Philosophy and Phenomenological Research
Vol. L, Supplement, Fall 1990

Causation in the Philosophy of Mind

FRANK JACKSON AND PHILIP PETTIT
Australian National University

Causation has come to play an increasingly important role in the philosophy of mind, reaching its apotheosis in the doctrine that to be a mental state of kind K is to fill the causal role definitive of that kind of mental state: the typology of mental states is a typology of causal roles. However, ironically, there is, from this very functionalist perspective, a problem about how to understand the causal role of mental properties, those properties which make a mental state the kind of mental state that it is. This problem surfaces in one way or another in the debates over the language of thought (for instance, in the argument that only if intentional states have syntactic-like structure can they play the required causal roles); over the explanatory role of broad content (for instance, in the argument that broad content is explanatorily irrelevant to behaviour because doppelgängers behave alike while possibly differing in broad content); and over the eliminativist implications of connectionism (for instance, in the argument that certain versions of connectionism falsify the propositional modularity component of the folk conception of the causes of behaviour). We wish, however, to reverse the usual order of discussion. Instead of entering directly into one or another of these fascinating debates, we want to raise the problem of how to understand the causal role of mental properties as an issue in its own right. We will then offer a solution to the problem which seems to us plausible independently of those debates. The final stage of our discussion will be a brief application of the proffered solution to argue that connectionism does not have the eliminativist implications sometimes associated with it.[1]

[1] For the application to how broad content can explain see, e.g., Martin Davies, 'Individualism and Supervenience', *Proceedings of the Aristotelian Society*, supp. vol. 60 (1986): 263-83, Frank Jackson and Philip Pettit, 'Functionalism and Broad Content', *Mind*, XCVII, 387 (1988): 381-400, and the references therein. The bearing of the issue about the causal role of mental properties, and of content properties especially, to the

1. The problem

How things were at some earlier time is succeeded by how things are at subsequent times, and we distinguish the way and extent to which how things were is causally responsible or relevant to how they are or will be. For instance, one aspect of how things were a little while ago is that there was a sharp drop in atmospheric pressure, and another aspect of how things were a little while ago is that a man with an odd number of freckles scratched his nose; the first aspect of how things were is causally relevant to the fact that it is now raining, the second is not.

But how things were, are or will be at a time is a matter of which properties are instantiated at that time. So our commonplace observation amounts to noting that we can and must distinguish a relation of (positive) causal relevance among *properties*. Those who hold to a fine-grained or relatively fine-grained conception of events, which broadly places them in the category of property instances, or of property instances of some favoured class of properties, will see this as really nothing more than the familiar doctrine that (singular) causation relates events.[2] But for those who hold to a coarse-grained conception of events which places them in the category of concrete particulars which have or instantiate properties, but are quite distinct from, and much more sparse than, properties or their instances, our commonplace must be seen as an addition to the story

debate over the language of thought is central in Jerry Fodor, *Pyschosemantics* (Cambridge: MIT Press, 1987). See also D. R. Braddon Mitchell and J. B. Fitzpatrick, 'Explanation and the Language of Thought', *Synthese*, forthcoming.

[2] For different versions of this approach see, e.g., Jaegwon Kim, 'Events as Property Exemplifications' in *Action Theory*, ed. M. Brand and D. Walton (Dordrecht: Reidel, 1976), pp. 159-77, David Lewis, 'Events', in *Philosophical Papers*, vol. II (Oxford: Oxford University Press, 1986), pp. 241-69, and David Sanford, 'Causal Relata', in *Actions and Events*, Ernest LePore and Brian McLaughlin, eds. (Oxford: Basil Blackwell, 1985), pp. 282-93. But note that property instances in these approaches need to be distinguished from property instances in the sense of the tropes of Donald Williams, 'The Elements of Being' in *Principles of Empirical Realism* (Springfield: Thomas, 1966). For in these approaches, when one and the same person at one and the same time says hullo loudly and thereby says hullo, the instance of saying hello is distinguished from that of a saying hello loudly – that is essential to allowing them to stand in different causal relations. Whereas Williams's tropes are absolutely specific; and so, if the saying hullo is a property instance in his trope sense, it is identical with saying hullo in some absolutely specific manner, and, therefore, in the case in question, to saying hullo loudly. On the absolutely specific nature of property instances on the Williams's scheme, see Keith Campbell, *Metaphysics* (Belmont, California: Dickenson, 1976), chap. 14. As a result of this point, a trope approach to the relata of the causal relation will, like the Davidsonian approach discussed next in the text, need to regard our commonplace as something to be added to the story about causation.

about causation being a relation between events. In addition to asking which events are causally relevant to which other events, we can and must ask which properties of events are causally relevant to which other properties. But surely this must, or should, have been an implicit ingredient in the story about singular causation all along.[3] Surely not even the most robust defender of a concrete conception of events supposed that *featureless* events might do some causing. Their events caused what they did because of how they were — that is to say, because of which properties they possessed.

Accordingly, we are going to take as a datum the idea that we can distinguish among properties in respect of their causal relevance to the obtaining of some effect or other. Exactly how to fit this fact into an event metaphysics of causation is left as a question for another time. We should emphasise that by 'causal relevance' in what follows we mean positive, actual causal relevance. We mean what might best have been called 'causal responsibility' except that 'relevance' has become somewhat entrenched in the literature, and 'responsibility' perhaps carries a connotation of sufficiency, whereas we are talking about the idea of a property being *a* factor, and typically one factor among many, in the causing of something.

