A Structure for Mental Causation

June 28, 2009

Abstract

This paper suggests a structure that makes room for a class of solutions to the mental causation problem.

1 The mental causation problem

The core of the mental causation problem, as I see it, is that there is more than one candidate for what makes the effect of a mental event happen. One of those candidates seems to have the best possible credentials for making the effect happen. The rest have less good credentials. The best candidate wins; hence only it makes the effect happen. And, finally, the best candidate is not mental. What makes your actions happen isn't your thoughts; rather it's your physical makeup. (I take it that calling this the "mental" causation problem is a bit of a misnomer, since the same problem comes up for anything outside of physics.)

What I'm going to do here is suggest a structure for getting out of the problem. I'll spend some time at the end discussing some problems for solutions based on the structure.

2 Background

I'm going to make some standard assumptions about the background for this problem.

- **Physicalism**: every actual concrete individual is made up of nothing but physical matter. "Physical" means: referred to by an ideal completed theory of what makes things happen in the natural world.

- **Closure**: physics is **closed** in the sense that for anything that happens, there is a purely physical cause, if it has a cause at all.
- **Causal relevance**: when one event causes another, the cause has properties in virtue of which it causes the effect. Events are "coarse-grained", that is, they are individuals that have many properties. Given an effect, some of the cause's properties matter to the existence and character of the effect, others do not. Call this relation on properties **causal relevance**.
- **Supervenience**: mental properties supervene on physical properties.

3 Property identity

A fairly popular strategy for dealing with the mental causation recently is to say that mental properties are identical to physical properties. Kim (2005) uses the mental causation problem to argue against non-reductivism and for identifying mental events with physical events. Heil (2003) holds that there are properties (or dispositions) and their existence and manifestation constitute truthmakers for propositions about various kinds of things, among them the mental ones. More recently[1] Heil suggests that this kind of view fits nicely with Davidson's anomalous monism (1970): it's not about *properties*, it's about which *descriptions* go with which laws.

There is at least a whiff of eliminativism about the property identity view. The view says there are no distinctively mental properties. It's not literally true of you that you have the property of thinking about Australia. One might then say: if nothing, really, has any mental properties, then there are no minds.

This is not a decisive consideration. According to Heil, it *is* literally true of you that you think about Australia. There are dispositions that make this claim true. They have physical descriptions as well.

Yet Heil's view is committed to a distinction between ways that the world makes sentences true: (a) its properties and objects are the referents of the predicates and referring terms of the sentences; (b) something else. Truth and existence look sort of second-class

[1] Presentation at the Southern Society for Philosophy and Psychology, Savannah, GA, April 2009.

on the second way: perhaps mental sentences are made true in some way similar to the way sentences about the average family are made true. There's still the whiff of eliminativism. I would prefer a robustly non-eliminative position, and so I think we need to say that mental properties are not identical to physical properties.

Non-reductivism and non-identity Non-identity is supposed to follow from irreducibility. The main arguments for irreducibility are the multiple realizability argument and Davidson's argument about the rationality of the propositional attitudes. Both arguments are controversial and troubled.

I favor a neoCartesian modal argument. Suppose that non-physical bearers of mental properties are conceivable; and suppose this is enough to show that they are logically possible. Now suppose that properties are individuated with respect to their logically possible bearers, so that if some possible individual a has property P_1 and lacks P_2, then P_1 is not identical to P_2. Then if some possible individual has mental properties and lacks any physical properties, then the mental properties are not identical with the physical properties.

This argument is consistent with physicalism as I describe it above. So a mental property can be non-physical, in this sense, without being in any way "spooky" and without any of the mental particulars in our world (or any world of which our physics is true) having any non-physical parts or accompaniments. It's also consistent with closure as I described it above.

