# *Probabilistic Causation and the Pre-emption Problem*

PETER MENZIES

Probabilistic theories of singular causation claim that singular causal relations can be analytically reduced to probabilistic relations.[1] Though they differ in detail, these theories depend on the guiding idea that a cause makes a difference to its effect by making it more probable than it would otherwise be in the circumstances.[2] Appropriately formulated, this idea applies uniformly to both deterministic and probabilistic causes.

I wish to argue that singular causation cannot be reduced to probabilistic relations. There is a problem—I call it the *pre-emption problem*—that shows that the guiding idea that a cause raises the probability of its effect does not accurately match our intuitions about singular causation. The pre-emption problem is illustrated by examples in which there are two processes leading to some effect, one of which goes to completion and brings about the effect, but in doing so cuts off or pre-empts the other process.[3] The significant feature of these examples is that the event that is the potential, but not actual, cause raises the probability of the effect, while the event that is the actual cause does not raise the probability of the effect at all. In §1 I describe one such example in the context of a very plausible way of spelling out the guiding idea of probabilistic theories.

An initially appealing response to the examples illustrating the pre-emption problem is to think that they can be overcome somehow by com-

[1] This assumption is made by the probabilistic theories of singular causation presented by Suppes (1970), Skyrms (1980), Lewis (1986a), Mellor (1991), and Eells (1991) among others.

[2] A notable exception to this claim is the probabilistic theory of Eells (1991). Eells argues that the idea that a cause raises the probability of its effect fits type-level (general) causation, but does not fit token-level (singular) causation. He thinks that the token-level causation should be explained in terms of probability trajectories rather than probability comparisons. Reasons of space preclude me from examining here Eells's unorthodox probabilistic theory of singular causation.

[3] Pre-emption examples of this kind are to be distinguished from examples of overdetermination. In the latter kind of example, there are two or more processes leading to the same effect, but all the processes go to completion, with no pre-emption of one process by another. These examples also pose serious difficulties for probabilistic theories of causation. But I have chosen to focus on the pre-emption cases because they tend to elicit particularly clear intuitions about causation.

Mind, Vol. 105 . 417 . January 1996 

plicating the analytic reduction of causation to probabilistic relations: perhaps the examples show that a cause need not raise the probability of its effect, but they do not show that it is impossible to reduce singular causal relations to probabilistic relations. In §2 I consider two theories which attempt to circumvent the pre-emption problem by complicating the guiding idea that a cause raises the probability of its effect. Despite their sophistication, these theories too are unsatisfactory, I argue, because they cannot deal with all the problem cases involving pre-emption.

In my view, the correct conclusion to draw from the failure of these theories is that singular causation cannot be analytically reduced to probabilistic relations: there is a residual feature of singular causation that cannot be captured solely by probabilistic relations.[4] The appropriate corrective to adopt, I believe, is to view the causal relation as a theoretical entity—like quarks, genes, and mental states. And it is to be defined in the same way as these other theoretical entities. In §3 and §4 I show how a standard account of the definition of theoretical entities can be applied to the case of singular causation. In §5 I modify this application of the standard account to accommodate the intuition that causal relations relate events in virtue of the properties exemplified by the events. In §6 I employ the resulting definition and identification of singular causation to solve the pre-emption problem. The final section considers some pressing objections to the proposed solution of the pre-emption problem.

In one sense, this new theory involves a radical departure from probabilistic theories in that it does not attempt to eliminate causal relations from ontology by reducing them to probabilistic relations. Nonetheless, the fundamental insight of probabilistic theories is not abandoned altogether, for the guiding idea of such theories is used as one of the central conditions in the theoretical definition of causation. The resulting theory of causation is a probabilistic theory in spirit, if not in letter.

## 1. The pre-emption problem

To explain the pre-emption problem it is useful to fix on one way of elucidating the idea that a cause makes its effect more probable than it would otherwise be in the circumstances. In my discussion I adopt David Lewis's suggested formulation of this idea, which avoids many of the shortcomings of other elucidations. It also forms the basis of Lewis's own probabilistic theory of causation, which is discussed in §2.

[4] Others have reached a similar conclusion for slightly different reasons: see Salmon (1984), Sober (1984, 1985), and Tooley (1987).



Lewis's formulation of the idea that a cause raises the probability of its effect starts with the following definition: *e probabilistically depends* on *c* if and only if (i) *c* and *e* are distinct events; and (ii) if *c* were to occur, the chance of *e*'s occurring would be *x*, and if *c* were not to occur, the chance of *e*'s occurring would be *y*, where *x* is much greater than *y*.[5]

This definition employs two notions which set Lewis's formulation of the guiding idea apart from others. The first is the notion of a probabilistic counterfactual, which is used in place of the more usual notion of conditional probability. Lewis thinks of these probabilistic counterfactuals as would-counterfactuals with consequents about the chance of the effect. They are to be interpreted so as to rule out so-called backtracking reasoning, according to which one reasons backwards in time to see what conditions must have obtained in order for the antecedent to be realised. Ignoring the possibility of backwards causation, one can state the intended non-backtracking interpretation as follows: to evaluate the truth-value in the actual world of a given counterfactual with a false antecedent that refers to a time *t* (call this the antecedent's *reference time*), hold fixed the history of the actual world more or less up to the reference time *t*, change the state of the actual world at the time *t* no more than is needed to make the antecedent true, and then consider whether the laws of nature, operating on the minimally changed state at *t*, predict states which make the consequent true.[6] The second notion that sets Lewis's formulation apart from other formulations is the notion of objective, single-case chance: the probabilistic counterfactuals which define probabilistic dependence concern the single-case chance of the effect. The objective character of chance distinguishes it from various kinds of epistemic probability; and its single-case character distinguishes it from finite and limiting relative frequencies.[7] A feature of chance which should be noted here is its time-dependence: the chance of some event can vary with time before the event occurs. The chances that appear in the probabilistic counterfactuals that define probabilistic dependence are to be understood,

[5] See Lewis (1986a, pp. 176–7). The first counterfactual in clause (i) may seem puzzling as its antecedent is true. In Lewis's view, however, a counterfactual need not have a false antecedent: a counterfactual with a true antecedent is true if its consequent is true as well.

[6] For more details of Lewis's account of the intended interpretation of counterfactuals see Lewis (1986a, pp. 33–66). The use of non-backtracking counterfactuals in the definition of probabilistic dependence overcomes two problems that arise with standard definitions in terms of conditional probabilities: a probabilistic dependence does not hold in cases where *c* and *e* are joint effects of a common cause; and the probabilistic counterfactuals are well-defined even when the probability of their antecedents is zero. See Lewis (1986a, pp. 178–9).

[7] For Lewis's account of chance see his (1986a, pp. 83–113).

Lewis says (1986a, pp. 176–7), as referring to a time immediately after the cause occurs.[8]

Lewis's definition of probabilistic dependence is immune to many of the difficulties plaguing probabilistic theories. But it is not immune to the pre-emption problem. To understand this problem consider the following example, which is a modification of some deterministic examples described by Lewis (1986a, pp. 200–1). Figure 1 depicts a system of neurons.

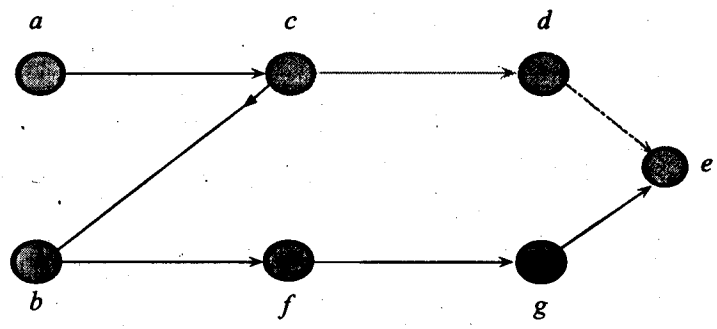


Figure 1

In this system one neuron may fire and stimulate another by way of a stimulatory axon (depicted by a forwards arrow), or it may fire and inhibit another by an inhibitory axon (depicted by a backwards arrow). A neuron that is both stimulated and inhibited does not fire. In Lewis's examples the causal connections are deterministic, but in this example the connections are to be probabilistic. Suppose that the stimulatory axons connecting neuron *a* with neuron *e* are extremely reliable, so if *a* were to fire it would be very probable that all the intermediate neurons and neuron *e* would fire. Suppose also the stimulatory axons connecting neuron *b* with neuron *e* are much less reliable, so if *b* were to fire it would be very improbable that all the intermediate neurons and neuron *e* would fire. Finally, suppose that the inhibitory axon between *b* and *c* is moderately reliable, so if *b* were to fire, it would be moderately probable that *c* would *not* fire.

The following events take place in this system of neurons. Neurons *a* and *b* fire at the same time. The unreliable process that begins with *b*'s firing very improbably goes to completion so that *e* fires. On the other hand, the reliable process that begins with *a*'s firing is cut short, because *b*'s firing just happens to inhibit neuron *c* from firing on this occasion. (The dotted lines in the figure represent those neurons that never fire.)

[8] I argued for this claim originally in my (1989a), in which the example illustrating the pre-emption problem was first described.