We are now in a position to state our problem. Perhaps we can say a priori that a number's being prime cannot be causally relevant to any physical occurrence, but most often the question of whether a property is causally relevant to some effect is an a posteriori one. It was, for example, an empirical discovery that the mass of a body is irrelevant to its rate of acceleration under gravity in a vacuum, and that the density of a medium is relevant to the speed of light through that medium.

---

[3] And is, we think, though under a different guise, in Donald Davidson's adumbration of the view that causation is a relation between events concretely conceived in his 'Causal Relations' *Journal of Philosophy*, 64 (1967): 691-703. For, first, he holds that singular causal relations hold in virtue of causal laws (while holding that exactly how to spell this out is no easy matter), and, secondly, in discussing the kinds of examples which lead other writers to make events property-like, he admits in effect that we can, when dealing with what is in his view one and the same event, discriminate which properties of the single event play a special role in a causal explanation: although the bolt's giving way is one and the same event as its giving way suddenly in his view, the special place we may well give the latter in explaining the tragedy is accommodated by giving the correlative property a special place in the causal explanation of the tragedy. Davidson may well wish to urge that this special place is a place in a causal *law*, not in a singular causal relation. But for our purposes what is central is the partition of properties into causally relevant and causally irrelevant ones with respect to some effect, not whether this partition is a topic in the theory of singular causation or the theory of causal laws. We are indebted here to a discussion with Peter Menzies.

How do we establish that some property or set of properties is causally irrelevant to some effect? An attractive answer is that we do so by completely explaining the effect in terms of properties distinct from that property or set of properties. This is the point behind the familiar argument — sometimes referred to as 'the shadow of physiology' argument — against dualist interactionist theories of mind.[4] It is observed that it is very plausible that in principle a complete explanation of each and every bodily movement of a person can be given in terms of their internal physiology, with their neurophysiology playing a particularly important role, along with interactions of a physical kind between their physiological states and their environment. There are no mysterious, unclosable-in-principle gaps in the story medical science tells about what makes a person's arm go up. The conclusion then is that the sort of properties that feature in the dualist story are causally irrelevant to behaviour; and we are led to the familiar objection to dualism that the interactionist variety of dualism has to give way to an epiphenomenalist variety — and so much the worse for dualism![5]

Our problem is that if the popular functionalist approach to mental properties is correct, the very same style of argument appears to be available to cast doubt on the causal relevance of mental properties. The shadow of physiology seems to raise a problem for functionalists as well as for dualists despite the fact that functionalism is compatible with a purely materialistic view of the mind. Take, for instance, content and our commonsense conviction that content is causally relevant to behaviour, our conviction that the fact that a certain state of mine has the property of being the belief *that p* or of being the desire *that q* is causally relevant to my arm moving in a certain way. (We will stick with this example from now on in order to avoid the difficult problem of *qualia* or raw feels. We take it for granted that a materialistic account of an essentially naturalistic variety can be given of intentional states and their contents.) How can that be, given the just discussed fact that a complete explanation in principle entirely in physiological terms of my behaviour is possible? For the kind of property content is identified with in the functionalist story will not appear anywhere in that story. What will matter at the various points in that story will be the physiological, and particularly neurophysiological, properties involved, whereas, as has so often been emphasised, what matters from the functionalist perspective for being a certain kind of mental

---

[4] See, e.g., Keith Campbell, *Body and Mind* (London: Macmillan, 1971), chap. 3.

[5] But see Campbell, *Body and Mind*, and Frank Jackson, 'Epiphenomenal Qualia', *Philosophical Quarterly*, 32, 127: 127-36, for reservations about the decisiveness of this argument when (and only when) directed at qualia.

state is not the nature of the state neurophysiologically speaking but rather the functional role occupied by that state. One way of putting the point is by saying that what drives behaviour is the physiological nature of the various states, not the functional roles they fill. How then can functional role, and so content according to functionalism, be a causally relevant property?

Some have concluded from considerations like these — so much the worse for functionalism as an account of content, in somewhat the same way that an earlier generation of philosophers concluded — so much the worse for dualism.[6] We think, however, that there is an important error in the line of thought that suggests that functionalism makes content (and mental properties in general) causally irrelevant or epiphenomenal. Before we say what it is, we need to say why, as it seems to us, two initially attractive responses to our problem fail. The first response appeals to a type-type version of mind-brain identity theory based on functionalism; the second to the fact that functional role is supervenient on physiology plus physical environment.

The identity theory response to our problem

The identity theory as originally presented was a type-type identity theory. Mental properties, including the possession of some particular content, were identified with neurophysiological properties.[7] Functionalists sometimes speak as if the familiar and correct point that the kind of functional role definitive of content (to stick with that example) can be, and most likely is, variously realized in different sentient organisms refuted this theory.[8] We agree, however, with the unrepentant type-type

---

[6] Most recently, Ned Block, 'Can the Mind Change the World?' in George Boolos, ed., *Meaning and Method: Essays in Honour of Hilary Putnam* (Cambridge: Cambridge University Press), forthcoming. See also Hartry Field, 'Mental Representation' reprinted in Ned Block, *Readings in Philosophical Psychology*, vol. 2 (Cambridge: Harvard University Press, 1981), pp. 78-114, see esp. pp. 88-96; Jerry Fodor, 'Introduction: Something on the State of the Art' in his *Representations* (Sussex: Harvester, 1981), but see his *Psychosemantics* (Cambridge: MIT Press, 1987), p. 140 for what we take to be something akin to the supervenience approach we describe below. The problem and the associated issues would, of course, be much the same for views which regard functional role as a major ingredient, along with evolutionary history or whatever, in determining content. To keep things simple, we will set these hybrid views to one side.

