Can Psychophysical Identity Save Mental Causation?

The popularity of psychophysical property identity has waxed and waned since its presentation by the Australian materialists of the 1950s. It went out of favour after the criticisms of Putnam and Davidson, but now it is back in fashion. The reason, it seems, is not because new shortcomings in those criticisms have been pointed out. But rather, because it has been thought that psychophysical identity is useful for solving difficult problems as to how there could be any consciousness or any mental causation in a physical world. In my view psychophysical property identity is false, and furthermore it wouldn't help solve these problems. My aim in this paper is to show that it wouldn't help us understand how there can be mental causation in a physical world.

The challenge to those who believe there is a problem in understanding mental causation is to provide an articulation of that problem. And this will involve articulating a causal notion that gives rise to the problem. It will be my main concern in this paper to show how implausible it is that any such notion can be provided. But in the course of considering a proposal that might be thought to give rise to a mental causation problem, I will argue that psychophysical property identity would not help solve the problem. The same holds if one talks of reduction instead of identity or if, acknowledging the implausibility of identity or reduction in the light of multiple realisability concerns, one talks instead of local reduction or local identity.[1] It is a local, or species-specific, version of psychophysical property identity that was until recently the favoured solution of Jaegwon Kim, the person who has done most to press the urgency of the problem.[2] What we have then may be seen as a solution in search of a problem. The task I am attributing to the former Kim and to those with similar views,[3] is to come up with an analysis of causation which sustains a

problem of mental causation to which psychophysical identity (or one of the above variants) is a solution.

The most pressing problem of mental causation, as Jaegwon Kim sees it,[4] is causal exclusion--whenever one would intuitively suppose there to be a mental cause of some event there is always a physical or neural cause that excludes the mental cause. The reason for this is that a physical event provides a "full causal account"[5] of the effect, or is "sufficient"[6] for the effect. Hence, there cannot be any work left over for the mental event to do once the physical event has done its causal work.[7]

The exclusion of the mental by the physical may have some plausibility for notions like causal sufficiency, but it is far less compelling for notions readily characterized as causal relevance that are seen as naturally permissive. I shall not directly challenge Kim's exclusion argument here. Rather, my claim is that if there is an exclusion problem for some causal notion, then a more general problem arises for that notion, namely that it attributes a causal role to physical but not mental entities (unless the mental entities are identical to those physical entities), and its failure to attribute a causal role to mental entities is counterintuitive. The problem is more general because a causal role may be denied to mental entities for conceptual reasons other than exclusion considerations. The exclusion problem might appear not to be subsumed under the general problem when it is presented as applying to cases in which there is a mental cause as well as a physical cause but the mental cause gets excluded. For on this presentation, we start out with what pretheoretically seems to be a causal notion allowing both physical and mental causes. However, the end result of declaring the mental entity a noncause is a causal notion that attributes a causal role to physical but not mental entities. When intuition declares that mental entities should sometimes play that causal role, we have the general problem.

Thus to demonstrate that there is a mental causation problem one would need to produce an analysis of some causal notion that applies to physical but not mental entities, and for which a causal role is intuitively required for mental entities.

To begin our search for suitable analyses of causal notions, we need to consider whether we are discussing causation as a relation between properties (or types), or causation as a relation between particulars. I shall focus my discussion on particulars, specifically instantiations of properties by spatial regions at a time. This is a natural choice of particulars when looking for the closest fit with a physicist's worldview.[8] And to simplify discussion I'll focus on moments in time rather than temporally extended intervals.

Next, we need to consider what causal relations we are discussing. Throughout this paper I shall assume we are discussing asymmetric causal relations rather than such symmetric relations as sharing a common cause. I shall assume that for every coherent causal concept or notion there is a unique causal relation. And, as the ensuing discussion will illustrate, I take it that causal terms and locutions can be used to pick out more than one causal notion, and a given causal notion can be picked out by more than one locution. Let us focus on distinctions among the principal ways 'cause' is used. We have 'a is the cause of b', 'a is a cause of b', 'a caused b', and 'a is causally relevant to b'. ('Causal efficacy' is often used too.) The most permissive causal notions are usually picked out by 'a is causally relevant to b'. Assuming that there are degrees of causal relevance, one can make a distinction between the utterly irrelevant and the somewhat relevant that is both sharp, in the sense that there is no arbitrariness about where to draw the line between them, and objective, in the sense that it is independent of anyone's judgement. 'a is causally relevant to b' might be taken as picking out a notion of being somewhat relevant in this sharp and objective sense. Or it might be taken as picking out the vague and subjective

notion of having a significant degree of causal relevance. 'a is the cause of b' picks out the most selective causal notion—one that will invariably be seen as not purely objective but subject to pragmatic factors in selecting one among many causally relevant items of special importance. 'a is a cause of b' should be taken as equivalent to 'a is positively causally relevant to b', the opposite of 'a is a detractor of b', and accordingly has both objective and subjective readings. The locution 'a caused b' is the most slippery since it could be taken to mean 'a is the cause of b' or 'a is a cause of b', or anything between on the spectrum of significance.

