is not to endorse such analyses. My point is simply that it was not a good objection to such functional-role analyses that the specification of the functional role of one mental state would invariably advert to other mental states. (For example, the specification of the functional role of belief would have to accommodate the fact that beliefs typically cause behaviour only in conjunction with other mental states like desires.) The concepts of different kinds of mental states are interconnected, and so have to be functionally analysed as collective wholes. But this fact is not a problem for the functional-role theorists who identified mental states, not with a functional role, but with the state that occupies the role. Provided the state that occupies the given functional role is not itself mental in character, the functional-role analysis is not vitiated by circularity.

Correspondingly, the proposed account of causation should be viewed as offering functional-role analyses of kinds of causal relations. To be sure, in specifying the functional role of a certain kind of causal relation, we must advert to other causal relations, such as those involved in the operation of interferers. However, the interdependence of causal concepts does not harm the theory, because a causal relation is identified, not with its functional role, but

with the process that occupies the functional role. Since such processes are causal-free entities, being temporally ordered sequences of events or states, the functional-role analyses of causal relations are innocent of any vicious circularity.

6. TWO CAUSAL MODELS OF BEHAVIOUR

In this section I shall consider how this functional-role acccount of causation can be applied to the problem of the causation of behaviour. I shall argue that the account permits different but non-exclusive causal models of behaviour and that these causal models pick out distinctive causal pathways to behaviour .

Let us consider the different ways in which we can understand the causal processes involved in our illustrative action of raising one's arm with the intention of signalling a taxi. As we have seen, one way of understanding these processes is in terms of a neurophysiological model that abstracts from complexity of the situation by conceptualising the individual agent as a nervous system governed by the idealised laws of neurophysiology. In this model, determining the causes of the individual's behaviour requires looking for a condition that makes a difference to the behaviour with respect to the most similar worlds generated by the model. Let us suppose that certain neuron firings in the individual's motor cortex make a difference to his behaviour in the counterfactual dependence sense explained earlier. The next stage is to determine whether the counterfactual dependence would pick out a process. It is plausible to think that it would pick out a process consisting of the familiar sequence of events: certain neurons fire in the agent's motor cortex, spikes run down to his spinal cord, certain motor neurons fire in his spinal cord, a spike runs along these motor neurons to muscles in his arm, these muscles contract, and so on. Finally, if this process obtains in the actual situation where the agent raises his arm, then it is true, relative to this model, that the neurons' firing in his motor cortex caused his behaviour.

As we have also seen, there is another way of understanding what goes on when an agent performs the intentional action of raising his arm. This is in terms of an intentional causal model which abstracts from the complexities of the situation by conceptualising the

individual as a rational agent governed by the idealised laws of rational intentional action. In this intentional model, determining the cause of an individual's raising his arm requires locating some condition that makes a difference to this behaviour with respect to the most similar worlds generated by the model. The obvious candidate for the condition are the mental states that are the agent reasons for raising his arm— perhaps his desire to catch a taxi-driver's attention and the belief that the best way of doing so is to raise one's arm. The next question to ask is whether the counterfactual dependence between his reasons and behaviour picks out a process. It is plausible to think that there is some process such as this: the agent makes a practical evaluation on the basis of his beliefs and desires, he forms an all-things-considered evaluation, he forms an intention to raise his arm, and he raises his arm. If this process exists in the actual situation in which the agent raises his arm, then it is true, at least with respect to this model, that his reasons caused him to raise his arm.