[7] See, e.g., J. J. C. Smart, 'Sensations and Brain Processes', *Philosophical Review*, LXVIII, 1959: 141-56, D. M. Armstrong, *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul, 1968), and, most explicitly, David Lewis, 'An Argument for the Identity Theory', *Journal of Philosophy*, 63, 1 (1966): 17-25.

[8] See, e.g., Hilary Putnam, 'The Mental Life of Some Machines', reprinted in his *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1975). Putnam is, of

theorists that the point about the possibility of different neurophysiological states realizing a given content in different species, or even in different members of the same species, or even in a given individual at different times, only shows that different properties may be a given content in different species, or in different members of the same species, or in the one individual at different times.[9] Does this mean that we should espouse a simple answer to our question about the causal relevance of being in a state with a certain content — namely, the answer that being in that state is precisely as causally relevant to the action it putatively explains as is the neurophysiological property the relevant content property is identical with?

We think that this reply evades the crucial question of concern. When I explain your behaviour by citing your belief that it is about to rain, I am surely explaining your behaviour in terms of something I know about you, or at least that I think that I know about you. I am not saying that there is some internally realized property, I know not what, which is causally relevant to your behaviour. That would be hardly more than a declaration that your action is not a random occurrence. I am rather explaining your behaviour in terms of something I know about you; and as I do not know, and know that I do not know, about the nature of your internal physiological states, it can only be the relevant functional role which I am citing as the property which you instantiate which is causally relevant to your behaviour. When we explain behaviour in terms of the contents of beliefs and desires, the properties we are invoking must be the known or guessed about functional roles, not the unknown nature of the occupiers of those roles. Moreover, even though type-type theorists identify a given belief content in a given organism on a given occasion with a neurophysiological state (type) rather than a functional state, they must and do hold that it is the functional role the state occupies, not the kind of neurophysiological state it is, which gives that state the belief content it has. Functional role is the final arbiter. The upshot is that we need to vindicate the causal relevance of functional role — it is what we know about and what in the final analysis matters — in order to justify the commonsense attitude to causal explanations in terms of content.[10]

---

course, no longer a functionalist, see, e.g., chap. 5 of *Representation and Reality* (Cambridge: MIT Press, 1989), but the point is widely accepted, see, e.g., Dan Dennett, 'Current Issues in the Philosophy of Mind', *American Philosophical Quarterly*, 15, 4 (1978): 249-61.

[9] See, e.g., David Lewis, 'Review of Putnam' reprinted in *Readings in Philosophical Psychology*, vol. 1 (London: Methuen, 1980), ed. Ned Block, 1980, pp. 232-33, and Frank Jackson, Robert Pargetter and Elizabeth Prior, 'Functionalism and Type-Type Identity Theories', *Philosophical Studies*, 42 (1982): 209-23.

[10] We take it that our objections here are essentially the same as Block's, 'Can Content

The supervenience response to our problem

Our problem was framed in the following terms. The whole causal story about the origins of behaviour can be told in terms of the neurophysiological nature of our internal workings combined with environmental considerations — where then is there room for functional properties to do any causal work? Ergo, functional properties are causally irrelevant. Our model was the familiar argument which forces dualists into embracing an epiphenomenalist position on the mind.

There is, however, a major difference between dualism and functionalism. Although both see properties other than neurophysiological ones as what is crucial to being minded, the properties functionalism sees as crucial *supervene* on physiology, or at least on physiology together with the relevant laws and, if we are dealing with broad functional roles, certain environmental and historical factors, whereas the properties dualists see as crucial are *emergent* ones. There is a sense, that is, in which the crucial properties according to functionalists, namely, the functional properties, are not *wholly* distinct from the neurophysiological ones. Although no functional property is identical with any neurophysiological one, enough by way of neurophysiological properties when combined with environmental facts (and perhaps laws of nature) fully determines the functional properties. The supervenience reply to our problem, thus, is the observation that from the fact that the whole causal story can be told in neurophysiological terms, and that no functional property is any neurophysiological property, it does not follow that the functional properties do not appear in the story. They appear in the story by supervening on the neurophysiological properties (in the same way though less transparently, that you and I being the same height appears in a story that includes your being 182 cms and my being 182 cms in height).

This reply wins the battle but not the war. Our problem is not how to reply to scepticism about whether functional properties are instantiated, but how to reply to scepticism about their causal relevance. Some philosophers have worried about whether we should acknowledge truth and reference as features of the world on the ground that neither the fact that some sentence is true nor the fact that some word has a certain reference plays a role in explaining the causal order of the natural world.[11] The reply to this worry is that truth and reference supervene on what does feature in the best explanations of the natural world. Similarly, the fact that a person's behaviour can be explained in full without explicit reference to

---

Change the World?'

[11] See, e.g., Michael Devitt, *Realism and Truth* (Oxford: Basil Blackwell, 1984), chap. 6.

the functional properties as such of their internal states does not show that we should be sceptics about their states instantiating functional properties. The functional properties supervene on the properties we do explicitly invoke in our explanations. It is, though, one thing to be reassured about the presence of certain properties, and another to be reassured about their causal relevance. The point about supervenience leaves open the question of the causal relevance of the functional properties.

It might be thought easy to close this question by appeal to the following principle: If being F is causally relevant to some effect E, and being G supervenes on being F, then being G is causally relevant to E. The idea would be that we solve our problem by observing (a) that physiological properties are non-controversially causally relevant to behaviour, (b) that functional properties supervene on them, perhaps in combination with other matters, and then (c) use the principle to obtain the desired result that functional properties, and so, contents, are causally relevant to behaviour.[12] The principle is, however, false. In general for any property or property complex which is causally relevant to the obtaining of some effect E, there will be indefinitely many properties which supervene on that property or property complex, and it would be absurdly generous to count all and sundry as causally relevant. Examples bear this general observation out.