4 The problem, again

Kim's is the standard formulation of the mental causation problem these days. Slightly modified, Kim's argument goes like this. Assume (for *reductio*) that a mental event causes some effect in virtue of having some mental property. By supervenience, the event and the effect have physical properties that are referred to by an explanation from completed ideal physics. By closure, the physical properties matter to the effect. Hence there are two candidates for which property instance causes the effect: the cause having its mental

property, or the cause having its physical property. The Exclusion Principle says that no more than one independent candidate can make the effect happen. The best candidate wins; so the mental event causes the effect only in virtue of having its physical property.

5 Independence and non-distinctness

In recent formulations, Kim's exclusion principle rules out "more than one sufficient cause" (Kim, 2005, 42); in his original formulation it rules out more than one "independent" cause (Kim, 1989, 239). As Bennett (2003) and others have pointed out, the assumption of supervenience entails that the mental cause and the the physical cause are not independent. The more recent formulation raises a further question: what if the mental cause isn't strictly distinct from the physical cause—what if, in some useful sense, they aren't strictly speaking more than one thing?

So here's my suggestion for a structure for mental causation: the mental is not identical to the physical, but it is not distinct from it either. (See Sanford (2005) for a nice discussion of the distinction between non-identity and distinctness.)

6 How to use the structure

There are lots of ways to implement this structure, that is, ways to set up our metaphysics of causation to satisfy the structure.

- Nomological covariation. If the physical properties P *just do* nomologically necessitate the mental properties M, then there's a sense at least in which the mental is not distinct from the physical.
- Constitution. Pereboom (2002, 2001) suggests that the causal powers of instances of mental properties are constituted by the causal powers of the instances of their realizers.
- Coincidence. Yablo (1992) argues that mental events coincide with physical events—they have all their "categorical" properties in common, but differ in their essential

properties. Causation, then, is governed by a "proportionality" constraint. So a mental event is suitably proportional to an action, while the physical event on which it supervenes is not.

- Intersection. Watkins (2002), Clapp (2001), and Shoemaker (1998) suggest that the causal powers of multiply-realized properties are the intersection of the causal powers of their realizers.[2]

- Union. Lewis (1983) suggests that properties *are* sets of actual and possible individuals. Then multiply-realized properties are just the union of their realizers. The set of all the red things properly includes the set of scarlet things, for instance. I don't think Lewis himself makes this suggestion; I do, in Dardis (2008).

And others.

Perhaps a general name for what is going on here is "overlap".[3] Let "generalities" name whatever it is that particulars have in common (for instance: properties, kinds, universals, types, dispositions, exactly similar tropes, powers, capacities, potentialities . . .), such that particulars make things happen in virtue of some of their generalities. Suppose generalities can *overlap*. Two items such that one overlaps the other are neither identical nor fully distinct. It is particularly straightforward to explain overlap if the items in question are set-like. Lewisian properties clearly can overlap, by way of set-inclusion. On the Watkins/Clapp/Shoemaker view, the causal power of a multiply realized property is a proper subset of the causal powers of each of its realizers, hence overlaps them. Constitution (the Pereboom/Kornblith view) looks like another promising way to explain overlap. It's less clear how to explain overlap on Yablo's coincidence strategy: coincident events are not identical, and they are not independent, but it's not clear whether we should say they are not distinct. The nomological covariation strategy has only non-independence, and so has even more work to do to explain what overlap might be.

What about the other kinds of generalities?

[2] I think Watkins is the originator of this idea; I learned of it from him in 1996 at John Heil's NEH "Metaphysics of Mind" summer seminar.

[3] (Harbecke, 2008, 165) defines "new compatibilism" about the mental causation problem as a family of solutions to the problem that deny the identity of mental things and physical things, and at the same time deny their distinctness. Harbecke's own view is a refinement of Yablo's.

The nature of properties and the rest is obscure. I want to suggest that the way to discover their nature is to work out a good philosophical theory of what they are. The test of a good theory is what it explains. One thing we *might* want of such a theory is that it provides a way to solve the mental causation problem. In this spirit, then, we can try out as an axiom about properties that properties can overlap. So, just as the set of red things properly includes the set of scarlet things, we can stipulate that the property of being red overlaps the property of being scarlet. We may want to say more than this. For instance, we might add that the property of being scarlet is *a part of* the property of being red. We would then take on the obligation to spell out what, if anything, this additional claim means.