It has sometimes been claimed that, in view of the central role that considerations of rationality play in the intentional model, it is not, properly speaking, a causal model at all. This is the position of many Wittgensteinians who argue that reasons cannot be causes. One Wittgensteinian argument is that agents' reasons are logically connected to their actions in a way that causes cannot be connected to their effects. Another argument is that the laws governing rational intentional action are cluttered with*ceteris paribus* conditions, unlike the strict laws governing genuine causal interactions. In my opinion, these arguments have long been refuted by Davidson (1963) and others. (See J. Bishop, 1989 for discussion.) The replies to these arguments I favour are these. First, some events are described in terms their of causes, but this does undermine the standing of the causal statements involving them. Sunburn is an event described in terms of its causes, but it is still true and informative to say 'Exposure to the sun caused Flora's sunburn'. Secondly, the laws governing straightforwardly physical causal interaction are no less cluttered with *ceteris paribus* conditions than the laws of rational action.

There are also compelling positive arguments in support of the claim that reasons can be causes. Davidson, for instance, offered one

simple but highly persuasive argument for the claim that, when agents act for reasons, their reasons must be the causes of their behaviour. Intentional action must occur, Davidson argued, not just *in the presence* of the agent's reasons for doing it but also *because* the agent had those reasons for doing it. This 'because' condition is required to avoid collapsing the distinction between intentional action on the one hand, and, on the other, behaviour that is rationally justified in the light of the reasons that the agent happens to have at the time. For example, I may stay silent during an important meeting simply because my laryngitis renders me dumb and yet, at the same time, have very good reason for keeping quiet. The 'because' condition is most simply understood as requiring that the agent's reasons were the causes of the agent's behaviour. (For more discussion of this point see J. Bishop (1989).)

On the basis of this defence of the thesis that reasons can be causes, Davidson set about developing a causal theory of intentional action. A number of conditions were generally agreed to be necessary conditons for the truth of the claim that an agent performs an intentional action A: the agent must be in a mental state R (a belief-desire complex); the state R makes it reasonable for the agent to do A; and the agent's being in the state R causes the action A. However, Davidson denied that these conditions were sufficient for intentional action because of the problem of 'deviant causal chains'. He gave the following as an example:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally (1973).

Davidson at first thought that a condition could be added to deal with such examples of deviant causal chains that would bring the necessary conditions up to sufficiency. But, after several failed attempts, he despaired of finding such a condition.

Davidson gave up hope too quickly, in my view. There is no need to supplement the three conditions to bring them up to sufficiency if causation is understood in accordance with the present account. It is very natural to view the example of the climber as one in which the climber's nervousness, induced by the realisation of his own thoughts, is an interfering factor in his practical reasoning. If so, determining whether there is a causal relation in the particular case requires considering whether there would be a counterfactual dependence between the agent's reasons and his behaviour with respect to the system that abstracts away from this interference. It seems reasonable to think that, in the absence of the nervousness, there would be such a counterfactual dependence. The next question is to ask whether this counterfactual dependence would pick out a process. Presumably, the processes of normal practical reasoning would occur without the interference of the nervousness and we could expect some sequence of events like this: practical evaluation of options, formation of all-things-considered evaluation, formation of intention to act, action. If we ask whether such a process exists in the actual situation in which the climber lets go of the rope, we see that it does not. For this reason we are reluctant to say that the climber's letting go was caused, non-deviantly, by his reasons. In similar fashion, the causal theory of action, formulated in terms of the three conditions above, can handle other examples of deviant causal chains, when it is conjoined with the proposed account of causation.

An important issue that needs to be settled at this point is whether it is possible to bring the intentional model and the neurophysiological model closer together. As we have seen, causal relations, on both models, are processes picked out by counterfactual dependences. It would simplify matters considerably if the processes picked out by the counterfactual dependences of the intentional model were in fact the very same processes cited by the neurophysiological model. This simplifying proposal would bring the two models into harmony, implying that they are simply two ways of specifying the same underlying processes. This proposal would clearly have far-reaching implications for the problem of mental causation.

