MENTAL CAUSATION

Tim Crane
Department of Philosophy
UCL
Gower Street
London WC1E 6BT
Tel. 020 7679 3074
Fax 020 7209 0554
tim.crane@ucl.ac.uk

Keywords: action, dualism, functionalism, materialism, physicalism

Contents

It is arguably an assumption of both commonsense and scientific psychology that mental states and events cause events in the physical world. Yet this fact is problematic both for physicalist and dualist theories of the mind.

Introduction

Does the mind have effects in the physical world? To believe it does is to believe in mental causation. It can be argued that we are committed to the existence of mental causation when we explain people's actions in terms of their thoughts, beliefs, intentions, desires and other propositional attitudes. For example, we might say that Jenny drank the whisky because she thought it would calm her nerves. To say that there is mental causation in this case is to say that the 'because' expresses a causal relation between Jenny's thought and her action, just as it does in non-mental cases, as when we say that the bridge collapsed *because* the bomb exploded beneath it. In other words, the thought, like the explosion of the bomb, is *causally efficacious*. So understood, mental causation is ubiquitous. Whenever we do something or think something because of something going on in our minds, this is a case of mental

causation. But what is the nature of this causation, and why have philosophers found it so problematic?

1. What is mental causation?

Mental causation is when a mental state (like a belief or intention) or a mental event (like an experience) has an effect, either a mental effect (like another thought or experience) or a physical effect, an effect in the physical world. In any case of causation, we can distinguish between the *relata* of causation (what is being related) and the *relation* itself. So, for example, when the explosion caused the bridge to collapse, we can distinguish between the cause (the explosion), the effect (the collapse) and the relation itself (causation) which links these two events. To say that there is mental causation is to say that at least one of the relata of some particular causal interaction is mental; just as to say that there is physical causation is to say that at least one of the relata of some causal interaction is physical. It is not necessarily to say that there is a distinctive kind of relation – a distinctive kind of *causation* – which holds in the cases where a mental entity is a cause. This is a possible position; but it is not required merely by the idea of mental causation. It is a consequence of this that we should not commit ourselves at the outset to a conception of causation (e.g. that it must involve contact action) which renders mental causation impossible to understand.

What, then, is it that mental and physical causation have in common? What is it that makes them both cases of causation? The answer to this question depends on the correct theory of causation, and it is important to emphasise that few theories of causation entail that causation must be a physical relation. Some theories say that A causes B when there is a Law of Nature linking A-type and B-type events; others that

A causes B when B is counterfactually dependent on A (i.e. if A had not existed, B would not have existed); and others say that A causes B when the probability of B is higher in the presence of A than it would have been otherwise (for all these options, see Sosa and Tooley 1991). Other theories deny that causation is a relation at all (Mellor 1995). But however they differ, these analyses can apply equally well to mental as to physical causes and effects. Indeed, no one analysis needs to be assumed in stating the problem about mental causation in section 3 below.

As well as discussing the nature of the causal relation, theories of causation also discuss what kind of entities are the relata of causation; that is, what kinds of entities are causes and effects. Some theories say that causes and effects are always events (like the explosion of the bomb or Jenny's drinking the whisky) while others say that they are facts (like the fact that the bomb exploded or the fact that Jenny drank the whisky) or states (the state of Jenny's having drunk the whisky). Others express this distinction as one between events and properties of events; events have many properties, but only some properties of events are 'causally efficacious'. Theories of 'agent causation', by contrast, claim that the fundamental phenomenon of mental causation is when agents, rather than their states or events involving them, cause things to happen – as when *John* breaks the window by smashing it. In what follows, we will consider only causation by events, or states/properties (where a state is understood as a thing's having a property at a time).

Sometimes mental causation is an essential part of a metaphysical theory of mind. So it is with *functionalism*, whose characteristic thesis is that mental states are individuated, or distinguished from one another, by the causal roles they play (see Block 1980). A functionalist holds that belief, for example, is the sort of state that is typically caused by perceptions and other beliefs, and is disposed to cause actions in

conjunction with desires. Functionalism therefore assumes that mental states are causes and effects – the mind is a causal mechanism – though there are various accounts of what this actually means.

2. History

Debates about the causal powers of the mind can be traced back to antiquity; but in their contemporary form they derive from Descartes's influential theory of mind and body. Descartes was a dualist – he thought that mind and body were distinct substances. For Descartes, a substance is a being which is capable of independent existence, one whose existence depends on nothing else. So to say that mind (or soul) and body are distinct substances is to say, among other things, that they are capable of independent existence.