7 Try it out on dispositions

In the rest of this paper I aim to try out the "overlap" idea on dispositions, specifically on Heil's "no-levels" account of dispositions. If the strategy works for dispositions, it can probably be made to work for the rest.

7.1 What are they?

Heil (2005) offers a framework for a certain conception of dispositions. They are actual intrinsic features of things. Subjunctive conditionals may tell us something about a disposition, but the disposition is something distinct from those sentences—its nature is not exhausted by some set of subjunctive conditionals. A disposition fully fixes what its manifestations can be like; in other words, the relation between a disposition and its manifestations is not contingent. In particular, it is not contingent on what laws of nature happen to hold.

According to Heil, dispositions are not grounded in something "lower level." Indeed, there are not *levels* of dispositions at all. If, by contrast, we thought that some dispositions are realized by other dispositions, the consequence would be that the higher level dispositions would not be causal powers. The argument is the mental causation argument (347-350). Heil sees a trilemma: either the effect is overdetermined by both the more and

the less fundamental property; or the lower level property is not by itself sufficient for the effect; or else one of the purported causes really isn't a cause. The first is implausible. The second violates closure. On the third alternative, since the more fundamental disposition is the better candidate for making the effect happen, the higher level disposition would be powerless. Since Heil assumes that dispositions are always "powerful", it then follows that there are no higher-level dispositions.

7.2 Using the overlap structure

The argument depends on the assumption that the more fundamental and the less fundamental properties are distinct (cf. the expression (349) "higher-level items"). Can we give this up?

Heil comments at the start of his piece (343) that "'Disposition' is a term of art: you can define dispositions as you please". So perhaps we have some liberty in setting up an account of dispositions that we can use in a "unified understanding of mind and world" (351).

As I argued above, the "no-levels" account as Heil articulates it has the whiff of eliminativism, and I count that as a strike against the account as a unified understanding of mind and world. What would happen if we say that dispositions can overlap?

The idea is to say that there are physical dispositions and also mental dispositions (and others, of course). The mental ones overlap the physical ones. It's so far left open what exactly this means. We don't have levels, since we don't have any dispositions (fully) distinct from the physical dispositions.

8 What "does causal work"?

So far, all we have is an answer to the exclusion argument: since the mental dispositions and the physical dispositions aren't distinct, they aren't independent causes, and so the exclusion argument doesn't apply to them. But we haven't begun to explain what it might mean to say that a given event is caused by (is the manifestation of) both a physical disposition and a mental disposition.

8.1 Overlap doesn't do the work

The general point is that the overlap structure by itself doesn't answer the question how mental (etc.) generalities (etc.) "do their work" or make things happen.

The exclusion argument was supposed to show that the supervening candidate *could* not cause the effect, since there can't be more than one non-overdetermining independent cause. Ok, so we have two candidates, which are not independent, and not overdetermining. It clearly doesn't yet follow that they *do* both cause the effect. In fact, there appears to be a perfectly good reason to say that they *don't*: the physical cause is sufficient, if anything is, to produce the effect. The mental candidate would be "'pretend' or 'faux'" (Kim, 2005, 62), nothing new or different from the physical cause. This point doesn't have anything to do with the dispositional account; we could make the same argument in a view that says that causation is driven by properties and laws.