The example shows that there can be causation without probabilistic dependence. In the example, *b*'s firing caused *e* to fire, but it did not increase the chance of *e*'s firing in the circumstances. In fact, it lowered the chance of *e*'s firing, because it made it fairly likely that the more reliable process, which had been started by *a*'s firing, would be cut short. In the circumstances in which *a* had fired and started the reliable process, *b*'s firing made *e*'s firing less likely than it would otherwise have been because it made it more likely that *c* would be inhibited from firing, so pre-empting the reliable process. In this case the causal connection between *b*'s firing and *e*'s firing is not matched by a corresponding probabilistic dependence.

The example also demonstrates the converse fact that there can be probabilistic dependence without causation. In the example, *a*'s firing did not cause *e* to fire even though it increased the chance of *e*'s firing in the circumstances. *a*'s firing increased the chance of *e*'s firing as it started the reliable process which had a fair chance in the circumstances of going to completion and causing *e* to fire. It is true that these circumstances included the fact that *b* had fired as well and had some chance of pre-empting the reliable process. (While the circumstances include the fact that *b* fired, they do not include the fact that *b* prevented *c* from firing or the fact that b caused *e* to fire, since these facts concern events occurring after the time of *a*'s occurrence.) Nonetheless, even taking these circumstances into account, the chances of *e*'s firing, given that *a* fired, were moderate. In contrast, if *a* had not fired, the reliable process would not have been initiated, leaving the unreliable process as the only way *e*'s firing could have occurred. But this would have meant that the chances of *e*'s firing would have been very small indeed. Accordingly, even in the circumstances in which *b* fired at the same time as *a* did, the chance of *e*'s firing was greater given that *a* fired than it would have been if *a* had not fired. So there is a probabilistic dependence between *a*'s firing and *e*'s firing, but it does not correspond to any causal connection between them.

The example demonstrates in one fell swoop that probabilistic dependence is neither necessary nor sufficient for singular causation.[9] The problem posed by the example does not turn on anything in this particular definition of probabilistic dependence. It can be shown more generally that any theory that identifies a causal relation between two events with a relation of probability raising between them is inadequate to our intuitions about causation because of the problems posed by examples of pre-emption.

[9] As we shall see in the next section, Lewis's probabilistic theory of causation states that probabilistic dependence is sufficient but not necessary for singular causation. Mellor's (1991) probabilistic theory of causation seems to be committed to the claim that something like probabilistic dependence is necessary and sufficient for causation.

Is there any way of defending the inherently plausible idea that a cause increases the chance of its effects against counterexamples such as the one above? Ultimately, I do not think there is, but I pause here to rebut some commonly suggested defences against the counterexample. Following Lewis, let us say an event is *fragile* if, or to the extent that, it could not occur at a different time or in a different manner.[10] It is often suggested that if one treats the cause and effect events as fragile events, one can save the guiding idea of probabilistic theories from the pre-emption problem. Suppose that the way and the time that *e* fires depends on which process causes it. Suppose, for example, that it fires vigorously and early if stimulated by the reliable process, and feebly and late if stimulated by the unreliable process. When *e*'s firing is taken as a fragile event—as an event which could not have occurred other than as a late, feeble firing—then it ceases to be true that *a*'s firing increased the chance of *e*'s firing while *b*'s firing lowered it. For *a*'s firing made it less likely that *e* would fire *feebly and late* while *b*'s firing made it more likely. The probabilistic relations of this example match the causal relations in the right way.

Phrased like this, the defence is not compelling. For it can simply be stipulated, as part of the example, that neuron *e* will fire in exactly the same manner and at exactly the same time whichever process causes it. Imagine, then, that things turn out as before: the unreliable process pre-empts the reliable process and improbably goes to completion, causing *e* to fire—but to fire feebly and late. It is entirely conceivable that this very same event—the event of *e*'s firing feebly and late—could have been caused by the reliable process. So there is no contradiction in thinking that *a*'s firing increased, and *b*'s firing lowered, the chance of this very event. Even when the effect is taken to be an extremely fragile event, it seems that a pre-empted potential cause may increase, and a pre-empting actual cause may decrease, its chance of occurring.

However, the strategy of taking the cause and effect to be fragile events can be given a slightly different, and more appealing, twist. Consider the thesis, first enunciated by Davidson,[11] that events are to be individuated in terms of their causes and effects. Suppose we understand this as implying that events cannot occur except as causally related to the events that are their actual causes and effects. Endorsing Davidson's thesis, on this understanding of it, amounts to taking events to be fragile in a special way. And it amounts to rejecting the key assumption of the counterexample that *e*'s firing could occur in exactly the same way whichever process brought it about. Understood in this way, the thesis rules out the possibility that *e*'s

[10] Lewis (1986a, p. 196). Lewis does not endorse the strategy of taking the effect event to be fragile: see his discussion on pp. 196–9.

[11] See Davidson's 'The Individuation of Events', in his (1980, pp. 163–80).



firing, though actually caused by *b*'s firing, could have been caused by *a*'s firing: for the causal criterion of identity implies that these different causes would individuate different events. By taking events to be fragile in this special way—so that they cannot have different causes and effects from the ones they actually have—one can reject the counterexample from the very outset.

However, it is essential to distinguish two readings of Davidson's causal criterion of identity: a weak, plausible reading and a strong, implausible reading. On the weak reading, the criterion is an *intraworld* criterion of identity, a criterion applying to the individuation of events within a world, primarily the actual world. The criterion on this reading states that no two events in the actual world have the same causes and effects. On the strong reading, the criterion is an *interworld* criterion of identity, a criterion applying to the individuation of events across worlds. The criterion on this reading states that an event cannot have one set of causes and effects in one world and a different set in another world.

The weak reading of causal criterion has much to commend it. Indeed, there is good reason for thinking that it is the reading that Davidson intends for the criterion.[12] But it is not strong enough for the defence of the idea that a cause increases the chance of its effect. For one could accept that the effect event in the example is distinguished from other events in the actual worlds by its causes and effects, but still ask whether that very same event in another world could have been brought about by some other cause. The defence against the counterexample requires the strong reading of the criterion as an interworld criterion of identity. On this reading, it would be absurd to ask whether the same event could have a different cause in another world from the cause it has in the actual world. However, this reading of the causal criterion is implausibly stringent in that it rules out questions that seem to be perfectly intelligible. The extreme implausibility of this reading is marked by the fact that it makes causal relations necessary. If events are individuated so that they cannot have different causes or effects from the ones they actually have, it follows that if a causal relation exists between events, it holds necessarily: for any world in which those events occur must be ones where they are related as cause and effect. This conflicts with the almost universal view that causal relations are contingent. This cost of the defence against the counterexample is excessive.

[12] In 'The Individuation of Events', Davidson formulates the causal criterion for the identity of events in terms of material equivalences. He states that an event *x* is identical to an event *y* iff for any event *z*, *z* caused *x* is materially equivalent to *z* caused *y* and *x* caused *z* is materially equivalent to *y* caused *z*. This indicates that he thinks of the causal criterion as intraworld criterion for individuating events in the actual world. See Davidson (1980, p. 179).

## 2. Two probabilistic theories

Although I have presented the pre-emption problem in terms of probabilistic dependence, as defined by Lewis, Lewis's probabilistic theory of causation is not committed to the claim that causation is the same thing as probabilistic dependence.[13] Lewis recognises the fact that there can be causation without probabilistic dependence. Indeed, his theory accommodates this fact by defining causation, not as probabilistic dependence, but in terms of the ancestral of this relation.

Lewis's definitions run as follows: a finite sequence of events $<a,b,c\ldots>$ is *a chain of probabilistic dependences* if and only if $b$ probabilistically depends on $a$, $c$ probabilistically depends on $b$, and so on. An event $a$ is a cause of an event $e$ if and only if there is a chain of one or more probabilistic dependences running from $a$ to $e$.

This theory deftly handles the fact that there can be causation without probabilistic dependence. In the example described in the last section, $b$'s firing caused $e$'s firing though there is not a probabilistic dependence between them. For there is a chain of probabilistic dependences connecting the two events. $g$'s firing probabilistically depends on $b$'s firing, because $g$'s firing belongs to the unreliable sequence of firings which would not have occurred if $b$ had not fired; and $e$'s firing probabilistically depends on $g$'s firing, because by the time $g$ fired the more reliable process had been cut off and so $e$'s firing would not have occurred if $g$ had not fired. This two-step chain of probabilistic dependences justifies, from the point of view of the theory, the claim that $b$'s firing caused $e$'s firing.

However, the important point to make about Lewis's theory is that while it recognises that there can be causation without probabilistic dependence, it fails to recognise that there can be probabilistic dependence without causation. In the example of the last section, $e$'s firing probabilistically depends on $a$'s firing, so we have a one-step chain of probabilistic dependences connecting the two events. Hence, by the theory, $a$'s firing counts as a cause of $e$'s firing. But this conflicts with the intuition that this is a spurious probabilistic dependence that does not correspond to a genuine causal relation.