Consider a machine with two weighing platforms set up to respond whenever the weight on one platform is half that on the other. In that case alone a circuit in the machine closes causing a bell to ring, and further suppose that on some particular occasion a weight of three grams is placed on one platform and a weight of six on the other causing the bell to ring. Clearly none of, one weight's being a prime number of grams, one weight's being one less than seven, or one weight's being divisible by three or the weight of the Prime Minister is causally relevant to the bell's ringing. And yet all these properties supervene on the properties that on the occasion were causally relevant — namely, one weight's being three grams when the other was six grams. As we might naturally say it, one weight's being three grams and the other's being six grams was relevant because three is half six and not, for instance, because three is a prime number, or because six is one less that seven. Or again suppose that the

---

[12] The principle would be a kind of converse of that espoused by Jaegwon Kim, see, e.g., 'Epiphenomenal and Supervenient Causation', *Midwest Studies in Philosophy* 9 (1984): 257-70, and, for a recent critical discussion of the surrounding issues to which we are indebted, Peter Menzies, 'Against Causal Reductionism' *Mind*, XCVII (1988): 551-74. The 'perhaps in combination with other matters' is included in (b) to cover the possibility that the functional properties are broad ones tailored to capture broad content. Also, a full specification of the supervenience base should include the relevant laws.

fact that someone lives in a particular suburb on the North side of town is causally relevant to their being happy about where they live. Their living *somewhere* on the North side of town supervenes on their living in the particular Northern suburb that they do live in, yet it need not be the case that their living somewhere on the North side of town causally explains their contentment. Perhaps they particularly dislike all the other Northern suburbs apart from the one they live in — in this case it would not be their living somewhere on the North side, but only their living just where they do in fact live which would be causally relevant. Or again, going to twenty committee meetings may be causally relevant to Jones's sorry state of mind in a way in which going to at least two is not, yet going to at least two supervenes on going to twenty.[13]

A solution to our problem

We can think of functional properties as a more complex and general case of dispositional properties, and as our problem has a simplified analogue in the case of dispositional properties, we will start with them.

Dispositional properties are causally relevant: a glass breaks because it is fragile; Fred is saved because his seat belt has the right degree of elasticity; Mary dies because the ladder she allows to touch power lines is a good conductor of electricity; a kingdom is lost because a monarch is intemperate; and so and so forth.[14] And yet a full account of how these various events come about can be given in terms of the dispositions's categorical bases rather than the dispositions themselves. It is this point that lies behind the familiar doctrine that dispositions are, as it is sometimes put, causally impotent.[15] How then can they be causally relevant?

---

[13] Examples such as these abound in the literature, but the focus is most often not so much on whether causal relevance among properties is transmitted over supervenience, but rather on whether it is events concretely conceived or whether it is property-like entities (be they called 'events' or not), which are the relata of the causal relation. See, e.g., Kim, 'Events as Property Exemplifications', Lewis, 'Events', Alvin Goldman, *A Theory of Human Action*, Princeton: Princeton University Press, 1970, and Sanford, 'Causal Relata'.

[14] Why does the first example in this list have so much less force than the others? Because being fragile is in part defined in terms of a certain relation to breaking; in consequence, being told that a glass broke because it is fragile is not particularly informative. Some have gone further and held that being fragile is no explanation. We disagree but do not pursue the point because the other examples will serve.

[15] See, e.g., Roger Squires, 'Are Dispositions Causes?' *Analysis*, 29, 1 (1968): 45-7, and Elizabeth Prior, Robert Pargetter, and Frank Jackson, 'Three Theses about Dispositions', *American Philosophical Quarterly*, 19, 3 (1982): 251-57.

It can be tempting to think that there is a simple solution to this puzzle.[16] A dispositional property may be properly invoked in a causal explanation despite its impotence provided that its categorical basis is causally relevant to what is being explained.[17] This is the analogue for dispositions to the identity solution to our problem discussed above, and we could simply repeat, suitably modified, our objections to that solution. However, there happens to be a simple and decisive counterexample to the solution as applied to dispositions. It is the case of conductivity.[18] The categorical basis in metals of the different dispositional properties of electrical conductivity, thermal conductivity, ductility, metallic lustre and opacity is essentially the same, namely, the nature of the cloud of free electrons that permeates the metal. Nevertheless, the person who dies because she allows her aluminium ladder to touch power lines does not die because her ladder is a good conductor of heat, or because it is lustrous or ductile or highly opaque; she dies because her ladder is a good electrical conductor. Although one and the same property is the categorical basis of all these dispositions, out of these dispositions it is only being a good electrical conductor which is causally relevant to her death. This is a contingent fact, of course. It might have been the fact that the ladder obscured someone's view which was crucial, in which case the ladder's opacity would have been the causally relevant property; or it might have been the opacity together with the good electrical conductivity which was the real problem, or . . . . The point of importance for us is that the fact that there is one categorical basis for the various dispositions does not mean that the various dispositions are alike in causal relevance.

We propose in place of an 'identity theory', that the causal relevance of dispositions can be captured in terms of what might be called invariance of effect under variation of realization. Here is a simple non-dispositional example to illustrate the central idea. Smith takes ten grains of arsenic which causes him to die about ten minutes later. Jones takes ten grains of arsenic which causes him to die in about ten minutes also. When is it right to say that the fact that they both died in about the same time is explained by the fact that they both took the same amount of arsenic? Well, suppose the time to die after taking a given dose of arsenic is given by a compli-

[16] One of us was tempted, see Prior, Pargetter, and Jackson, 'Three Theses about Dispositions'.