Just to drive the point home, consider a particular account of the laws of nature, the (Hume)/Mill/Ramsey/Lewis (MRL) view. According to Lewis's formulation of the MRL view, a law of nature is a sentence from a strongest simplest theory of everything (one of the "best systems" of the world). We can define the notion of "causal work" or "making something happen" in terms of the laws of nature. (This isn't, of course, Lewis's view.) An event causes another event in virtue of its having property P just in case a description of that effect follows logically from a best system together with a description of the event, and the description of the cause refers to property P. Clearly the only properties that contribute to "causal work" according to this definition are ones referred to by the laws of nature, which will be the laws of fundamental physics.[4]

8.2 What does "doing the causal work" mean?

8.2.1 Humean answers

The textbook account of Hume's answer to this question is, of course, "nothing at all." The only thing it could mean, according to Hume, is "necessary connexion", but, it turns

[4]Thus a broadly Humean picture of what "making happen" means can be just as "fundamentalist" as any other picture, and hence just as exposed to the mental causation problem.

out, those words don't actually correspond to any idea (or: they don't mean anything).

Humeans other than Hume answer the question with fancy versions of "constant conjunction". They can, as we just saw, be fundamentalists. They don't have to be. Consider again the MRL view of laws, properties, and "doing the causal work". There is no metaphysical obstacle to broadening our conception of which sentences yield nomic properties. Lewis (1983) demonstrated the need for "natural properties" by observing that there is no obstacle to building up a theory of the world around *any* of the abundant properties, and hence that the "best system" account radically underdetermines our concept of law. Once we have "natural" properties, the number of best systems becomes manageable. But this would seem to open the way to developing additional best systems around properties that are *nearly* natural. Without trying to explain what that might mean, let's suppose that psychological predicates pick out nearly natural properties. So now we can say that an event causes another because it has a physical property, and because it has a mental property, since both properties are "nomic", that is, related by "best systems" to properties of the effect.

8.2.2 Non-Humean answers

Humean accounts of causation and law are not so popular these days. We want, it is felt, something like that "necessary connexion"; we want to know how the cause *makes* the effect happen.

One way to respond to this demand is to show that the relation between causes and effects is necessary. There is a variety of proposals that make this demonstration: Sellars (1948) argued that properties are individuated by what they do, and hence what they do is essential to them; Shoemaker (1984) argues that causal powers necessitate their effects; Armstrong (1983) argues that a law expresses a necessitation relation between universals (although he also argues that whether this relation holds is itself contingent); Bird (2001) offers the "down-and-up" argument to show that salt's dissolving involves exactly the same physical mechanism as water's being able to dissolve, and hence it's not possible for *salt* not to dissolve in *water*. Ellis (2001) argues, as does Heil, that what dispositions do is essential to them.

8.2.3 Skepticism: necessity isn't "connexion"

Suppose that these arguments do establish that the connection between an instance of one of these generalities and an effect is a necessary one. Then Hume was wrong that there is nothing necessary here. But do we now know what the *connexion* is? I don't think we do. Take Bird's lovely demonstration. Salt is a crystalline molecule held together by ionic bonding. The molecules in the lattice will experience electrostatic attraction and repulsion. Hence Coulomb's law is true of salt. Similarly, Coulomb's law is true of water. Hence the law that requires salt to dissolve in water follows from what salt and water *are*. Hence dissolving is necessary. Now: why do atoms experience electrostatic attraction and repulsion? The answer takes us down a level, to electrons and their charge, and the geometry of the molecules. How does charge *work*, i.e., how does electrostatic attraction bring together a negative and a positive ion? Well, there's more to say about the structure of what charge does. But, I think, in the end, there is an end to what we can say. At that end, the answer to, "how does that *work*?" is, "it does".[5]

Notice also that a purely Humean account of causation and laws has the resources to hold that the connection between cause and effect is necessary. If properties are individuated with respect to the laws that govern their instances, then there is room for a position on which property instances necessitate their effects. Take Lewis's conception of properties as sets of actual and possible individuals, and take some law of nature, for instance Coulomb's law. If the properties referred to by Coulomb's Law are sets the members of which are exactly the actual and possible individuals that satisfy the law, then the law turns out to be necessary and the properties necessitate their effects. This explanation of why the connection is necessary leaves what I take to be the disturbing conclusion of Hume's critique of causation untouched: there is no *connexion* between the cause and effect, even though the effect necessarily happens, given the cause.

[5]Unless all of physics is grounded in some deeper necessity (as Einstein hoped that it all reduces to geometry).