Why do we judge that $a$'s firing does not cause $e$'s firing? Intuitively, the reason is that the process leading from $a$'s firing is broken or cut short before it reaches neuron $e$. Lewis's theory does not deliver the right result

[13] Lewis's probabilistic theory of causation is described in his (1986a, pp. 175–84). It is a generalisation of his earlier counterfactual theory of causation, which covered some but not all cases of probabilistic causation. For the earlier theory see his (1986a, pp. 159–72).



about $a$'s causal status because it does not adequately capture the intuitive idea that causation requires an unbroken process to link cause and effect. Is it not possible, though, to construct some more sophisticated version of Lewis's theory that does adequately capture this idea? This is the thought I expressed in an earlier paper (Menzies 1989a) in which I tried to show how a probabilistic theory, based on Lewis's, could capture the idea that causation requires the existence of an unbroken process connecting cause with effect.

The basic idea of this amended probabilistic theory is to define the concept of a causal process in terms of probabilistic relations and then to use this notion to define causation. The fundamental definition of a causal process runs as follows: there is *an unbroken causal process* running from event $c$ to event $e$ if and only if, for any finite sequence of $n$ ($n \geq 0$) times $<t_1,t_2,\ldots,t_n>$ between the time $c$ occurs and the time $e$ occurs, there is a sequence of actual events $<x_1,x_2,\ldots,x_n>$ occurring at these times respectively such that $<c,x_1,x_2,\ldots,x_n,e>$ constitutes a chain of probabilistic dependences. Then, as a first approximation to a definition of causation, I suggested that an event $c$ is *a cause* of an event $e$ if and only if there is an unbroken causal process linking $c$ with $e$. Informally, the idea is that if $c$ is a cause of $e$, then, whichever times we select between the time $c$ occurs and the time $e$ occurs, we can find corresponding actual events which form a chain of probabilistic dependences connecting $c$ with $e$. This suggestion has the virtue of disallowing neuron $a$'s firing as a cause of neuron $e$'s firing: since $b$'s firing cuts short the reliable process by preventing neuron $c$ from firing, there are no actual events occurring after $c$ has failed to fire to link up with the final effect in a chain of probabilistic dependences.

As it stands, the theory is not quite correct. In its present form, it does not deliver the result that $b$'s firing is the cause of $e$'s firing. The condition on an unbroken causal process is stated in terms of a universal quantification over finite sequences of times occurring between the time of the cause and the time of the effect, and included in the domain of this quantification is the empty sequence of times. The condition states that every such finite sequence of times has to be such that events occurring at those times form an appropriate chain of probabilistic dependences. As applied to our example, the condition on an unbroken causal process requires that, for the special case of the empty sequence of times between the time $b$ fires and the time $e$ fires, there should be a probabilistic dependence between $b$'s firing and $e$'s firing. But we have seen that this is not the case. Therefore, one condition for the existence of an unbroken causal process running from $b$'s firing to $e$'s firing is not met.

It is not possible to dispose of this problem by eliminating the empty sequence of times from the domain of quantification without complicating the theory in other places. In my (1989a) paper I chose to deal with this problem by further complicating the analysis of causation: I suggested defining causation in terms of the ancestral of the notion of an unbroken causal process in the same way that Lewis defined causation in terms of the ancestral of the relation of probabilistic dependence. This was done as follows: a finite sequence of events <*a*,*b*,*c*...> is *a chain of unbroken causal processes* if and only if there is an unbroken causal process running from *a* to *b*, unbroken causal process running from *b* to *c*, and so on; an event *c* is *a cause* of an event *e* if and only if there is a chain of unbroken causal processes running from *c* to *e*. This handles the difficulty about the example because it is reasonable to think that *b*'s firing is connected to *e*'s firing by a chain of unbroken causal process even if not by a single such process.

I now think that this amended probabilistic theory of causation is unsatisfactory for a number of reasons. One reason—perhaps a less than conclusive reason in some eyes—has to do with the fact that theory implies the *a priori* impossibility of action at a temporal distance. The theory rules out action at a temporal distance by virtue of the fact that it requires that there be a continuous unbroken causal process linking cause with effect.[14] That there is no action at a distance in the actual world is made plausible not just by the force of commonsense intuitions but also by the appeal of field theories in physics. However, it is one thing to say that there is no action at a distance in the actual world, quite another thing to claim that action at a temporal distance is a conceptual impossibility. The theory is far too strong in ruling it out on *a priori* grounds. It is possible, after all, to describe consistent thought experiments involving action at a temporal distance. There is no contradiction in supposing that a certain event *c* could increase the chance of a temporally removed event *e* without there being an unbroken causal process running between them.

Another reason for my dissatisfaction with the amended theory—and this is the more conclusive reason—is that there is a relatively simple kind of case which poses a serious difficulty for it. This kind of case is illustrated by an example of Lewis's, which differs from the earlier neuron example in three respects: first, the causal connections between the

[14] The theory in question does allow, as degenerate cases, instances where cause and effect are instantaneous events occurring at the same time and instances where cause and effect are events with partly, or completely, overlapping temporal durations. See my (1989a, pp. 657–9). But these are not cases of causal action at a distance—or gappy causation—in which cause and effect are spatiotemporally distant events with no intervening process connecting them.



firings of the neurons are deterministic; secondly, the main process going from cause to effect runs more quickly than the pre-empted alternative process; and thirdly, the alternative process is cut off, not by a branch process diverging from the main process, but by the effect itself after the main process has gone to completion.[15] Figure 2 depicts the new example.

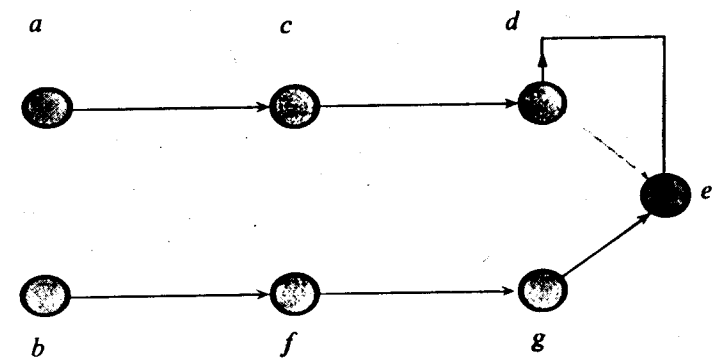


Figure 2

Lewis actually presents the example as a problem for his restricted counterfactual analysis of causation. But it is also a problem for the analyses presented in this section—his generalised probabilistic analysis and my emendation of it.[16] The problem lies in the fact that in this example, unlike the earlier one, there is no chain of probabilistic dependences running from *b*'s firing to *e*'s firing. Suppose, for example, that we try to establish a two-link chain of probabilistic dependences with *g*'s firing as our intermediate event. *g*'s firing probabilistically depends on *b*'s firing straightforwardly, because it is part of the main process which would have had no chance of getting underway if *b* had not fired. But the catch is that *e*'s firing does not probabilistically depend on *g*'s firing, since the chance of *e*'s firing would have been 1 whether or not *g* had fired: for if *g* had not fired, the alternative process would not have been cut short and would have ensured *e*'s firing in any case. No matter how many links we choose for our chain, the same problem

[15] For the example see Lewis's discussion of late pre-emption (1986a, pp. 203–4).

[16] More generally, it is a problem for any counterfactual theory of causation or causal explanation that tries to deal with the pre-emption problem by requiring that there be a chain, however complicated, of counterfactual dependences linking cause with effect. For one such theory of causal explanation see Ruben (1994).

will arise in trying to establish the final probabilistic dependence in the chain.[17]

## 3. The theoretical definition of causation

Neither theory examined in the preceding section is successful in reducing causation to probabilistic relations. It is unlikely, in my view, that any attempt along similar lines will succeed. For there is, I think, a residual element in our concept of causation which resists capture solely in terms of probabilistic concepts. This residual element consists in the idea that the causal relation is something which, in some sense, underlies and supports probabilistic relations: the causal relation is a deeper structural feature of reality which explains the more superficial probabilistic relations. How exactly is this residual element to be explained more precisely?

We can provide a more precise explanation by treating the concept of causation as the concept of a theoretical entity. We can explain the concept, then, in the same manner as the concepts of other theoretical entities such as quarks, genes, and species. There are a number of different accounts of the way in which concepts of such theoretical entities can be defined. To my mind the best account is David Lewis's refinement of the ideas of Ramsey and Carnap.[18] In this section of the paper I outline

[17] I had chosen to set this problem aside in my (1989a) article because I had thought that it could be treated by invoking Kim's (1975) conception of events, according to which the time at which an event occurs is an essential or constitutive feature of the event: a neuron's firing at $t$ and its firing at $t'$ are different events. This is an important point in the present context because the alternative process runs more slowly than the main process so that, if the alternative process had not been cut short, $e$'s firing would have been delayed. Thus, if $b$ had not fired, $e$ would have had no chance of firing at the specific time that it did, in which case $b$'s firing, after all, increased the chance of $e$'s firing, considered as a particular event with a specific time of occurrence.

Now I find this way of dealing with this kind of counterexample to be unsatisfyingly *ad hoc*. Clearly, we discount $a$'s firing and allow $b$'s firing as the cause of $e$'s firing in this example for the same reason as we discount $a$'s firing and allow $b$'s firing as a cause of $e$'s firing in the first neuron example. The correct analysis of causation should reflect the fact that the second example is similar in this respect to the first example, rather than appeal to a special feature of the second example to wriggle out of the difficulty.