As attractive as this simplifying proposal may be, there are two decisive reasons against it. First, we have seen, that the issue of how the processes of the intentional model are identified is crucial to rescuing the causal theory of action from the problem of 'deviant causal chains'. It would be a serious mistake, however, to make the causal theory of action hostage to the scientific discoveries of neurophysiology. Our practical mastery of the concept of intentional action indicates that we already know very well what we mean by it, so the analysis of its meaning cannot involve any unknown empirical information about the types of neurophysiological pathways. We know the difference between an intentional action and an involuntary response without knowing any details of neurophysiology. This indicates that the solution to the problem of 'deviant causal chains' is to be solved at the functional level of intentional psychology, rather than the level of neurophysiological realisation.

This brings us to the second reason for rejecting the suggestion. The functionalist argument from multiple realisability implies that complete psychological processes, leading from mental states to behaviour, can be multiply realised: for example, the psychological process involved in an individual's practical reasoning. So empirical information about the neurophysiological processes going on in an individual engaged in practical reasoning will give no more than the facts about how the process of practical reasoning is realised in that individual. In other individuals, a process with the same functional characteristics may be realised very differently in physical terms. Discoveries about the physiological basis of behaviour will not, then, yield the the kind of understanding of processes that is required in the intentional model. That can only be supplied by information about the functional level of psychological processes.

7. THE EXCLUSION ARGUMENT AGAIN

We are finally in a position to return to the Exclusion Argument. Let us begin by considering how the proposed account of causation bears on each of the assumptions of the argument, focusing especially on the exclusion assumption.

First, the assumption of the supervenience of mental states on physical states. The account of causation does not bear directly on the issue of whether mental states supervene on physical brain states. So consistently with this account, we can accept the supervenience assumption.

Second, the assumption of the causal closure of the physical world. We saw that this assumption should be read, in connection with the mental causation problem, as the thesis that every piece of physical behaviour has a complete neurophysiological causal history. The present account of causation does not contradict this, as it allows that there is a causal model of behaviour —the neurophysiological model— according to which neural events cause behaviour by way of complete, non-gappy neurophysiological processes. (This model of behaviour does not, however, exclude other models such as the intentional model, according to which an agent's reasons cause behaviour by way of distinctive psychological pathways.)

Third, the assumption of the causal relevance of mental properties. The present account of causation vindicates this assumption, when it is read as pertaining to the intentional model of behaviour. For in this model the counterfactual dependences that pick out causal relations rely on the causally related events' exemplifying mental properties. The causal relations, so identified, differ from those picked out in the neurophysiological model by the counterfactual dependences linking events that exemplify neural properties rather than mental properties. It makes sense to say, then, that when an agent’s mental states cause behaviour, they do so in virtue of the mental properties exemplified by these states.

Fourth, the assumption of the non-identity of mental properties and neural properties. The account of causation supports this assumption because it implies that, within a given causal model, mental properties can have causal roles independent of the neural properties on which they supervene. As we shall soon discover, the mental property of having a certain reason for performing an action may have a causal role independent of the neural property upon which it supervenes. In view of their difference in causal role, the properties cannot be identical.

Fifth, the assumption of the homogeneity of mental and physical causation. The theory does not repudiate this assumption, though it may appear to do so. (One might ask: are not the psychological processes identified as causal relations in the intentional model different in character from the physical processes identified as causal relations in the neurophysiological model?) However, the precise formulation of the assumption states that the *concept* of mental causation is the same as the *concept* of physical causation. This is, indeed, the case on the proposed account. It is the very same concept of causation that picks out these different processes in the different models. The unitary concept represents causation in terms of a counterfactually defined functional role, a role that can be realised by different processes in different kinds of models.

Finally, the exclusion assumption, which states that no event has more than one causal history. It is this assumption that is falsified by the present account of causation. The account, as we have seen, allows that a single event—in our example, the event of an individual's raising an arm—can have two different complete causal histories. Relative to the intentional model, the agent's reasons may be truly said to cause the agent's raising of the arm. Relative to the neurophysiological model, it is neurons' firing in the motor cortex which may truly be said to cause the behaviour. These are different causal relations, consisting in distinct processes, which can coexist because they are posited by two models, neither of which excludes the other.