[17] We agree with D. M. Armstrong, *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul, 1968), p. 85f, though for reasons different from his, that it is necessarily true that a disposition has a categorical basis. However, the argument needs only the weaker doctrine that there is in fact a categorical basis. Moreover, we can regard the term 'categorical basis' as a tag phrase, and so do not need to buy into the debate about exactly how 'categorical' it must be.

[18] We owe the example to Peter Menzies, 'Against Causal Reductionism', who owes it in turn to David Lewis. We are much indebted to them for it. They should not be held responsible for the use we make of it.

cated formula involving body weight, and that this formula gives in the case of Smith and Jones very different times to die for a given identical dose *except* when the dose is ten grains. In that case the explanation of their taking the same time to die would be their both taking ten grains, and not their taking identical doses. After all, if the one and only case where the same dose is followed by the same time to die is the single case where the doses are both ten grains, it is a fluke — the fluke that the sameness in doses happened to be constituted or realized by their both taking ten grains — that the sameness in dose was followed by their taking the same time to die. Only if its being ten grains in both cases does not matter to their dying in about the same time, that is if they would take about the same time to die after the same dose pretty much regardless of the dose provided it was lethal, is it correct to explain their dying in the same time as being due to the doses being the same. We can view the matter in terms of realizations. There are many ways of realizing taking the same doses — by both taking ten grains, by both taking nine grains, ... If any of a good range of these realizations, including the actual one, would lead to death in the same time for each person, then it is correct to explain the sameness in times to die in terms of the sameness of doses, and taking the same dose is causally relevant to dying in the same time. For then the doses each being 10 grains is not what is crucial for Smith and Jones dying in the same time, but rather the doses each being the same number of grains.

We suggest a similar approach to causal explanation by citing dispositional properties. The reason being a good conductor of electricity is causally relevant to Mary's death is that it did not matter (within reason) what the categorical basis of that disposition was, for provided the causal role definitive of good electrical conductivity was occupied by a state of the ladder she would have died. We move from the non-contentious causal relevance of the categorical basis to the causal relevance of the disposition via the facts that (a) the actual categorical basis was causally relevant to the death by electrocution, and (b) had the good electrical conductivity of the ladder had a different categorical basis, then that basis would have been causally relevant to the death. And, of course, the reason opacity, say, is not causally relevant to her dying is that it might easily have been realized without her dying — as would, for instance, have been the case had the ladder been wooden.

The explanatory interest of an explanation in terms of a dispositional property is now clear. We are often interested not merely in how something in fact came about but also in how it would have come about. That is why, paradoxically, we can sometimes improve an explanation by, in a sense, saying less. An elevator has a safety device which holds it at a given

floor if more than ten people step into it at that floor. Twenty people step into it on the ground floor and as a result it does not move. In explaining what has happened to the disappointed customers, it will be better for me to say that the reason that the elevator is not moving is because more than ten people stepped into it, than to say that it is not moving because twenty people stepped into it. How so — after all, that twenty people stepped into the elevator entails, but is not entailed by and so is logically stronger than, that at least ten people stepped into it? The answer, of course, is that in giving the explanation in terms of at least ten, I tell the customers what would have happened had, say, fifteen people stepped into the elevator.

Our account of how functional properties, and so in particular content, can be causally relevant to behaviour will by now come as no surprise. A certain piece of behaviour will have a certain property, say that of being in the direction of a certain cup of coffee, as a result of the concatenation of very many neurophysiological states which will have given rise to that piece of behaviour by virtue of their natures, that is, by virtue of the neurophysiological properties they instantiate. But, of course, there will be other ways that behaviour with the property of being towards the coffee could have been caused, other neurophysiological ways, or even, other non-neurophysiological ways if we allow ourselves Martian speculations. Is there anything interesting that we can say about resemblances between these various actual and possible ways of getting behaviour towards the coffee? The answer is that it may be that many of these ways, including the actual way, are united by the functional properties they realize, and in particular by the functional properties definitive of contents that they realize. In that case, an explanation in terms of content-bearing states will apply and its explanatory interest will lie in the fact that it tells us about what would happen in addition to what did happen. That is how the content properties may be causally relevant.

Have we really laid the demon to rest: the metaphysics of causation?

The intuition that functionalist accounts of content make content epiphenomenal is a strong one.[19] We have encountered the following response to our defence of the causal relevance of content properties. "You have shown how the fact that a certain piece of behaviour follows the instantiation of certain content properties need not be a fluke. For (a) it is not a fluke that the behaviour follows a certain concatenation of neurophysiological states, (b) this concatenation is, at the least, a major part of what the relevant functional properties supervene on, and (c) it may be that many different complexes of neurophysiological states alike in the having

[19] As Ned Block and Paul Boghossian convinced us.

the relevant functional properties supervening on them would also be followed by behaviour exemplifying the feature we are seeking to explain. (Often the behaviours will count as different under some natural taxonomy, but this is, of course, consistent with their being alike in the respect of interest). But all that that shows is the non-flukey nature of a certain sequence, and the explanatory value of content ascriptions. It does not show that content properties conceived functionally do the *driving* of behaviour. The fact remains that that is done by neurophysiological (or least relatively intrinsic structural or syntactic) properties; yet surely the commonsense intuition that cries out for vindication is that content drives behaviour."