8.2.4 Overlap, dispositions, and process accounts of causation

Put skepticism aside. Suppose there is some account of "making happen" that holds of generalities. I know of roughly three such accounts in the recent literature: (1) counterfactual accounts, like Lewis's, and more recently Woodward's; (2) mechanism accounts, like Glennan's; and (3) process accounts, like Dowe's. I myself find counterfactual accounts unpromising, since it seems as though the truth of the relevant counterfactuals has to depend on something else, like laws, or dispositions. Similarly mechanisms seem to me to be explained on the basis of laws or dispositions, and not the other way around. So for the remainder of this paper let me consider what it would be like to put a process account of causation together with a view that holds that dispositions may overlap.

The process account of causation hold that when one event causes another, there is a causal process that connects them; a causal process transmits a conserved quantity between the cause and the effect. Conserved quantities are defined in terms of physical properties, like mass-energy, momentum and charge (Dowe, 2008, Section 5).

Putting this together with the basic dispositional account, we get, one event causes another when the first has a disposition, and the manifestation of that disposition involves the transmission of a conserved quantity. So take a case of mental causation: the desire for the cheesecake causes Suzy to reach for the cheesecake. There's some collection of physical dispositions at work here; let's call the collection P. P's causation of the arm to move involves conserved physical quantities.

Now suppose that M supervenes on P, and overlaps it. If M can make E happen, then there is some quantity conserved when E's follow M's. Two constraints seem necessary:

- **Non-additivity**: the quantity conserved in the M/E transaction and the quantity conserved in the P/E transaction are not independent. In particular, they must not be "additive". If, say, the total energy of the physical basis for the desire is conserved as it moves the hand toward the cheesecake, *and* some quantity is conserved in the desire/action causal transaction, the quantities don't add up; the effect doesn't end up with *more* of anything than either quantity bestows;

- **Non-triviality**: the existence of this conserved quantity should be non-trivial. The conservation laws in mechanics are non-trivial in the sense that they are a structural element in the theory of mechanics (cf. John Norton's complaint, (Dowe, 2008, Section 6.2)).

About **Non-additivity**, since M and P overlap, they aren't distinct; so we could think of the two conserved quantities as stemming from something like two alternative book-keeping schemes *for the same transaction*. I don't think **Non-additivity** is automatically satisfied given overlap. There are two sciences here; we would have to explain on their basis why the two quantities are not independent. (By contrast, *without* overlap, such an explanation would be a great deal harder, since if the two properties are independent, they would appear to make independent contributions to causation.)

If the **Non-triviality** constraint is satisfied, then there will have to be a science of the domain in which M falls—in this case, psychology must be a science. And this science must describe psychology as involving causation and conserved quantities.[6]

Is there any chance whatsoever of psychology being like this? (Dowe, 2008, section 6.6) writes, "in any case, to suppose that the conserved quantity theory will deal with causation in other branches of science also requires commitment to a fairly thorough going reductionism, since clearly there is nothing in economics or psychology that could pass for a conservation law". Maybe Dowe is right about this. But what if he's wrong? Here are two thoughts.

(1) The norm for dispositions is that they exercise their powers only together with other dispositions. This is all the more true for non-fundamental dispositions. Suppose fragility is a disposition, and suppose a particular glass is fragile. There are lots of ways to get it to shatter, and lots of ways for it to shatter. And there are all kinds of ways to prevent it from shattering even if struck. A disposition like this thus exercises its powers in a complex web-like way. If there were a quantity conserved in mental causation, perhaps we should expect it to be distributed over such webs, and perhaps very difficult to discover.

[6] I suppose it could be a matter of fact that there is a conserved quantity in mental/physical transactions, but that we are epistemically closed off from it (as McGinn suggests about consciousness), but I don't see any reason to believe this.