Descartes was criticised in his lifetime for making mental causation hard to understand, most famously by Princess Elisabeth of Bohemia (see Descartes 1985 vol.3). Princess Elizabeth asked how substances so different as minds and body could affect one another; Descartes claimed not to see the difficulty, and the debate between them was left unresolved.

A more effective criticism of Descartes's dualism came from Leibniz. A central thesis of Descartes's physics is that matter is a substance whose characteristic (essential) attribute is extension in space. God has endowed matter with a certain *quantity of motion*, and the total quantity of motion is preserved in all physical interactions: an interaction never diminishes or adds to the total quantity of motion in the world. Thus Descartes believed in the conservation of quantity of motion, but also believed that mental causation was consistent with this law of nature. His reasoning behind this was that the mind causes things to happen in the body by changing the

direction of motion of the animal spirits (a rarefied form of matter) at the pineal gland in the brain. So the mind can change the direction of motion of matter and not alter the total quantity of motion: mental causation is consistent with the conservation laws, as Descartes understood them.

Leibniz did not challenge the validity of this reasoning, but the correctness of Descartes's conservation laws. According to Leibniz, what is conserved in the physical world of matter is not quantity of motion but quantity of *momentum*, mass times velocity. Since velocity is a vector of speed and direction, the mind cannot alter the direction of motion of the animal spirits without altering the quantity of momentum in the physical world. Therefore mental causation is inconsistent with the correct conservation law: the conservation of momentum (see Woolhouse 1993, Chapter 9).

Leibniz's alternative to Descartes's dualism was his doctrine of pre-established harmony, sometimes called *parallelism*. This is the view that mind and body do not interact causally, but operate in parallel (hence: harmony) in accordance with the will of God who initiated the harmony (hence: pre-established). This doctrine is a form of *epiphenomenalism*: the view that the mind has no effects in the physical world. Another form of epiphenomenalism is the *occasionalism* of Malebranche, which holds that the mind cannot act in the physical world on its own, but needs the help of God's action on each occasion of interaction. Each movement of the body by the mind is, in effect, a miracle. Epiphenomenalism need not deny that there are causal relations between mental phenomena and other mental phenomena. But it must deny that there are any causal relations between mental phenomena and matter.

The naturalistic philosophy of the 19th and 20th centuries did not generally see mental causation as a problem. Many naturalists are materialists, and materialists

identify the mind with something material, the brain. By identifying the mind with the brain, materialism can allow mental causes to cause material things, because mental causes are just a species of material cause. In the 20th Century, sometimes the term 'physicalism' was used as a synonym for materialism, while sometimes the term was meant to indicate the special ontological and epistemological authority which *physical science* has in telling us about the material world. The supposed difference between materialism and physicalism could then be put like this: materialism holds that everything is matter, whereas physicalism says that everything is physical, where being physical is being the subject-matter of physical science. Therefore, if physics talks about various things which are not matter, physicalism can recognise the existence of something which materialism cannot. Since arguably the fields and forces of contemporary physics are not matter in any normal sense, physicalism seems to be the preferable theory. In what follows, therefore, I will talk of physicalism rather than materialism.

3. Mental causation as a problem for dualism

The problem of mental causation which originated in the 17th century re-emerged in the 20th century as part of the arguments for a specific form of physicalism, the identity theory (see Feigl 1958). Defenders of the identity theory argued that there were no philosophical, *a priori* objections to identifying mental phenomena with states of the brain; the truth of this claim must be established empirically. Identity is here understood literally: the claim is that a mental state is the very same thing as a state of the brain. ('Pain = c-fibre firing' became a common, though empirically false, way of illustrating the claim.) Later theories went further, and argued that the identity theory could be demonstrated by philosophical argument, rather than simply shown to

be coherent (see Lewis 1966, Armstrong 1968, Davidson 1970). The general form of this argument is as follows:

(1) Premise one: mental causes have physical effects.

(2) Premise two: the physical world is causally closed; that is, all physical effects have physical causes which are enough to bring them about. (I shall call this 'the completeness of physics'; see Papineau 2000)

(3) Conclusion: mental causes are physical causes.

Different proponents of the argument elaborate and defend it in different ways, to make it strictly valid. So, for example, some say that an extra premise denying the existence of mental-physical *causal overdetermination* must be added. (Overdetermination is when an effect has two or more causes, each of which is enough to bring the effect about, and each of which would have brought it about if the other(s) hadn't.) Others say that the second premise must be re-formulated to make it compatible with indeterminism, since as it stands it is a deterministic claim. And others say that the first premise must be definitive of the nature of mental states, not just a fact about them (Lewis 1966). But here we can put to one side these clarifications of detail, and restrict ourselves to the general form of the argument.