Now of course it is true that some non-flukey sequences are not causal. That possibility lies at the heart of classical epiphenomenalism. According to epiphenomenalism, a certain kind of mental event regularly precedes a certain kind of brain event which leads on to the behaviour we associate with that mental event; but this is not because the first event causes the second but because both are caused by a third, earlier brain event. But it is essential to this story that according to classical epiphenomenalism the mental event is indeed caused by the earlier brain event. But if caused then distinct, whereas a key part of our account of how content properties are causally relevant to behaviour is that they are not completely distinct from the relevant neurophysiology; they instead supervene on it. We would be in trouble if our story was that the neurophysiological properties are causally relevant both to the content properties and to the behaviour. But our view is rather that the connection between neurophysiology and content is that the latter supervenes on the former, and supervenience is incompatible with causation. More precisely, enough by way of neurophysiology and the relevant laws together possibly with environmental setting and history (how much of the latter two you need to include depends on whether and to what extent the content is broad) *logically* fixes the content, and therefore is not *causally* responsible for it. Accordingly, as the neurophysiology is a proper part of what the content logically supervenes on (we might put this by saying that the content contingently supervenes on the neurophysiology), the neurophysiology is not causally relevant to the content. This is why the content is not possessed a moment after the relevant neurophysiological facts obtain, as would have to be the case were the connection causal.

Nevertheless, there is more to say about the objection, for behind it lies an attractive view about the metaphysics of causation.[20] Suppose that in a laboratory in Russia electron A is acted on by a force of value four and

---

[20] One author (F. J.) finds it more attractive than does the other (P. P.).

accelerates at rate seven (all in some suitable units). At the same time in a laboratory in America electron B is also acted upon by a force of value four and as a result it too accelerates at rate seven. Suppose that the sameness of the resultant accelerations is in no way dependent on the fact that the impressed forces were of value four. All that mattered (within limits, of course) for the sameness of the accelerations was the sameness of the impressed forces. Then clearly the sameness of the impressed forces is causally relevant to the sameness of the resultant accelerations. The sameness of the first causally explains the sameness of the second. (The situation is in essentials the same as in the arsenic example described earlier.)

Suppose, however, we think of causation as a matter of production or efficacy which does not reduce sooner or later to nothing more than nomological sequence: according to this view, a sequence is nomological because of underlying causal productivities, not conversely.[21] Then it is plausible that in some sense the sameness of the impressed forces does not actually *produce* the sameness of the resultant accelerations.[22] Consider electron A. It is acted upon by a force which both has the property of taking the value four and the relational property of being the same in value as the force acting on electron B. Does the latter fact actually have any influence on the way the electron moves off under the impact of the force? Surely not. All the work is done by the force acting on A taking the value four; how things are with B, which after all is a very long way away, is surely in *some* sense irrelevant. Perhaps the sharpest way of putting the point is the Occamist one. Supposing that the force taking the value four produces the acceleration of value seven in both cases is enough to explain (because it entails) the fact that sameness of impressed force on A and B is in fact followed by sameness of acceleration by A and B. There is no need in addition to give having the same impressed force per se a productive or efficacious role with respect to the sameness of the resultant accelerations.[23]

---

[21] This is a view forced on us if (but *not* only if) we accept the non-Humean idea that there can be strongly singularist causation in the sense of one event causing another which does not fall under a law, either deterministic or indeterministic. For defences of strongly singularist causation, see G. E. M. Anscombe, 'Causality and Determination', in *Causation and Conditionals*, E. Sosa, ed. (Oxford University Press, 1975), pp. 63-81, and Michael Tooley, 'The Nature of Causation: A Singularist Account', forthcoming.

[22] Of course, from a purely nomic point of view, and provided the details are sufficiently filled out, the sequence: same forces, same accelerations, may be just as 'good' as the sequence: force four, acceleration seven.

[23] We are here in agreement with Block, 'Can the Mind Change the World?'.

The idea then is that we can distinguish as a special case of causal relevance among properties, causal efficacy. Every case where an instance of F is causally efficacious with respect to an instance of G is a case of causal relevance, but some cases where an instance of F is causally relevant to an instance of G are cases of relevance without efficacy.[24] The objection under discussion can now be put as follows: our defence of the explanatory role of content from the functionalist perspective only shows causal relevance (and indeed we used that very term to describe matters earlier); it does not show that content properties are causally efficacious with respect to behaviour, and it is the latter which is integral to the common intuition about content's role with respect to behaviour.

Our reply turns on the point that the Occamist thought that lies behind distinguishing causal efficacy as a special case of causal relevance has far-reaching ramifications. It has been widely noted how plausible is the idea that everything about the way our bodies move, including everything by way of the causal relations involved, supervenes on how things, including the laws, are at the most fundamental micro-physical level. If this is right, then the Occamist attitude combined with the view of causation which does not reduce it to nothing more than nomological sequences, enjoins us to restrict relations of causal efficacy to certain properties in fundamental Physics — which properties exactly is a matter for empirical science — and to see all the causal relevancies 'higher up' as, strictly speaking, non-efficacious.[25] For we do not need to believe in any fundamental efficacies over and above those between properties at the micro-level in order to explain the regularities, actual and counterfactual, all the way up, because supervenience tells us that they are fixed by how things are at the bottom (*if* there is a bottom). But then the neurophysiological properties are not causally efficacious in the special sense any more than are the content properties. And more generally there will not be a contrast between the causal relationship that content and functional properties generally have to behaviour, and the causal relationship that taking arsenic has to death, that lying in the sun has to getting hot, that rising inflation has to falling living standards, and so on and so forth. These cases will be all alike in being cases of causal relevance without causal efficacy. Ergo, the functionalist account of content does not downgrade its causal role, rather it leaves

---

[24] Elsewhere, we refer to cases of causal relevance without causal efficacy as cases of causal programming, see 'Functionalism and Broad Content'; see also 'Program Explanation: A General Perspective' *Analysis*, forthcoming.