There are similarities between this kind of causal situation and examples of causal overdetermination. Our causal judgements about overdetermination examples are formed in the roughly same kind of way as our causal judgements about mental causation. Consider the example of the building that catches fire because of a short circuit in the faulty wiring and a bolt of lightning that hits the building at the same instant. To figure out whether the short circuit is a cause of the fire, we ask: would there be process picked out by a counterfactual dependence between the short circuit and the fire in a hypothetical scenario involving the building but no lightning bolt; and, if so, does this process hold in the actual situation? It is plausible to answer 'Yes' to both questions. We ask a similar question to work out

whether the lightning bolt caused the fire. We ask: would there be a process picked out by a counterfactual dependence between the lightning bolt and the fire in a hypothetical scenario involving the lightning and the building but no short circuit; and, if so, does this process exist in the actual situation? Again it is reasonable to answer 'Yes' to both questions. Accordingly, it is appropriate to say that the house fire was overdetermined by the two causes.

Despite these similarities, I think we must agree with Kim that there are vital differences between the examples of mental causation and the examples of overdetermination. The difference is that a supervenience relation links the mental and physical causes of behaviour in cases of mental causation, whereas the causes of overdetermination examples are not so linked. It is this difference which explains why the existence of multiple causal pathways seems coincidental in overdetermination examples but not in cases of mental causation: if mental state M supervenes on brain state P, then it is not surprising that, when there is causal pathway from M to some behaviour B in the intentional model, there is a causal pathway from P to B in the neurophysiological model.

Another significant difference between the examples of mental causation and the examples of overdetermination is that the multiple causes in the examples of mental causation are picked out within different models, whereas the multiple causes of overdetermination examples are picked out within the same model. This difference makes it clear why it is unsatisfactory to claim, as Kim and others do, that the postulation of multiple causal pathways to the same behaviour involves an unnecessary duplication. This claim unsatisfactorily begs the question in presupposing that one causal pathway suffices for the explanation of a phenomenon, making other causal pathways explanatorily redundant. This line of thought fails to recognise that our claims about the causation of behaviour are made relative to models and that different models involve different kinds of abstraction that shape the identification of different, non-competing causal processes.

In view of these vital differences, I agree with Kim that we should not assimilate the cases of mental causation to cases of overdetermination. Accordingly, we cannot see the mental causation

cases as falling within the exclusion assumption's explicit exception clauses relating to overdetermination cases. They represent an altogether new kind of counterexample to the principle.

There is one last defence of the exclusion assumption that Kim can fall back on. It is the defence that an application of the causal inheritance principle to the case of mental causation entails the exclusion assumption. If a mental state M is a second-order functional state realised by a first-order physical state P, then the causal inheritance principle implies that the causal role of M is identical with that of P. Consequently, there cannot be two different causal pathways from M and P to the behaviour B.

This defence is not effective, I claim, because the causal inheritance principle on which it relies is, in fact, false. To see this consider our earlier example, illustrated in Figure 3. Let us suppose that the mental state R of having reasons for raising an arm is a second-order state realised by the first-order neural state NF of having certain neurons fire in the cortex. This is represented in the diagram below by R's inclusion of NF. But R is not coextensive with NF, because it can be realised by other states besides NF. The diagram makes it clear that the causal role of R is not identical with that of NF with respect to the intentional model. There is, relative to the most similar worlds generated by the intentional model, a counterfactual dependence between R and A, but no such dependence between NF and A. The diagram represents the situation in which

R œ->$_I$ A and ~R œ ->$_I$~A
NF œ->$_I$ A and ~(~NF œ ->$_I$~A)

So in the intentional model, the neural firing NF does not convey as much causal information as the mental state R. Indeed the neural state NF does not even make a difference to the effect relative to this model.