In any case, Dowe (1993) has constructed a counterexample to my (1989a) theory which cannot be handled by treating the effect as a particular event with a specific time of occurrence. The counterexample is a case of overdetermination in which two events initiate deterministic processes which separately bring about the same effect: my (1989a) theory implies, contrary to intuition, that neither event is a cause of the effect, since neither is linked by a chain of unbroken processes to the effect.

[18] Lewis's account is set out in his (1972), and his (1983a, pp. 78-95). Also see the classic works by Ramsey (1931, pp. 211–36) and Carnap (1966, pp. 958–66).



Lewis's account of theoretical definition by way of its best-known application to the concepts of mental states and show how a similar theoretical definition can be framed for the concept of singular causation. This definition will elucidate the residual element of the concept that is not successfully captured by probabilistic theories. The definition presented in this section will be modified in §5.

How are the meanings of terms denoting mental states to be explained? Lewis argues that the meanings are to be explained in terms of the role they play in a mature folk psychology. One can think of folk psychology as a kind of theory consisting of platitudes that are common, albeit tacit, knowledge among us. The platitudes concern the causal relations between mental states, sensory stimuli, and behavioural responses: they say how certain perceptual stimuli and mental states, individually or in combination, typically cause certain mental states; and how certain mental states, individually or in combination, typically cause certain behavioural responses. Suppose all the platitudes concerning the causal roles of mental states are conjoined into a single sentence—call it *the postulate of folk psychology*. Then it is a simple matter to give a recipe for analysing mental state terms. Let's take as our example the mental state term "pain". The term is implicitly defined by the postulate of folk psychology, which says that pain is a state which occupies a certain typical causal role—call it the pain-role $R$. One can convert this implicit definition into an explicit one by analysing "pain" as "the state which typically occupies the pain-role $R$".[19]

The first step in providing a similar definition of the causal relation is to set down the central tenets of our folk theory of causation—the platitudes about causation which are common knowledge among us. There are many such platitudes: for example, it typically coincides with a temporal ordering of events so that causes precede effects; it typically coincides with the means-end relation so that if an effect is an end, its causes are means to it; causes explain their effects.[20] The *postulate of the folk theory of causation* will consist of a conjunction of all such platitudes; or better, a long disjunction of all conjunctions of most of the platitudes. The most important platitudes—the ones which are crucial to the concept of causation—will be elements in all the conjunctions in this long disjunction. In the subsequent discussion I concentrate on a simple formulation of the postulate of the folk theory of causation, a formulation which takes the postulate to be a conjunction of three crucial platitudes.

[19] More accurately, since the causal role of pain involves other mental states, "pain" must be interdefined with other mental state terms in a kind of package deal. For details see Lewis (1972) and (1983a, pp. 78–95).

[20] Mellor calls these the connotations of causation (1991, p. 226): they represent the kinds of platitude figuring prominently in the folk theory of causation.

The *first crucial platitude* is that the causal relation is a relation holding between distinct events.[21] Although the concept of distinctness between events is a technical one, I take it that it is a familiar one: there are certain dependence relations that can hold between events, the existence of which mean that one event is, in some sense, part of the other. One non-causal kind of dependence occurs when an event depends on a more specific event that implies it: for example, a ball-bearing's weighing more than 5 grams depends on the more specific event of its weighing 10 grams. Another dependence of this kind occurs when an event characterised in terms of a functional role depends on the event that occupies the role: for example, a person's immunity to infection is a dispositional state that is implicitly understood in terms of a functional role and as such depends on the specific event that occupies this role, in this case the event of the person's having antibodies to the infection in his bloodstream. Yet another kind of non-causal dependence occurs when one event is a proper part of another event: for example, the event of my writing the letters "Larr" depends on the event of my writing the word "Larry". A final kind of non-causal dependence relation occurs when an extrinsically characterised event depends on an intrinsically characterised one: for example, the extrinsically characterised event of Xanthippe's being widowed depends in this way on the more intrinsically characterised event of Socrates' dying. When two events are related by one or other of these kinds of dependence, they are not distinct from one another. Hence, to say that two events are distinct is to say that none of these dependence relations holds between them.[22]

The *second crucial platitude* is that the causal relation is an intrinsic relation between events. By this I mean roughly that it is a relation between events determined by the intrinsic properties of the events and of what goes on between them. In taking this to be a platitude of the folk theory of causation, I assuming that the folk theory simply contradicts Hume: the commonsense conception presupposes that when events are causally related, there is some connection between them which is determined by the intrinsic nature of the events and of the local spatiotemporal region containing them: the causal relation is not an extrinsic relation depending on a regularity, or any other pattern of events happening outside the local spatio-temporal region of the causally related events. This platitude is particularly important in guiding judgements about cases of pre-emption. For it implies that if a causal relation exists between events, this causal relation is not affected by the existence of other alternative processes. For example, if there is a causal relation between *b*'s firing and *e*'s firing in the neuron examples described above, this causal relation would still hold whether or not there is an alternative process which could bring about the same effect. In this context, the platitude expresses the idea that a causal relation is independent of the presence of alternative causes.[23]

[21] In saying that the causal relation relates events, I do not mean to imply that the events in question are always changes: the term "event", as I use it here, is meant to include changeless states of affairs, facts, and even omissions.

[22] For further discussion of the four kinds of non-causal dependence that can exist between events see my (1988). The fundamental intuition expressed by the platitude that the causal relation relates distinct events is the intuition that the causal relation is a dependence relation of a fundamentally different kind from the relations just enumerated.



What counts as an intrinsic relation? This question deserves some discussion in view of the central role that the notion will play in the definition of causation. It is possible to give a philosophical explication of the concept of an intrinsic relation in terms of a special conception of properties and relations—the *natural properties and relations*, as they are called by Lewis (1983b and 1986b). Natural properties and relations are sparse in number: there are only as many as required to characterise the way things are in reality comprehensively and without redundancy. These properties and relations do not correspond in any fashion to the predicates of any natural language. They are highly specific and not at all gruesome or disjunctive in character: they carve nature at its joints so that sharing them makes for true similarity. If physicalism is true, then physics, in some ideally completed form, gives us an inventory of the natural properties and relations instantiated in the actual world. But these are not all the natural properties and relations that there are. There may be worlds where physics is different or nonphysicalist worlds where the fundamental properties and relations are not physical at all.

The notion of an intrinsic relation can be explicated in terms of this conception of natural properties and relations. It turns out that relations can be intrinsic in two ways. A relation is intrinsic in the first way if it supervenes on the natural properties of its relata. More precisely, let us say that two objects are *duplicates* just in case they have exactly the same natural properties. Let us also say a relation is *intrinsic to its relata* if and only if, whenever *a* and *a'* are duplicates and *b* and *b'* are duplicates, then both or neither of the pairs <*a*, *b*> and <*a'*, *b'*> stand in the relation. The traditional internal relations, such as similarity and difference in intrinsic respects, are intrinsic in this way. There is, however, a different way in which a relation can be intrinsic that applies to external relations such as spatiotemporal distance and causation. A relation is intrinsic in this way if it supervenes, not just on the natural properties of its relata, but on these taken in conjunction with the natural relations holding between them. Let

[23] The formulation of this second platitude owes something to Bigelow and Pargetter (1990, pp. 271–5), which emphasizes the local, intrinsic character of the causal relation in opposition to regularity and probabilistic theories of causation.

us say that <$a$, $b$> and <$a'$, $b'$> are *duplicate pairs* if and only if $a$ and $a'$ have the same natural properties, and so do $b$ and $b'$, and also the natural relations between $a$ and $b$ are exactly those between $a'$ and $b'$. Then let us say that a relation is *intrinsic to its pairs* if and only if, whenever <$a$, $b$> and <$a'$, $b'$> are duplicate pairs, then both or neither of the pairs must stand in the relation. Relations which are neither intrinsic to their relata nor to their pairs are *extrinsic relations*. They correspond to relations such as having the same owner which in some sense involve extraneous objects besides the events that are their relata.

In stating the platitude that causation is an intrinsic relation, I mean more precisely that it is a relation intrinsic to its pairs. It is not a relation that is intrinsic to its relata, since the causal relation does not supervene on the natural properties of its relata alone. But it does supervene on the natural properties of its relata, when they are taken in conjunction with the natural relations holding between them. In saying that causation is a relation intrinsic to its pair, I mean to imply that it is not an extrinsic relation, as it would be under a Humean conception. For Humean theories make causation an extrinsic phenomenon to the extent that they take causal relations to depend on features of reality extraneous to the causally related events, in particular on regularities holding globally throughout the world. The distinctive mark of our intuitive conception of causation, I claim, is that it takes causal relations to be determined by the natural properties of the relata and the natural relations holding between them, taken in isolation from everything else happening in the world.

The *third crucial platitude* that will play a role in the postulate for the folk theory of causation relates the causal relation to the relation that exists between two distinct events when one event increases the chance of the other. As we have seen, the causal relation does not always coincide with this relation of increase of chance between distinct events. Nonetheless, it is a striking fact that the two relations coincide for the most part. Aside from cases involving pre-emption and overdetermination, one event causes another event just when the two events are distinct and the first event increases the chance of the second event.[24] It is striking how a probabilistic theory of causation, appropriately formulated, follows the contours of our intuitive conception of causation with amazing accuracy—with the exception of the problem cases to do with pre-emption and overdetermination.