[25] For a defence of the view, to put it in our terms, that the answer science delivers is that causal efficacy is a relation between forces, see John Bigelow and Robert Pargetter, 'The Metaphysics of Causation', *Erkenntnis*, forthcoming.

it in the excellent company of everything except for certain members of that most exclusive of clubs, the properties of fundamental physics.[26]

We suspect that the thought behind the view that we functionalists have made content epiphenomenal is that we have somehow taken the 'push' out of content. But consider someone being torn apart by an imbalance of forces acting on him (as happens if you step into Space without a space suit on). The imbalance of forces has plenty of push but plausibly is not efficacious, for the simple reason that it is a 'convenient fiction'. It is plausible that the resultant force in the familiar parallelogram of forces is a convenient fiction. It is the component forces which really exist (or rather certain of the component forces, the component forces in a parallelogram of forces can of course themselves be resultants in some other parallelogram of forces), and so it is they at most which can stand in relations of causal efficacy.[27]

Application to an Argument for Eliminativism

We can now see the mistake in an interesting and initially appealing line of argument for eliminativism about the propositional attitudes.

Eliminativists see the apparatus of beliefs and desires with their associated contents as part of an ancient (and so prima facie suspect, but that is another story) theory — dubbed 'folk psychology' — which we invoke to explain and predict inter alia and especially behaviour. But to explain behaviour is to say something about the causes of behaviour and, runs the argument we wish to reply to, what folk psychology says about the causes of behaviour may turn out as a matter of empirical fact to be mistaken in an important respect, a respect important enough to justify describing what has happened as a refutation, rather than, say, an elaboration, of folk psychology. What is meant here by 'as a matter of empirical fact' is

---

[26] We are not, of course, saying that most of our commonsense convictions about causal connections expressed in everyday language are false. When we use terms like 'efficacious' and 'productive' in everyday talk, they mean roughly what we are using 'causal relevance' for (perhaps restricted to causal relevance between relatively intrinsic properties, see Lewis, 'Events'). Our thesis is a thesis in (a posteriori) Metaphysics which holds, not that most of our convictions are mistaken, but rather that what makes the true ones true is a relation between properties in fundamental Physics. We take this general way of looking at the matter to be consonant with D. M. Armstrong's species of realism about universals, as expressed for instance in his *A Theory of Universals* (Cambridge: Cambridge University Press, 1978), see particularly chap. 24 for the connection with causality. What becomes of the doctrine that dispositions are causally impotent on this metaphysics of causation? *If* 'causally impotent' is given the special sense given to 'causally inefficacious', then the doctrine is true; but it is also true that a disposition's categorical basis is impotent unless specified micro-physically.

[27] On the existence of component forces, see John Bigelow and Robert Pargetter, 'Forces', *Philosophy of Science*, forthcoming.

not as a matter of abstractly possible empirical fact — it is common ground (or ought to be) that it is logically possible that the causal story about our behaviour be incompatible with folk psychology. What these eliminativists have in mind is the causal story implied by certain connectionist views about information processing in the brain, which they take to be very much live options. Eliminativists see folk psychology as committed to beliefs and desires being properly described as propositional attitudes. This combined with the idea that folk psychology is an explanatory theory leads to the doctrine that the folk are committed to the idea that the internal causes of behaviour can be illuminatingly divided up in terms of the propositions which are the objects of our beliefs and desires. Folk psychology carries with it its own way of taxonomizing the causes of behaviour in terms of contents given typically by indicative natural language sentences prefixed by 'that' — propositional modularity, as it is sometimes called. The eliminativist argument is that if developments in neuroscience confirm certain connectionist views, then this will show that propositional modularity is false, and so will be nothing less than an empirical refutation of the folk taxonomy of the causes of behaviour, and so of folk psychology with its apparatus of beliefs and desires.[28]

One might quarrel with one or another detail of the eliminativists's account of folk psychology, but the general picture is highly plausible. For consider Jill, who believes that a book relevant to her current research has arrived in the library and also believes that it will rain later today. We folk do distinguish these two beliefs precisely because they differ in content, and that is a matter at least very closely connected with the propositions expressed by the embedded sentences.[29] And further we do distinguish the causal role that the two beliefs play with respect to her behaviour. Unless we have reason to attribute somewhat bizarre desires to Jill, the belief about a book relevant to her current research is most likely to be appealed to in order to explain her going to the library, and the belief that it will rain later is most likely to be appealed to in order to explain her taking an umbrella to work.

---

[28] The most explicit development of this argument that we know is in William Ramsey, Stephen Stich and Joseph Garon, 'Connectionism, Eliminativism and the Future of Folk Psychology', in *Philosophy and Connectionist Theory*, W. Ramsey, D. Rumelhart, and S. Stich, eds. (New Jersey: Erlbaum, forthcoming), but see also Paul Churchland, *Scientific Realism and the Plasticity of Mind* (Cambridge University Press, 1979), § 18 ff, and 'Eliminative Materialism and the Propositional Attitudes' *Journal of Philosophy*, 78, 1981, 67-90.

[29] See David Lewis, *On The Plurality of Worlds* (Oxford: Blackwell, 1986), for arguments that the connection between the objects of beliefs and the embedded sentences in our reports of belief is more complicated than one might have hoped.

Our reply to the eliminativist argument takes this general picture for granted. We grant that we folk distinguish the two beliefs by distinguishing their propositional objects and that we folk give the distinguished beliefs distinct causal roles in explaining Jill's behaviour precisely in accord with their distinct propositional objects (and the same goes for desires, of course). Our quarrel is with the claim that there is an incompatibility between this picture and certain connectionist views about information processing in the brain.