(2) A lot of psychological causation is causation involving propositional attitudes, states of mind that are intentional, that is, about other things. It may be that some concept of information is key to understanding intentionality; and it may be that that concept of information is connected to conservation concepts. It may even be that some concept of information is key to understanding consciousness, as Dretske thinks, and so something similar might work for mental causation involving consciousness.

9 Conclusions

If mental properties are distinct from physical properties, the mental causation problem looks hopeless; if they are identical with them, the cost of mental causation is eliminativism. So, I've suggested, it's time to try out saying that mental properties are neither identical with nor distinct from physical properties: they overlap.

Overlap by itself does not yield causation or "making happen". More is needed. Our ordinary epistemic access to this extra bit is through science: where we find what Fodor calls "special sciences," sometimes (but probably not always) we are inclined to think that the special science properties are involved in making things happen.

A philosopher with Humean inclinations will not go a lot farther than that: "making happen" is a matter of some science telling us what happens. An anti-Humean philosopher can go farther. She might have a positive theory of what "making happen" amounts to. The theory will add to what the Humean philosopher says: science shows us where to look for these higher-level "making happen" relations, and then, when we have found them, we will also find the extra metaphysical bits that make up "making happen." Even if the anti-Humean philosopher doesn't have a positive theory of what "making happen" is (she might think of it as a primitive), she has a choice as to whether all of it is located at the fundamental level, or whether it can occur at more than one level. If a fundamental disposition and a higher-level disposition overlap, the anti-Humean philosopher can say that both the fundamental disposition and the non-fundamental disposition make effects happen.

I myself, in Dardis (2008), propose a neoHumean solution to the mental causation

problem: (1) properties are roughly Lewisian, and so overlap is clearly defined for them; (2) the "gold standard" for causal relevance is to be linked by strict laws; (3) mental properties can be linked by strict laws if those laws are qualified to compensate for ways in which instances of the same kind of mental event can be microphysically different; finally, (4) "causal work" is tied to the existence of a science.

I hope to have shown in this paper that the structure underlying this proposal is an interesting and useful one that may profitably be implemented in a wide variety of metaphysical settings.

References

Armstrong, D. M. (1983). *What is a Law of Nature?* Cambridge University Press, Cambridge.

Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs*, 37(3):471–497.

Bird, A. (2001). Necessarily, salt dissolves in water. *Analysis*, 61(4):267–74.

Clapp, L. (2001). Disjunctive properties: Multiple realizations. *The Journal of Philosophy*, 98(3):111–36.

Dardis, A. (2008). *Mental Causation: The Mind-Body Problem*. Columbia University Press, New York.

Davidson, D. (1970). Mental events. In *Essays on Actions and Events*, pages 240–260. Oxford University Press, New York.

Dowe, P. (Fall 2008). Causal processes. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Ellis, B. (2001). *Scientific Essentialism*. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge.

Harbecke, J. (2008). *Mental Causation: Investigating the Mind's Powers in a Natural World*. ontos verlag, Frankfurt.

Heil, J. (2003). *From an Ontological Point of View*. Oxford University Press, Oxford.

Heil, J. (2005). Dispositions. *Synthese*, 144:343–356.

Kim, J. (1989). Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives*, 3:77–108.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press, Princeton.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge University Press, Cambridge.

Pereboom, D. (2002). Robust nonreductive materialism. *The Journal of Philosophy*, 99(10):499–531.

Sanford, D. (2005). Distinctness and non-identity. *Analysis*, 65(4):269–74.

Sellars, W. (1948). Concepts as involving laws and inconceivable without them. *Philosophy of Science*, 15:287–315.

Shoemaker, S. (1984). Causality and properties. In *Identity, Cause and Mind: Philosophical Essays*, pages 206–233. Cambridge University Press, Cambridge.

Shoemaker, S. (1998). Realization and mental causation. *Proceedings of the Twentieth World Congress of Philosophy*, 9:23–33.

Watkins, M. (2002). *Rediscovering Colors: A Study in Pollyanna Realism*, volume 88 of *Philosophical Studies*. Kluwer Academic Publishers, Dordrecht.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, 101(2):245–280.