Such causal locutions are explicitly two-place and commonly pick out two-place notions. But all of them may be uttered in contexts in which they are understood as elliptical for a higher-place notion. Indeed sometimes causal locutions are explicitly higher-place, invoking relativity to such further items as interests, an enquirer, or contrast classes.[9]

What causal notion is picked out by a causal term and what causal term is selected in an enquiry also depend on the nature of the enquiry. We can expect different causal notions to be involved for such different purposes as interpreting moral precepts or legal statutes, defining functionalism, and serving in a causal analysis of notions like reference, perception, memory and action. When worried about whether there is any mental causation in a physical world, I think it is clear that the concern is not over whether there are mental causes given a suitable interest, enquirer, or contrast class. And the question whether there can be mental causation in a physical world is not thought to depend on anyone's judgement. The causal notions we seek as potentially giving rise to a problem of mental causation must therefore be objective two-place notions. Another desirable feature of a causal notion that generates the problem is that it be tied to the fundamental laws of nature. It is generally thought that things are made to

happen in virtue of fundamental laws governing the universe. So it will be helpful in setting up a problem of mental causation to do so by way of a causal notion that is tied to these fundamental laws. And in fact all of the proposals I consider here have this feature. But I shall not regard it as essential, since I do not wish to rule out the conceptual possibility that an agent, natural or supernatural, can make things happen by some act of the will independently of these fundamental laws.

Let me summarise the goals of this paper. They are first to identify analyses of objective two-place causal notions. (I take this goal to have an interest extending beyond concerns about mental causation.) Second, for each such analysis, to consider whether it attributes a causal role to mental entities. Third, for those analyses attributing a causal role to physical but not mental entities, to examine whether this is counterintuitive and hence gives rise to a mental causation problem. And fourth, in any cases in which a mental causation problem might be thought to arise, to consider whether it is at all plausible that the problem can be solved by invoking psychophysical property identity and thereby allowing the mental entities to inherit the causal role of the physical.

I propose to examine now whether there are any analyses of objective, two-place causal notions. The principal candidates for analysing causation are counterfactuals, rough laws, and strict laws. The most commonly discussed counterfactuals are those taking the form 'if a hadn't occurred b wouldn't have occurred'. Such counterfactuals are indeterminate when interpreted literally, and are subjective when the indeterminacy is removed in the standard way by specifying that the counterfactual circumstances are those obtaining in the closest possible world to the actual world. Rough laws are those containing a ceteris paribus or "in normal circumstances" clause. As it is hard to see how an objective

assessment can be made of the point at which a generalisation has so many exceptions that it ceases to count as a law, or of what constitute normal circumstances, any analysis given in terms of rough laws will be subjective.[10] By contrast, an analysis given in terms of strict laws, i.e. laws without any vague terms or ceteris paribus clauses, will be objective. I shall begin by examining the prospects for analyses in terms of strict laws, and then turn to objective counterfactuals.

The most promising candidates for strict laws are those relating complete intrinsic physical properties of regions of space, by which I mean those properties detailing all the fundamental physical parameters at every point in a region. (From now on whenever I refer to a property of a region of space I shall mean intrinsic property.) The fundamental laws of physics will not relate such complex properties, but strict physical laws relating such properties will be derivable from the fundamental laws together with some specification of the physical property that is to feature in the antecedent of the law. If the fundamental laws are deterministic these derived laws will be deterministic. To simplify the discussion I shall assume determinacy.

Suppose we have a small region of space R with complete physical property P1 at time t1 and complete physical property P2 at a slightly later time t2. Now in order for a derived law with P2 in its consequent to be strict, the physical property P1 in the antecedent must be reinforced with physical properties of a vast surrounding region of space. Assuming that radiation and particles travelling at the speed of light can influence R at t2, but that nothing can travel faster than the speed of light,[11] the region of space R# whose physical properties are relevant at t1 to the contents of R at t2 will consist of R together with the entire region lying within a distance of c(t2-t1) from R, where c is the speed of light. Sometimes this region R# is expressed as the time-slice of the

backwards light-cone having R as vertex. In order for the law to be strict the antecedent must give a complete physical property P1# of R#, for if even a single physical parameter is left unspecified at some point in the region, the range of possible values that parameter will have will make a difference to the complete physical property P2 at t2. We can now see that R#'s having P1# at t1 is sufficient for R's having P2 at t2, and that P1# cannot be replaced by any physical property that does not entail P1#, if sufficiency is to be preserved.