[24] Randomly occurring events also constitute exceptions to this generalisation: it may be that two events are distinct, one increases the chance of the other, but there is no causal relation between them because the event whose chance is increased is a purely random occurrence. This kind of case should be handled, in my opinion, in much the same way as the difficult cases of pre-emption.



These three crucial platitudes, then, will be taken to form the basis of the postulate of the folk theory of causation.[25] On the basis of this postulate, we can frame an explicit definition of the causal relation. The definition runs: the *causal relation is the intrinsic relation that typically holds between two distinct events when one increases the chance of the other event.*[26] (The notion of increase in chance, used here, is to be understood in the counterfactual manner elucidated in Lewis's theory of §1.)[27]

[25] Is it a problem that the postulate of the folk theory of causation, in contrast to the postulate of folk psychology, does not define the causal relation in terms of its causal role? It is, indeed, true that Lewis says that mental states, and theoretical entities more generally, are to be defined in terms of their causal role. (See his 1972, p. 249.) But this seems to me to overgeneralise a special feature of the mental state case. I cannot see anything in Lewis's style of theoretical definition which would make its applicability depend on the platitudes taking this particular form. In supposing that the postulate of a theory need not characterise theoretical entities in terms of their causal role, I follow the lead of Tooley (1987), who adopts the Lewis-style approach to define causation, though in somewhat different terms from the definition offered here, and Jackson (1992), who uses the approach to define values.

[26] There are systematic similarities between this definition of causation and a definition of causation briefly discussed by Lewis (1986a, p. 206–8). Lewis canvasses the possibility of defining causation in terms of a notion of quasi-dependence as a solution to the difficulties that examples of late deterministic pre-emption pose for his original counterfactual theory of causation. According to Lewis, one event *quasi-depends* on another if and only if they are connected by a process which in its intrinsic character is like processes occurring in the same world or other worlds governed by the same laws of nature, where the majority of these processes exhibit a pattern of counterfactual dependence. Using this notion of quasi-dependence, he suggests the following definition of causation: one event causes a distinct event if and only if there is a chain of stepwise counterfactual dependences or quasi-dependences between them. The common idea behind Lewis's definition of causation in terms of quasi-dependence and the definition suggested above is that causation is defined—partly in the former case and wholly in the latter case— in terms of what happens in certain typical cases which exhibit an appropriate counterfactual or probabilistic dependence. There are, of course, obvious differences between the definitions. One is that Lewis's definition is framed to handle deterministic rather than probabilistic causation. Another is that, even when Lewis's definition is modified in the obvious way to handle probabilistic causation, it does not handle all the problem cases. For example, modified in the obvious way to handle the probabilistic case, it says that in the first neuron example $a$'s firing caused $e$'s firing because there is a probabilistic dependence between them, even though not a quasi-dependence. Most importantly, Lewis's definition of causation does not use the "platitudes" approach to defining theoretical terms.

[27] An important objection: isn't this definition unsatisfactory because it is satisfied by the relation of being-an-increase-in-chance-between-distinct-events, which we know cannot be the causal relation? The answer to this question is that the relation of being-an-increase-in-chance-between-distinct-events does not completely satisfy the definition, for it is not an intrinsic relation in the intended sense, viz. a relation intrinsic to its pairs. The relation is properly understood, I would argue, as supervening on global patterns of relative frequencies. (I am grateful to David Lewis for pointing out this objection to me; and suggesting a way to answer it.)

## 4. A comparison with the theoretical definition of mental states

It is instructive to compare this definition of the causal relation with the definition of a mental state term, say the term for pain: pain is the state that typically occupies the pain role *R*.

One similarity between the definitions is obvious. In each case, the theoretical entity—the pain state or the causal relation—is defined in terms of a characteristic phenomenon—a causal role or an increase in chance between distinct events—but one distinct from the theoretical entity itself. Even in cases where a person's realisation of the pain role signifies that he is in pain, the pain is something over and above the pain role; and likewise even in cases where one event increases the chance of a distinct event, and this signifies a causal relation between the events, the causal relation is something distinct from the increase-in-chance-between-distinct-events relation. The distinctness of the theoretical entity from the characteristic phenomenon is emphasized by the fact that both definitions allow that the theoretical entity may be present even though the characteristic phenomenon does not occur and that it may be absent when the characteristic phenomenon does occur. The reason that such divergences are permitted is that both definitions identify the theoretical entity in question with the entity which *typically*, but not *invariably*, accompanies the realisation of the characteristic phenomenon.

The presence of the "typically" modifier in the definition of pain implies that a person is in a mental state, not because he is in a state which occupies the pain role in him, but because he is in a state that typically occupies the pain role for the rest of us. These two states need not be the same, as is illustrated by Lewis' example of the madman (1983a, pp. 122–32). This person is in a state which for the rest of us is caused by cuts, burns, and pressure and causes groans and writhing. But in the madman, the state is caused by moderate exercise on an empty stomach and causes him to cross his legs and snap his fingers. In short, the state has a causal role in him that differs from the causal role it typically has for the rest of us. Nonetheless, Lewis's account correctly predicts that we would say that he is in pain, since he is in the state that *typically* occupies the causal role of pain. The modifier "typically" in the definition of causation works in the same way. Its presence in the definition implies that a causal relation can hold between distinct events even when one event does not increase the chance of the other because they are related by the intrinsic relation that *typically* coincides with the chance-increase-between-distinct-events relation. As we shall see in §6, the causal relations in the pre-emption cases discussed earlier conform to this pattern.



This is a good place to try to forestall a misconception about the role of the "typically" modifier. It might be thought that the introduction of the modifier in the definition of causation is intended to deal with the problematic cases of pre-emption by excluding them from consideration on the grounds that they are atypical. If so, the definition of causation is trivial: any definition can be made watertight against counterexample simply by the introduction of the "typically" modifier to exclude the atypical cases.

This objection rests, however, on a misconception of the role of the "typically" modifier in the definition. Its role is not intended to exclude the problem cases from consideration: an obvious defect of that strategy would be that it leaves unexplained our causal intuitions about the excluded cases. Rather the modifier is intended to define the causal relation as the underlying relation that typically holds when a characteristic phenomenon is present. The presence of this modifier no more trivialises the definition of causation than it does the definition of the mental state of pain. It would be a mistake to criticise Lewis's definition of pain as the state that typically occupies the pain-role on the grounds that it excludes the madman's pain as atypical. Part of the point of the definition is to explain why the madman's pain does count as pain even though he does not manifest the characteristic phenomenon of pain behaviour. Similarly, part of the point of the definition of causation is to explain how there can be causal relations even when the characteristic phenomenon of causation—one event's increasing the chance of another event—is not present.

A second similarity between the proposed theory of causation and Lewis's theory of mental states is that both allow in the same way for *a posteriori* identifications of the theoretical entities they define. Lewis's theory of mental states provides a model for how such an identification can be made. As we have seen, the theory defines the mental state of pain in terms of the typical occupant of the causal role *R*. Now suppose—this is not an outlandish supposition in the light of what we know at present—that neurophysiology discovers that the occupant of the causal role is a certain neural state *N*. Then it is easy to see how pain might be identified with this neural state by way of the following inference: pain is the typical occupant of the causal role *R*; neural state *N* is the typical occupant of the causal role *R*; therefore, pain is the neural state *N*. The truth of the two premisses guarantees the truth of the psychophysical identification. The two premisses differ, however, with respect to the status of their putative truth: if the first premiss is true, it holds true *a priori* in virtue of the meanings of the words involved; but if the second premiss is true, it holds true as an *a posteriori* matter, as a fact discovered by neurophysiology.

The theoretical definition of the causal relation, presented above, opens the door to a theoretical identification in much the same way. The definition says that the causal relation is the intrinsic relation that typically holds between distinct events when one increases the chance of the other. What is this relation? If the term for the causal relation is not to be empty, there must be some way of specifying what relation the term denotes in the actual world. A number of different proposals for identifying the causal relation with physically specifiable relations have been made. The best known is David Fair's (1979) proposal that the causal relation is the relation of transfer of energy-momentum from cause to effect. Fair observes that, with many commonly recognisable causal relations, there is a flow of energy or momentum from the cause to the effect: in fact, if the cause and effect belong to a closed system, the conservation laws enable us to identify the quantity of energy or momentum which is transferred from cause to effect. Fair makes it clear that he is presenting a theoretical identification rather than a conceptual definition of causation. To conform with Lewis's account of theoretical identification, Fair's proposal might be reconstructed as the conclusion of an argument of the following kind: the causal relation is the intrinsic relation that typically holds between distinct events when one increases the chance of the other; the relation of energy-momentum transfer is the intrinsic relation that typically holds between distinct events when one increases the chance of the other; therefore, the causal relation is the relation of energy-momentum transfer. I do not say that Fair advances this argument, or even mentions increase in chance between distinct events. I merely say that if the two premisses were true, they would provide a compelling reason for accepting Fair's proposal. As in the case of the argument supporting the identification of mental states, the first premiss is to be understood as an *a priori* truth and the second premiss as an *a posteriori* truth.