Why do eliminativists see an incompatibility between connectionism and folk propositional modularity? Well, if beliefs are anything like stored sentences in the brain, then it is plausible that there will be in Jill's brain two distinct bits of storing, one of the sentence about the book, the other of the sentence about the rain, and eliminativists observe that consequently in this case we can sensibly suppose that the first bit of storing has a special causal relationship to Jill's movements towards the library that the second lacks, and that the second bit of storing has a special relationship to her umbrella-taking that the first lacks. But, runs the argument, if certain versions of connectionism are correct, it will be impossible to isolate in any way one part of the brain or its activities and see this as one of the beliefs, at the same time as isolating something else in the brain and seeing that as the other belief. Information processing is a completely holistic and distributed matter on these versions of connectionism. There will be nothing in the brain, at the neurophysiological level or at the level of cognitive architecture, to be isolated as one belief as opposed to the other.[30] How then can the folk hypothesis about distinctness of causal roles be true?

Our reply is that our approach to the causal relevance of content properties in terms of invariance of effect under variation of realization of that content, shows how one and the same underlying state (be it widely distributed or localized) can realize two different contents, one of which is, and the other of which is not, causally relevant to a given piece of behaviour. We do not need to find distinct states at, say, the brain level — distinct encodings, or whatever — to be the two beliefs in order to vindicate the commonsense conviction that my belief that p may differ in its causal relevancies from my belief that q. For on our approach, a certain content is causally relevant to a certain effect if (a) a state occupying the role definitive of that content is causally relevant to that effect, and (b) had that

---

[30] See Ramsey et al., 'Connectionism, Eliminativism and the Future of Folk Psychology', and Andy Clark, *Microcognition* (Cambridge: MIT Press, 1989), and the references therein to the connectionist literature. Clark is with us in denying that connectionism implies eliminativism.

content been differently realized, then other things being equal the counterfactual realizer state would have been causally relevant to that effect. Now it is clearly a live possibility that a single state S be such that it occupies the role definitive of different contents, C_1 and C_2, and yet for some effect E other ways of realizing C_1 would ceteris paribus have been causally relevant to E, whereas other ways of realizing C_2 would not. Indeed, the dispositional correlate of this live possibility actually obtains in the case of conductivity discussed earlier (and remember we noted that it is wrong to conclude from the fact that we cannot pick out distinct underlying states of Mary's ladder to be the bases of the various distinct dispositional properties of her ladder, that all the dispositions are equally causally relevant to what happened to her).

This story fits well with our everyday approach to our case of Jill who believes both that a book relevant to her current research has just arrived in the library and that it will rain. Although she has both beliefs, we take it that it is the first that is causally relevant to her movement towards the library, because we take it that she would have moved towards the library whether or not she had had the belief that it will rain; whereas she would have taken an umbrella whether or not she had believed that a book especially relevant to her research had just arrived in the library, and so this belief is not why she took an umbrella.[31]

It might well be objected that when we explain Jill's movement towards the library in terms of her belief that a book especially relevant to her research has just arrived, we are giving that belief an active role in the story.[32] It is not a standing condition but rather a state being activated in the context of a set of standing conditions. The same point applies when we explain Fred's survival in terms of his seat belt having the right degree of elasticity. The seat belt presumably had that right degree of elasticity from the day of its manufacture, but something happened at a certain moment during the accident which brought that degree of elasticity into play in a way which led to his survival. A fair question, therefore, is whether in the supposed connectionist case where there is no isolating one belief from the other in different encodings, we can give Jill's belief that a book especially relevant to her research has just arrived in the library an active role in explaining her movement towards the library without at the same time giving the intuitively irrelevant belief that it will rain the same role? But consider our ladder example again. Mary's aluminium ladder was a good conductor of electricity from the day it was made. When its

---

[31] We are supposing that the important problems of overdetermination and causal preemption are separate ones from those under discussion here.

[32] As Andy Clark reminded us.

being so was actively causally relevant to her death by electrocution what happened was that the nature of the cloud of free electrons in the matter of the ladder occupied the role definitive of being a good electrical conductor in a special way. The set of subjunctive conditionals definitive of that role contained a member which, in addition to being itself true, had a true antecedent and a true consequent. It was true that for some salient-to-being-an-electrical-conductor conditional 'had so and so happened, then such and such would have happened' that, not only was it the case that it obtained by virtue of the nature of those electrons, in addition, on the occasion in question, so and so actually happened and, by virtue of the nature of those electrons, such and such followed in a way which contributed to Mary's death. In brief, the disposition manifested itself; and the crucial point is that, although the underlying basis for being a good electrical conductor in the ladder is one and the same as that for, for instance, being opaque and being a good heat conductor, being a good electrical conductor can, and did in the case we are imagining, manifest itself without the other dispositions manifesting themselves. The same can be said in the case where Jill's belief about the book played an active role in getting her to the library. It did so by virtue of certain of the inputs and outputs salient in the specification of the functional role corresponding to having that belief actually obtaining, and, of course, that can happen without the differently specified inputs and outputs constitutive of having the belief that it will rain actually obtaining.

Conclusion

Functionalism specifies mental properties in terms of causal roles. The irony is that it then appears to be the case that functionalism deprives mental properties of causal relevance. It appears that it is the properties in virtue of which the relevant states occupy the relevant causal roles, and not the roles themselves, which are causally relevant to behaviour. Our aim in this paper has been to rebut this beguiling argument, and to do so in a way which shows the flaw in the equally beguiling argument that connectionism supports eliminativism.[33]

---

[33] This paper arose out of discussions engendered by the notion of a program explanation in 'Functionalism and Broad Content'. In addition to the acknowledgments already made, we can remember the changes forced by talking to Martin Davies and Robert Pargetter. No doubt there are more than we can remember.