The general form of the argument is that in order to reconcile mental causation with the completeness of physics, we have to identify mental and physical causes. So if all mental phenomena have some physical effects – a widely held assumption, but an assumption nonetheless – then all mental phenomena are physical phenomena. The reasoning behind this argument is simple: if there are mental causes of physical effects, then how is this compatible with these effects having adequate physical

causes, as the completeness of physics says they must? Either, it seems, the completeness of physics is false or epiphenomenalism is true. In other words, if the completeness of physics is accepted, then mental causation is a deep problem for mind-body dualism. The problem is resolved, it seems, by identifying the mental and the physical causes.

Can a dualist respond to this problem? Is physicalism the only adequate response? Perhaps the dualist can deny the premises. The first premise of the argument is the existence of mental causation. As we have seen, a dualist could deny this premise by being an epiphenomenalist. But epiphenomenalism is very hard to believe: the view that our minds make our bodies move does not seem to be a theoretical claim, but a datum that theory should account for. Can a dualist deny the completeness of physics? Here matters are more complicated. The completeness of physics is not normally understood as a law of physics (like Newton's laws or the Schroedinger equation) but as a metaphysical speculation based on the laws of physics. A dualist could deny that this speculation is a consequence of the laws of physics. This is widely thought to be contrary to received opinion among philosophers of science; but the issue is still controversial (see Papineau 2000 and Cartwright 2000 for opposing perspectives).

The physicalist conclusion is that mental causes are identical with physical causes. So long as all mental phenomena have some physical effect at some point, then physicalists can conclude that each mental phenomenon is identical with some physical phenomenon. This is an *identity theory* of mind and brain. There are two types of identity theory: the 'type identity theory', which identifies mental properties or types, and the 'token identity theory', which identifies mental tokens or particulars. Which identity theory one accepts might depend on one's views of the relata of

causation (see above): if one held that properties or states are causes, for instance, then one would conclude that the type identity theory is true (Lewis 1966), but if one held that events were causes, then one would hold a token identity theory (Davidson 1970).

4. Mental causation as a problem for physicalism

Since one of the general motivations for a physicalist theory of mind derives from the causal role of the mind, it is surprising therefore to discover that mental causation creates problems for physicalism as well as for dualism. The reason for this is that there is a form of physicalism (called 'non-reductive physicalism'), which denies the identity theory. Since the identity theory was what enabled physicalists to solve the problem of mental causation, it is not surprising if those physicalists who reject the identity theory encounter that problem in a new form.

Some physicalists deny the identity theory because it entails the thesis that all creatures who are in the same mental state must be in the same physical state too, and this thesis is empirically implausible, given the diversity of organisms. Consider, for example, the variety of creatures who are capable of being in pain, and the variety of their physical constitution, and then consider how likely it is that all these creatures share a physical state or property when they are in the same mental state (the point derives from Putnam 1975a). Therefore, non-reductive physicalists say that we should not *identify* mental properties or states with physical properties or states. But they nonetheless endorse a weaker form of physicalism, to the effect that all particular objects and events are physical, even if not all properties and states are physical. (This is the so-called 'token identity theory' mentioned in section 3 above.) The resulting view is non-*reductive*, because it does not 'reduce' mental states to physical states, as

the type-identity theory does, by identifying them; but it is still *physicalism* because it gives an ontological priority to the physical in saying that all particular objects and events are physical. There are no non-physical objects or events.

How does this affect the question of mental causation? This depends on how non-reductive physicalism regards the relata of causation. If causation is a relation between events, then non-reductive physicalism has no difficulty accounting for mental causation in physicalist terms, since all events are physical, even if not all properties are (this is Davidson's (1993) position). But some philosophers argue, for reasons independent of the philosophy of mind, that properties or states are causes, not events. One reason for holding this is from reflection on commonsense examples: if throwing a brick broke a window, then it is not the event of throwing the brick *as such* which had this effect, but rather the throwing of a brick with certain properties (its weight, its velocity etc.). For, if the brick had been made of rubber, or had been thrown with less force, it might not have broken the window. Therefore, it is concluded that strictly speaking, causes are properties or states (i.e. things having properties); or, to put it another way, causes have their effects *in virtue of* their properties. But if causes are states/properties, then non-reductive physicalists must deny the identity theory of mental and physical causes, and therefore they cannot employ the argument discussed in section 3 above. If they are not epiphenomenalist, then they must accept premises one and two and reject the conclusion.