A third similarity between the proposed theory of causation and Lewis's theory of mental states is that both define the theoretical terms in question in terms of non-rigid definite descriptions. The non-rigidity of the descriptions means that it is a contingent matter what the theoretical terms refer to, a matter that can vary from one world to another. For example, we have good reason for thinking that the mental state term "pain", defined as the occupant of the causal role *R*, refers to some kind of neural state in the actual world; but in another possible world it could refer to a pulsation of the ectoplasm, or a movement of an incorporeal Cartesian spirit. Similarly, the term "causal relation", defined as the intrinsic relation that typically holds between distinct events when one increases the chance of the other, may refer to different things in different worlds. If Fair is correct, then the term refers to the relation of energy-momentum



transfer in the actual world; but in another possible world, it might denote a completely different intrinsic relation, say one that does not involve a continuous linking process. The issue of whether "pain" is a non-rigid designator is a controversial matter.[28] But I take it that if the neural state that actually occupies the pain role in the actual world were to occupy a different role in another possible world, it would no longer deserve the name of pain. Similarly, if the intrinsic relation which satisfies the description of the causal relation were to satisfy a completely different description in another world—say, the description "the intrinsic relation that typically holds between distinct events when one *decreases* the chance of the other"—then it would no longer deserve to be called the causal relation. In these cases, it is the role-state rather than the realiser-state which carries more weight.[29]

## 5. *The definition relativised to properties*

The causal relation was defined in §3 in a way that abstracted from the nature of the causally related events: the definition made no mention of the properties or other features of the events. This way of proceeding overlooks the powerful intuition that causal relations relate events in virtue of specific features of the events themselves, in particular the properties they exemplify.

A simple example illustrates the intuition. I throw a stone at a window causing it to break. This conforms nicely to the theory of causation presented in §3, since there is an intrinsic relation connecting the event of my throwing the stone and the window's breaking, a process consisting in a transfer of energy-momentum from cause to effect. If the causal relation between these events existed simply in virtue of this intrinsic relation, it would not matter which properties the events exemplified. But it does matter: only certain features of the set-up are relevant to the causal rela-

[28] See Kripke's claim that "pain" is a rigid designator in his (1972). I think that Kripke's claim, which is based on an appeal to self-evidence, has been satisfactorily countered by Lewis (1994).

[29] The definition of causation offered here could be seen in a different light—a light according to which our term for the causal relation is a rigid designator. Thus, the definition could be seen as merely fixing the reference of the term: "causation" refers to that relation, whatever it is, which *in the actual world* typically holds when one distinct event increases the chance of another. (The justification for this view might be similar to that usually given for the view that the theoretical definition of pain serves merely to fix the reference of the term "pain".) But for the reasons mentioned briefly at the end of this paragraph I do not favour this way of viewing the definition.

tion. For example, what colour eyes I have, how old the stone was, how clean the window was are not causally relevant; but the momentum of the stone, the direction in which it was thrown, the composition of the glass are all relevant. The fact that some properties but not others are relevant to a causal relation shows that the existence of a causal relation between events depends to some extent on the properties exemplified by the events.

Another example illustrates the way in which our intuitions about the causal relevance of properties dominate our causal judgements. In a game of billiards I hit the cue ball with great force, aiming at a particular pocket; but the ball hits the metal rim of the billiards table and is deflected off the table; it then ricochets against several walls, finally bouncing back onto the table where it rolls into the pocket.[30] It is reasonable to think that an intrinsic relation—the transference of energy-momentum—links all the stages of this process. But it is implausible to think that all the links in the process are positive causal ones. In particular, the ball's being deflected by the metal rim seems to be negatively relevant to its falling into the pocket. It is more appropriate to say that the ball fell in the pocket *despite* the fact that it was deflected by the rim, rather than *because* it was deflected by the rim. So in this case there is no positive causal link between the events even though there is an appropriate intrinsic relation linking the events, in abstraction from any properties they exemplify. The example illustrates the way in which our causal intuitions are shaped by judgements about the causal relevance of the properties exemplified by the relevant events.

The definition of causation must be modified in some way to reflect the role of the properties exemplified by the events. The most natural way of doing this is to let the properties exemplified by the cause and effect events restrict the class of situations that the "typically" modifier operates over. Hitherto, this class has been implicitly understood to be the total class of situations in which one event increases the chance of another distinct event: the causal relation was then characterised as the intrinsic relation that typically holds throughout this class of situations. The current suggestion is that the concept of causation can be understood as relativised to more restrictive subclasses of situations, subclasses identified by the properties exemplified by the cause and effect events. One theoretical advantage of this strategy is that it leaves open the possibility of a richer and more fine-grained set of identifications of causation: rather than talk of the causal relation *simpliciter*, we can talk of one kind of causal relation relative to one subclass of chance-

[30] This example is a modification of Deborah Rosen's well-known golfball-and-squirrel example described in her (1978).



increases, and a different kind of causal relation relative to another subclass of chance-increases.[31]

Consider a situation in which a causal relation exists between two distinct events $c$ and $e$, where $c$ exemplifies the property $F$ and $e$ exemplifies the property $G$. Then the suggestion that these exemplified properties $F$ and $G$ should play a part in determining the class of situations that fix the causal relation can be cashed out in the following way:

> if $c$ and $e$ are distinct events and the event $c$ exemplifies the property $F$ and $e$ exemplifies the property $G$, the causal relation between $c$ and $e$ is the intrinsic relation (of the appropriate degree of naturalness) that typically holds between distinct $F$-type events and $G$-type events when the $F$-type event increases the chance of the $G$-event.

Two features of this definition require further comment. The first feature is the reference to an intrinsic relation of an appropriate degree of naturalness. The notion of an intrinsic relation was explained in §3 in terms of natural properties and relations. But the naturalness of these properties and relations comes in degrees.[32] Among the properties and relations instantiated in the actual world, the fundamental properties and relations of physics are perfectly natural, while those invoked by the higher-level sciences are natural to corresponding lesser degrees. An intrinsic relation of an appropriate degree of naturalness relative to a given level of scientific description is one that supervenes on the properties and relations with the degree of naturalness that applies at that level of description. Relative to this level of description, such relations are not too disjunctive in character; and to the extent that they are disjunctive they are not disjunctions of very heterogeneous relations.

The second feature of this definition that requires comment is that the class of situations in which an $F$-type event increases the chance of a distinct $G$-type event is to include actual and possible situations and they are to be varied with respect to their background conditions. The purpose of this requirement is to ensure that the identification of the causal relation

[31] Lewis (1983a, pp. 122–32) similarly restricts the definition of mental states, observing that the non-rigidity of mental state terms like "pain" may begin in the actual world. For example, it may be that pain, conceived of as the state which typically occupies the pain role, is one state in humans, another state in robots, and yet another state in Martians. To accommodate this possibility, Lewis's theory relativises the concept of pain to a population. To say that a human is in pain is to say that he is in the state which typically occupies the pain role for humans; to say that a Martian is in pain is to say that it is in the state which typically occupies the pain role for Martians. The obvious theoretical advantage of relativising the concept of pain in this way is that it provides for a richer range of possible identifications of mental states.

[32] Lewis (1983b, 1986b) emphasizes the fact that properties and relations have varying degrees of naturalness.

is counterfactually robust—to ensure that the identification would hold under varied counterfactual assumptions about background conditions. The rationale of the requirement is obvious enough. If we think that a causal relation exists between two events because of the properties they exemplify, then it should be just the properties exemplified by the events, and none of the other features of the situation, that are relevant to identifying the causal relation: accordingly, variation in background conditions should not affect the identification of the relation. This requirement has the effect of ruling out cases in which the causal relation between two events is identified with an intrinsic relation that just happens to hold on the few occasions on which an *F*-type event increases the chance of a *G*-type event. This relation would not count, on the present account, as the causal relation between *c* and *e* if it did not also typically obtain throughout the whole class of situations—both actual and possible situations conforming to the same laws but differing in their background conditions—in which there is a probabilistic dependence between an *F*-type and a *G*-type event.[33]

This modified account of causation has some interesting consequences that do not bear directly on the issue of pre-emption, but are worth noting because they show how the theory is independently justified. One consequence concerns the highly discriminating character of our judgements about causation. Lewis's probabilistic theory has trouble explaining away the fact that it implies that certain events are causes when they are not intuitively judged to be causes. An example of Jonathan Bennett's (1988, pp. 69–71) illustrates the difficulty: a life-guard saves the life of a man who would have otherwise drowned; the man goes on to live a healthy life and dies of old age many years later. If the lifeguard had not intervened, the man would have drowned and so would not have died of old age many years later. So, on Lewis's theory, the lifeguard's intervention was a cause of the man's death. Lewis defends this unsatisfactory result by saying that his theory is intended to provide us with an account, not of *the* cause, but simply of *one of the* causes (unselectively speaking). Yet, it is counterintuitive to say that the lifeguard's intervention was so much as *a* cause of the man's subsequent death of old age. The present theory of causation, however, provides a better explanation of our intuitions about this case. If the relevant property of the lifeguard's intervention is some property such as saving a person *x* from drowning, and the relevant property of the man's death is some property such as being *x*'s death from old age, then the theory predicts that these events are not causally related. If we hold these properties of the events fixed but change any details of the setting, we can see that the probabilistic dependence vanishes. The probabilistic dependence that holds between these two particular events is not counterfactually robust as it depends on the specific facts obtaining in this particular situation. Because the probabilistic dependence holding between the events is not counterfactually robust, the present theory implies, correctly, that the lifeguard's saving the man's life is not a cause of the man's subsequent death.[34]