In order to support the exclusion assumption, the causal inheritance principle would have to hold in all models, and in particular the intentional model. Since the principle fails to hold for this model, it cannot offer any support for the exclusion assumption.

Consequently, given the crucial role this assumption plays in the Exclusion Argument, we have no reason for taking this argument to be compelling, either in its special application to mental causation or in its more general application to other explanatory domains.

BIBLIOGRAPHY

- Bishop, J. 1989: *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press**.**
- Cartwright, N. 1983: *How the Laws of Physics Lie*. Oxford University Press.
- Cartwright, N. 1999: *The Dappled World*. Cambridge University Press.
- Crane T. 1995: 'The mental causation debate', *Proceedings of Aristotelian Society*, Supplementary Volume.
- Davidson, D. 1963: 'Actions, reasons, and causes', *Journal of Philosophy*, 60. Reprinted in Davidson1980.
- Davidson, D. 1973: 'Freedom to act', in *Essays on Freedom of Action*, ed. T. Honderich. Routledge and Kegan Paul. Reprinted in Davidson 1980.
- Davidson, D. 1980: *Essays on Actions and Events*. Oxford University Press.
- Frankfurt, H. 1969: 'Alternate Posibilities and Moral Responsibility', *Journal of Philosophy*, 66, 829-39.
- Giere, R. 1988: *Explaining Science: A Cognitive Approach*. Chicago University Press.
- Jackson, F. and Pettit P. 1990: 'Program explanation: a general perspective'. *Analysis*, 50, 107-117.
- Jackson F. 1996: 'Mental causation', *Mind*, 105, 377-413.
- Joseph, G. 1980: 'The many sciences and the one world', *Journal of Philosophy*, 77, 773-90.
- Kim. J. 1984: 'Epiphenomenal and supervenient causation', *Midwest Studies in Philosophy*, 9, 257-270.
- Kim, J. 1989: 'Mechanism, purpose, and explanatory exclusion', *Philosophical Perspectives*, 3, 77-108.
- Kim, J. 1998: *Mind in a Physical World*. MIT Press.
- Lewis, D. 1966: 'An argument for the identity theory', *Journal of Philosophy*, 67, 427-46. Reprinted in Lewis 1993.

· Lewis, D. 1980: 'Mad pain and martian pain', in *Readings in Philosophy of Psychology*, Vol 1. ed N. Block. Reprinted in Lewis 1983.

· Lewis, D. 1983. *Philosophical Papers, Vol, 1*. Oxford University Press.

· Lewis, D. 1986: *Philosophical Papers, Vol. 2*. Oxford University Press.

· Lewis, D. 1999: 'Causation as influence', Whitehead lectures at Harvard.

· Menzies, P. 1988: 'Against causal reductionism', *Mind*, 98, 551-574

· Menzies, P. 1989: Probabilistic Causation and Causal Processes: A Critique of Lewis', *Philosophy of Science*, 56, 642-64.

· Menzies, P. 1996: 'Probabilistic causation and the pre-emption problem', *Mind*, 104, 85-117.

· Menzies, P. 1998: 'Are Humean doubts about singular causation justified?', *Communicaton and Cognition*, 31,1-26.

· Menzies, P. 1999: Intrinsic versus extrinsic conceptions of causation', forthcoming in *Laws and Causation: Australasian Studies in the History and Philosophy of Science*, ed. H. Sankey. Kluwer.

· Suppe, F. 1989: *The Semantic Conception of Theories and Scientific Realism*. University of Illinois Press.

Earlier versions of this paper were read at a Philosophy of Mind conference in Sydney in January 1999, at a conference on Australian Metaphysics in Grenoble in December 1999, and at a seminar in the Research School of Social Sciences, Australian National University in February 2000. For their comments and objections, I am indebted to many, but especially to Geoff Brennan, James Chase, Hugh Clapin, Michael Devitt, Frank Jackson, Daniel Nolan, Lloyd Reinhardt, Michael Smith, and Daniel Stoljar.