To put it another way: suppose there is mental causation, and the completeness of physics is true. And suppose properties/states are causes, and that the identity theory is false. Then it is hard to see how there can be mental causation in the light of the completeness of physics, *even if* every mental event is a physical event. This is the

problem of mental causation for non-reductive physicalists (see Heil and Mele 1993 for a variety of statements of this problem, and responses to it).

Non-reductive physicalists have tended to respond in one of two ways to this problem; either by developing the notion of causation involved in the debate, or by developing the doctrine of physicalism. Those who wish to develop the notion of causation might say, for example, that mental causes are causally *relevant* to physical effects, although not causally *efficacious*. (For similar ideas, see Dretske (1988) and Jackson and Pettit (1988).) One difficulty with these approaches is that it is hard to see them as more than *ad hoc* responses to the problem in hand; it can seem as if a specific notion of mental causation is being tailored simply to solve the problem. Some more ambitious approaches have therefore motivated their solution with detailed independent accounts of causation itself (Yablo 1992 is a particularly detailed attempt).

The other kind of approach takes causation for granted, but further develops the idea of non-reductive physicalism (see Loewer 2001). This approach assumes Jackson's (1998 chapter 1) definition of physicalism, employing possible worlds: any minimal physical duplicate of our world is a duplicate *simpliciter*. It also assumes that causation is counterfactual dependence between facts or states of affairs. Jackson's definition yields the metaphysically *necessary determination* of the mental by the physical: given the way the physical facts actually are, the mental facts could not have been otherwise (see also Lewis 1993). It follows that if the mental facts had been different in some way, then the physical facts would have been different, even if the mental and the physical facts are not identical. So, in particular, a mental cause M of a physical effect E causes E even though the completeness of physics guarantees the existence of a physical cause P which is enough for E – because P necessarily

determines M, as well as causally sufficing for E. If M had not been the case, then E would not have been the case, since if M had not been the case, P would not have been the case and therefore (arguably) E would not have been the case either. By appealing to this (admittedly problematic) idea of metaphysically necessary determination, physicalists attempt to solve the problem of mental causation without appealing to the identity theory.

5. Mental causation and cognitive science

Insofar as cognitive science is committed to a form of non-reductive physicalism, denies epiphenomenalism and upholds the completeness of physics, then it has to give an account of mental causation. One of the most influential theories of the foundations of cognitive science, Jerry Fodor's Representational Theory of the Mind (RTM), presupposes that mental states involve causally related sequences of mental representations, or symbols in a *language of thought.* The main argument for the RTM is based on the idea that the logical and rational relations between thoughts must have an underlying causal mechanism (Fodor 1987). The causal mechanism of such thought-processes is argued to involve mental representations with a structure that mirrors the logical structure of thoughts; the representations have a semantic and a syntactic (i.e. causal) structure.

Critics have questioned whether the RTM renders the content of thought causally idle: since the causal role of mental representations is discharged by the syntactic structure of the representations, what causal role does this leave for the content of thought? And if the content of thought is epiphenomenal, does this make it theoretically dispensable? Defenders of the RTM have responded by claiming that the causal efficacy of content is guaranteed by the fact that it *supervenes* on the syntactic

structure of the brain; that is, that there is no difference in content without a difference in syntax. But if syntactic structure is an aspect of the local physical structure of the brain, this defence puts the RTM in conflict with the widely accepted doctrine of externalism, since according to externalism (see Putnam 1975), the content of our thoughts does not supervene on the local physical structure of our brains. Fodor (1995) attempts to resolve this apparent contradiction.

References

Armstrong, D.M. (1968) *A Materialist Theory of the Mind* London: Routledge and Kegan Paul.

Block, Ned (1980) What is functionalism? *Readings in the Philosophy of Psychology* ed. Ned Block, London: Methuen.

Cartwright, Nancy (2000) The completability of science *The Proper Ambition of Science* eds. M.W.F. Stone and Jonathan Wolff, London: Routledge.

Davidson, Donald (1970) Mental events *Experience and Theory* eds. L. Foster & J. Swanson, London: Duckworth; reprinted in Donald Davidson (1980) *Essays on Actions and Events* Oxford: Oxford University Press.

Davidson, Donald (1993) Thinking causes *Mental Causation* eds. John Heil and Al Mele, Oxford: Oxford University Press.