Another consequence of the present theory concerns the asymmetry Bennett (1993) has noted between hasteners and delayers of events. He has observed that we count hasteners of events as causes, but not delayers. Here's an example of this asymmetry. A doctor gives a patient a large dose of morphine that hastens his death: the patient would have died of cancer a few days later but the lethal dose of morphine kills him sooner. The doctor's intervention, intuitively speaking, caused the patient's death. Now consider another doctor who revives a patient by stimulating his heart after it has failed; a few days later, however, the patient dies of heart failure. In this case, the doctor's intervention did not, intuitively speaking, cause the patient's death. It is difficult to explain this asymmetry on Lewis's theory because there is a probabilistic dependence in both cases: on both occasions, if the doctor had not intervened, the chance of the patient's dying at the exact time that he did would have been less than it actually was. On the other hand, the present theory of causation provides a ready explanation of the asymmetry in our intuitions. The important point is that the probabilistic dependence in the first case is robust under varied counterfactual assumptions about background conditions, granted that the causally significant property of the cause is the doctor's giving a large dose of morphine and the relevant property of the effect is the patient's dying. By contrast, the probabilistic dependence in the second case is not robust in this way if we suppose that the relevant property of the first event is the doctor's stimulating the patient's heart and the relevant property of the second event is the patient's dying of heart failure. There will rarely be

[33] This requirement of counterfactual robustness is not imposed just to obviate a certain kind of objection to the relativised definition of causation. It also answers to an intuitive feature of our ordinary practice of providing causal explanations of events. It seems that in our explanations of particular events, we show a preference for citing causes which are linked to the events by processes which would obtain under varied counterfactual assumptions about the background conditions. Jackson and Pettit (1992) discuss this preference for counterfactually robust causal explanations, which they say convey comparative causal information.

[34] Murali Ramachandran (1995) tries to provide an account of the selectivity of causal discourse in terms of "sufficiency" counterfactuals. What is common to the present proposal and Ramachandran's proposal is that they both require the robustness of certain counterfactual constructions. There are, however, many differences, the most important being that Ramachandran's proposal is framed only for deterministic causation.

probabilistic dependences between events exemplifying these properties, which indicates that the probabilistic dependence holding in the particular situation depends on the specific, accidental facts of that situation. The present theory provides an explanation of Bennett's asymmetry by way of the difference in counterfactual robustness of the probabilistic dependences.

## 6. Solution to the pre-emption problem

Suppose we are considering whether a causal relation exists between events *c* and *e*, where *c* is an *F*-type event and *e* is a *G*-type event. The present theory says that the first thing we must find out is whether there is an intrinsic relation (of an appropriate degree of naturalness) that typically holds when an *F*-type event increases the chance of a *G*-type event. If there is and this relation holds between *c* and *e*, then they are causally related; if there is no such relation, or there is but it does not hold between *c* and *e*, then they are not causally related. This procedure of determination requires that we be able to identify the intrinsic relation that constitutes the causal relation between *c* and *e*. It may be asking too much to require a detailed knowledge of the relation; often enough, an incomplete knowledge suffices to determine whether the relation exists between the specific events *c* and *e*. In any case, the evaluation of a particular causal claim, on this account, calls upon two kinds of tacit knowledge: first, it calls upon the tacit *a priori* knowledge of how the concept of the relevant kind of causal relation is defined; secondly, it calls upon the tacit *a posteriori* knowledge of how that kind of causal relation is identified.

Consider the way in which the theory applies to the neuron examples described in earlier sections. In discussing these examples, let us assume that the causally significant features of all the events is simply their being neuron firings: the precise manner and time of the firings will not be judged to be significant. To determine which neuron firing counts as a cause of neuron *e*'s firing, we need to determine whether an intrinsic relation typically holds between two neuron firings when one increases the chance of the other. In the case of a causal relation between two proximate neuron firings—where there are no intermediate neuron firings—neurophysiology enables us to identify the causal relation with a process of the following kind: a pulse runs down the first neuron's axon, terminating in the transmission of a neuro-transmitter across a synaptic connection with the second neuron, which transmission lowers the electrical potential of the second neuron below a critical



point where it suddenly collapses entirely, initiating its own pulse.[35] A causal relation between non-proximate neuron firings—where there are intermediate neurons—can be identified with a spatiotemporally continuous chain of several of these processes linked together. Obviously, few lay people are in a position to make these identifications, requiring as they do detailed *a posteriori* knowledge of neuro-physiology. However, most lay people are in a position to recognise that a spatiotemporally continuous chain of processes of some such kind is required, at least in the actual world, for the existence of a causal relation between neuron firings.

Even very fragmentary knowledge of how a causal relation between neuron firings is to be identified enables us to arrive at appropriate judgements about the neuron examples of earlier sections. In the first example, we think that there is a causal relation between *b*'s firing and *e*'s firing because we judge that they are linked by a continuous process of the kind that typically holds when one neuron firing increases the chance of another's firing. And we think that there is no causal relation between *a*'s firing and *e*'s firing because they are not linked by such a process. A similar explanation can be given of the different causal status of *a*'s firing and *b*'s firing in the second neuron example illustrating late pre-emption. The account handles this example in much the same way as the first, treating early and late pre-emption cases, and probabilistic and deterministic pre-emption cases in a uniform manner.

## 7. Some objections considered

I conclude by considering some objections. They revolve around the heavy reliance the conceptual definitions of causation, in both the unrelativised and relativised forms, place on the qualifier "typically". Can pres-

[35] This description of the process by which one neuron firing causes another is taken from Churchland (1984, pp. 129–31). The description raises the question: is the description vitiated by the fact that it employs causal terminology such as "transmit", "lower", and "initiate"? To be sure, the terms used in the description of the underlying process are properly understood as causal, but this does not vitiate the identification of the process with the causal relation in question. I want to leave open the possibility, at least on the relativised model of causation, that the process identified with a causal relation may itself involve causal relations. Of course, these lower-level causal relations would have to be identified, perhaps in terms of still lower-level causal relations. Ultimately, if we are to avoid an infinite regress, the process of identifying causal relations has to reach a point where no further causal relations are appealed to. This will be the point at which the unrelativised model of causation would yield a unique non-causal identification of the causal relation.

sure be brought to bear on the definitions at this point? Consider, for example, the following possible case that *appears* to pose a problem for the account. A possible world does not contain many systems of neurons, but all the systems it does contain have the simple structure depicted in Figure 1. The systems of neurons of this world are governed by the same laws as were imagined to hold earlier. Suppose that what happened in the earlier example turns out, by coincidence, to occur all the time in all the systems of neurons in this world: neurons *a* and *b* fire at the same time, the unreliable process started by *b*'s firing cuts short the reliable process started by *a*'s firing, and this unreliable process just happens to go to completion and bring about *e*'s firing. It might be thought that identifying the causal relation between two neurons firing with the relation that *typically* exists between two such firings when one increases the chance of the other is bound to lead to the wrong results here. For it seems, in this scenario, there is *never* any process connecting the *a*-firings with the *e*-firings even though the former kind of event increases the chance of the latter.

There is a very natural way of countering this objection, which has already been mentioned in passing. In specifying the relativised causal relation between neuron firings as the relation that typically holds when one neuron firing increases the chance of another, I intended to include both actual and possible situations involving neuron firings. The only restriction on the possible situations is that they should be governed by the same laws as the actual situations. But obviously if we take account of all the possible instances of firings in the systems of neurons, we can see that in the majority of these instances *a*'s firing and *b*'s firing occur independently of each other, as do the processes they initiate; and even in those possible instances where *a* and *b* fire together, the majority will be ones in which the process started by *a*'s firing goes through to completion and brings about *e*'s firing while the unreliable process started by *b*'s firing does not. We know this because it is stipulated as a basic feature of the example that *a*'s firing increases the chance of *e*'s firing and *b*'s firing lowers it; and these chances must be reflected in the possible, even if not the actual, instances of the neuron firings. Accordingly, one can reason that in the majority of these instances, *a*'s firing will be linked to *e*'s firing by an appropriate process that can be identified as the relation that holds when one neuron firing causes another.

Is it possible to construct a further counterexample to the account that resists this type of response? Consider, for example, the following possible case that again *appears* to present a difficulty to the theory. Another possible world contains only systems of neurons like that depicted in Figure 2, illustrating late deterministic pre-emption. There is, however, a slight difference between the systems of neurons of this world and the



kind of system depicted in Figure 2. With the new kind of system of neurons, it is not an accident that neurons *a* and *b* fire together; rather it is a consequence of the laws of nature that neuron *b* cannot fire without neuron *a* also firing. I have depicted this in Figure 3 by representing *a*'s firing and *b*'s firing as effects of a common deterministic cause, *x*'s firing.

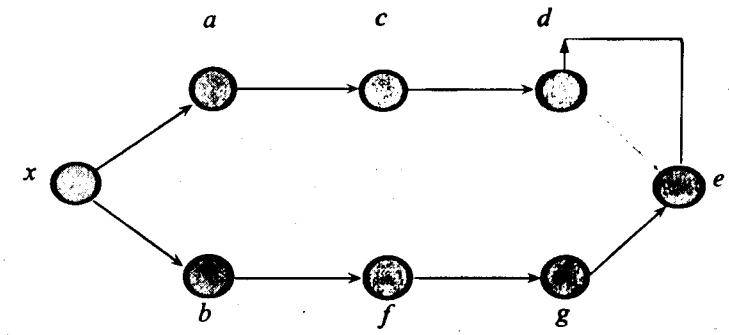


Figure 3

This kind of example might be claimed to present a special difficulty for the present theory because *a*'s firing neither raises nor lowers the chance of *e*'s firing, and the same is true of *b*'s firing. This follows from the fact that all the connections between the firings are deterministic, so that the chance of *e*'s firing is 1 whatever happens, provided that *x* fires. But if both *a*'s firing and *b*'s firing leave the chance of *e*'s firing unaffected, how are we to isolate the relation to be the identified as the causal relation between neuron firings? This question cannot be answered by appealing to possible instances of firings in the neuron system: since all such neuron firings are, by stipulation, governed by laws that ensure that *b* can never fire without *a* firing as well, we know that if both *b*'s firing starts up the main process, *a*'s firing will start up the alternative process, in which case neither *a*'s firing nor *b*'s firing will increase the chance of *e*'s firing.

This objection is not as insuperable as it looks at first sight. Even though it is true that neither *a*'s firing nor *b*'s firing in this kind of neuron system increases the chance of *e*'s firing, there are other pairs of neuron firings in the system that are so related that one increases the chance of the other. Consider, for example, the fact that *a*'s firing increases the chance of *c*'s firing, *b*'s firing increases the chance of *f*'s firing, and *f*'s firing increases the chance of *g*'s firing, and so on. These relations exist, I am assuming, because the events occurring as part of one process do not depend on the events occurring in the other process; for example, *f*'s firing depends on *b*'s firing but not on *a*'s firing, *c*'s firing, or any other event in

the alternative process. Accordingly, we have plenty of instances of chance increases between neuron firings and these can be used to identify the causal relation between neuron firings. However this relation is identified precisely, it is clear that it holds, not only between *b*'s firing and *f*'s firings, and between *f*'s firing and *g*'s firing, but also between *b*'s firing and *e*'s firing. There is, then, no problem about identifying the correct causal relations in this example.

To sum up: the present theory of causation has many notable virtues. First of all, it yields the right results for the examples that posed difficulties for the more orthodox probabilistic theories, as well as the more complicated examples just considered.

Secondly, the theory does not rule out the possibility of action at a temporal distance. Even though it is unlikely that the actual world could contain action at a temporal distance, the account is consistent with its conceptual possibility. The non-rigid character of the conceptual definition of causation allows that in a remote possible world one event might be a cause of a temporally distant event without there being any intervening process connecting them.

Thirdly, the theory neatly captures the idea that the causal relation is a local feature of a cause-effect pair by taking a causal relation between two events to consist in an intrinsic relation that would exist regardless of whatever else occurred in the surrounding neighbourhood. The theory implies, for instance, that the causal relation between *b*'s firing and *e*'s firing in the first neuron example obtains in virtue of the intrinsic nature of the events and the process connecting them, independently of the presence of alternative causes of *e*'s firing.

Fourthly, the theory captures the powerful intuition that a causal relation exists between events in virtue of features of the events, especially the properties exemplified by the events. This powerful intuition lies, I think, at the heart of the appeal of regularity accounts of causation that state that causal relations must fall under laws of nature, picked on the basis of properties exemplified by the related events. These theories do not, however, cope well with cases of pre-emption because they make causal relations depend too directly on global, extrinsic considerations. In contrast, the present theory handles these cases by expressly construing causal relations as intrinsic relations. Nonetheless, it is able to capture the intuition in question because it characterises these intrinsic relations in global, extrinsic terms as those intrinsic relations that hold throughout a class of situations identified by the properties exemplified by the cause and effect.

The theory of causation presented here conforms to the spirit of probabilistic theories in that it makes essential use of the idea that a cause raises the probability of its effect. But the theory avoids the mistake of other



probabilistic theories of trying to eliminate causation in favour of probabilistic relations. On the contrary, the account dignifies causal relations with the same status that is accorded to theoretical entities in general. A proper appreciation of pre-emption seems to demand no less. [36]

*Philosophy Discipline* PETER MENZIES
*School of History, Philosophy, and Politics*
*Macquarie University*
*North Ryde,*
*Sydney, NSW 2109*
*AUSTRALIA*
*pmenzies@laurel.ocs.mq.edu.au*

## REFERENCES

Bennett, J. 1988: *Events and Their Names*. Oxford: Oxford University Press.

——1993: "Event Causation: The Counterfactual Analysis", in E. Sosa and M. Tooley (eds.), *Causation*, pp. 217–33. Oxford: Oxford University Press.

Bigelow J. and Pargetter, R. 1990: *Science and Necessity*. Cambridge: Cambridge University Press.

Carnap, R. 1966: *Philosophical Foundations of Physics*. New York: Basic Books.

Churchland, P. 1984: *Matter and Consciousness*. Cambridge, Mass.: MIT Press.

Davidson, D. 1980: *Essays on Actions and Events*. Oxford: Oxford University Press.

Dowe, P. 1993: "Singular Causation", unpublished paper presented at 1993 meeting of Australasian Association of Philosophy.

Eells, E. 1991: *Probabilistic Causality*. Cambridge: Cambridge University Press.

Fair, D. 1979: "Causation and the Flow of Energy". *Erkenntnis*, 14, pp. 219–50.

Jackson, F. 1992: "Critical Notice of S. Hurley's *Natural Reasons*". *Australasian Journal of Philosophy*, 70, pp. 475–88.

Jackson, F and Pettit, P. 1992: "In Defense of Explanatory Ecumenism". *Economics and Philosophy*, 8, pp. 1–21.

[36] I am grateful to David Braddon-Mitchell, Phil Dowe, Brian Garrett, Frank Jackson, Martin Leckey, Graham Oppy and Philip Pettit for discussions of the ideas in this paper. I am also grateful to David Lewis, Hugh Mellor, Mark Sainsbury and a referee for this journal for helpful written comments on earlier versions of this paper.

Kim, J. 1975: "Events as Property Exemplifications", in M. Brand and D.Walton (eds.), *Action Theory*, pp. 159–77. Dordrecht: Reidel Publishing Company.

Kripke, S. 1972: "Naming and Necessity", in D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, pp. 253–355. Dordrecht: Reidel.

Mellor, D.H. 1991: "On Raising the Chances of Effects", in his *Matters of Metaphysics*, pp. 225–34. Cambridge: Cambridge University Press.

Menzies, P. 1988: "Against Causal Reductionism". *Mind*, 97, pp. 551–74.

——1989a: "Probabilistic Causation and Causal Processes: A Critique of Lewis". *Philosophy of Science*, 56, pp. 642–63.

——1989b: "A Unified Account of Causal Relata". *Australasian Journal of Philosophy*, 67, pp. 59–83.

Lewis, D. 1972: "Psychophysical and Theoretical Identifications". *Australasian Journal of Philosophy*, 50, pp. 249–58.

——1983a: *Philosophical Papers, Volume 1*. New York: Oxford University Press.

——1983b: "New Work for a Theory of Universals". *Australasian Journal of Philosophy*, 61, pp. 343–77.

——1986a: *Philosophical Papers, Volume 2*. New York: Oxford University Press.

——1986b: *The Plurality of Worlds*. Oxford: Basil Blackwell.

——1994: "Reduction of Mind", in S. Guttenplan (ed.), *A Companion to Philosophy of Mind*. Oxford: Blackwell.

Ramsey, 1931: "Theories", in R. Braithwaite (ed.), *The Foundations of Mathematics*. London: Routledge and Kegan Paul.

Ramachandran, M. 1995: "Introducing the 'Sufficiency' Analysis of Causation", unpublished paper.

Rosen, D. 1978: "In Defense of a Probabilistic Theory of Causality". *Philosophy of Science*, 45, pp. 368–86.

Ruben, D-H. 1994: "A Counterfactual Theory of Causal Explanation". *Noûs*, 28, pp. 465–81.

Salmon, W. 1984: *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Skyrms, B. 1980: *Causal Necessity*. New Haven: Yale University Press.

Sober, E. 1984: *The Nature of Selection*. Cambridge: MIT Press/Bradford Books.

——1985: "Two Concepts of Cause", in P. Asquith and P. Kitcher (eds.), *PSA* 1984, pp. 405–24. East Lansing, Michigan: Philosophy of Science Association.



Suppes, P. 1970: *A Probabilistic Theory of Causality*. Amsterdam: North Holland Press.

——1984: *Probabilistic Metaphysics*. Oxford: Basil Blackwell.

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