Descartes, René (1985) *The Philosophical Writings of Descartes* (3 volumes) translated by J. Cottingham, R. Stoothof and D. Murdoch, Cambridge: Cambridge University Press.

Dretske, Fred (1988) *Explaining Behavior* Cambridge, Mass.: MIT Press.

Feigl, Herbert (1958) The 'Mental' and the 'Physical' *Minnesota Studies in the Philosophy of Science* eds. H. Feigl, M. Scriven and G. Maxwell (Minneapolis: University of Minnesota Press. (Reprinted as a monograph by the same publisher, 1967.)

Fodor, Jerry (1987) *Psychosemantics: the Problem of Meaning in the Philosophy of Mind* Cambridge, Mass.: MIT Press.

Fodor, Jerry (1995) *The Elm and the Expert* Cambridge, Mass.: MIT Press.

Heil, John and Mele, Al, eds. (1993) *Mental Causation* Oxford: Oxford University Press.

Jackson, Frank (1998) *From Metaphysics to Ethics* Oxford: Oxford University Press.

Jackson, Frank and Pettit, Philip (1988) Functionalism and Broad Content *Mind* 97 (381-400).

Lewis, David (1966) An argument for the identity theory *Journal of Philosophy* 63 (17-25).

Lewis, David (1993) Reduction of mind *A Companion to the Philosophy of Mind* S. Guttenplan ed. Oxford: Blackwell (412-431).

Loewer, Barry (2001) From physics to physicalism *Physicalism and its Discontents* Carl Gillett and Barry Loewer eds. Cambridge and New York: Cambridge University Press.

Mellor, D.H. (1995) *The Facts of Causation* London: Routledge.

Papineau, David (2000) The rise of physicalism *The Proper Ambition of Science* eds. M.W.F. Stone and Jonathan Wolff, London: Routledge.
Putnam, Hilary (1975) The meaning of 'meaning' *Mind, Language and Reality* Cambridge, Cambridge University Press.
Putnam, Hilary (1975a) The nature of mental states *Mind, Language and Reality* Cambridge, Cambridge University Press.
Sosa, Ernest and Michael Tooley (eds.) (1991) *Causation* Oxford: Oxford University Press.
Woolhouse, Roger (1993) *Descartes, Spinoza, Leibniz: The Concept of Substance in Seventeenth Century Metaphysics* London: Routledge.
Yablo, Stephen (1992) Mental Causation' *Philosophical Review* 101 (245-280).

Bibliography

Crane, Tim (1995) The mental causation debate *Proceedings of the Aristotelian Society* Supplementary Volume 69 (211-236).
Horgan, Terence (1993) From supervenience to superdupervenience: meeting the demands of a material world *Mind* 102 (555-586).
Jackson, Frank (1996) Mental causation *Mind* 105 (377-413).
Kim, Jaegwon (1993) *Supervenience and Mind* Cambridge: Cambridge University Press.
Kim, Jaegwon (1998) *Mind in a Natural World* Cambridge: Mass.: MIT Press.
McLaughlin, Brian (1992) The rise and fall of British emergentism *Emergence or Reduction?* A. Beckermann *et al* (eds.) Berlin: De Gruyter.
Pietroski, Paul (2000) *Causing Actions* Oxford: Oxford University Press.

Glossary

Dualism: In the philosophy of mind, dualism is the doctrine that mental and physical entities are distinct from one another. Substance dualism is the doctrine that minds and bodies are distinct substances; property dualism is the doctrine that mental and physical properties are distinct properties.

Epiphenomenalism: The doctrine that the mind has no effects in the physical world.

Externalism: In the philosophy of mind, externalism is the view that the nature of some of a thinker's states of mind is essentially determined by facts external to the thinker's body (hence the slogan: 'the mind isn't in the head').

Functionalism: In the philosophy of mind, the doctrine that mental states are distinguished from one another by what they are typically caused by and what they cause (their typical or characteristic 'causal role').

Occasionalism: The seventeenth century doctrine that the mind cannot act directly on the physical world, but only with the help of God on each occasion of apparent mental causation.

Overdetermination: Two causes overdetermine an effect when they both bring the effect about and either would have done so equally well in the absence of the other.

Physicalism: The doctrine that everything is physical, or is composed out of or determined by something physical.

Supervenience: A phenomenon X supervenes on a phenomenon Y when there can be no difference in X without a difference in Y. So, for example, a thinker's mental states supervene on their brain states when there can be no difference in the mental states without a difference in the brain states.