If we take the notion of sufficient cause in its literal sense of a cause the occurrence of which guarantees the occurrence of the effect, we see that any sufficient cause of an instantiation of a complete physical property would have to be such an instantiation of a complete physical property in a huge region of space. Such a notion of sufficient cause between these property instantiations does not correspond to any causal notion familiar from ordinary talk. Clearly it does not allow mental sufficient causes, but it is unlikely this would be thought counterintuitive, and hence it is unlikely that this notion generates a problem of mental causation. Furthermore, if anyone were to think that there ought to be such mental sufficient causes, this could not be achieved by invoking psychophysical property identity. For a mental property cannot be identified with a complete physical property occupying a huge region of space.

There are two ways this notion of sufficient cause needs modifying in order to match more closely any ordinary causal notion: we need to take a property of a smaller region in the antecedent of the law, and we need to take much less detailed properties. A natural way of accommodating the first point is to consider the notion of being a nonredundant part of a sufficient condition, or inus condition, to borrow Mackie's terminology. (Unlike Mackie, I'm not regarding an inus condition as relative to a causal field, i.e. to a set of background conditions.)[12] As we have just seen that nothing short of R#'s having

P1# at t1 is sufficient for R's having P2 at t2, it follows that the instantiation of a complete physical property P1* by any subregion R* (including R*=R) of R# at t1 will be an inus condition of R's having P2 at t2.

To accommodate the second suggested modification, let us now examine causal relations between instantiations of properties that may be macroscopic as well as complete physical. Consider first the case in which the effect but not the cause is the instantiation of a macroscopic property, B, such as being occupied by a pain, or some neural property. Assuming that B supervenes on P2, it follows a fortiori that R*'s having P1* at t1 is an inus condition of R's having B at t2. Next consider the case in which the cause but not the effect is the instantiation of a macroscopic property. For any R*, R*'s having a macroscopic property A at t1 could not be an inus condition of R's having P2 at t2. For replacing a complete physical property in R* at t1 with A (or even with an incomplete physical property) would no longer leave a property of R#'s at t1 sufficient to guarantee R's having P2 at t2, since P2 is maximally sensitive. So no instantiations of macroscopic properties at t1 can be inus conditions of instantiations of complete physical properties at t2.

Consider now the case in which both cause and effect are instantiations of macroscopic properties. When we take the instantiation of macroscopic property B at t2 as effect, it might seem that we are still left with nothing short of the complete physical property P1# of R# at t1 as sufficient cause. For if the physical parameters are left unspecified at even a single point in the property instantiated by R#, one can imagine an extreme filling in of one of those parameters preventing the instantiation of B occurring at t2. Think of a mass so great that gravity sucks all matter into it, preventing the instantiation of any macroscopic property such as B occurring at t2. So a specification of physical parameters is required for every point in R#. However, certain parameters could be specified

as falling within a range of values, rather than precisely, opening up a variety of instantiations of properties of R# (that supervene on P1#) at t1 that would qualify as sufficient causes of R's having B at t2. And this also opens up prospects for macroscopic properties appearing in both antecedent and consequent of strict laws, and hence of inus conditions holding between instantiations of macroscopic properties. For there will be some regions R* where the complete physical property can be replaced by some less detailed property, perhaps a macroscopic property, without giving up the sufficiency of the antecedent in determining R's having B at t2.

However, such an inus condition seems hopeless as an analysis of any common causal notion. That is because if R*'s having A at t1 is to be causally related to (and ex hypothesi, an inus condition of) R's having B at t2, it would be thought that it must also be causally related to (and ex hypothesi, an inus condition of) R's having complete physical property P2 at t2, since R has B in virtue of having P2. And we have just seen that the latter inus condition cannot hold.

I turn now to consider the prospects for objective counterfactual analyses of causal notions. The statement 'a is an inus condition of b' entails but is not entailed by the statement 'if a hadn't occurred b might not have occurred' (as I shall shortly illustrate), where this counterfactual is to be interpreted as requiring the world to run on in accordance with the fundamental laws after leaving the circumstances outside the region occupied by a unaltered while replacing a inside that region by any event that is not a. Such an interpretation differs from the standard interpretation of the counterfactual 'if a hadn't occurred b wouldn't have occurred', on which the counterfactual circumstances, inside and outside

the region taken together, are characterised as those obtaining in the possible world most similar to the actual world.

To illustrate, if R*'s having P1* at t1 is an inus condition of R's having P2 at t2, then R*'s having P1* at t1 is nonredundant in guaranteeing that R has P2 at t2. So if R* had not had P1* at t1, then R might not have had P2 at t2. However, this counterfactual does not entail the corresponding inus conditions statement. For we saw earlier that R*'s having A at t1 cannot be an inus condition of R's having P2 at t2. Yet the corresponding counterfactual clearly obtains. Holding fixed the circumstances obtaining outside R*, if R* hadn't had A at t1, then R might not have had P2 at t2.

Consider now the following objective counterfactual analysis of causal relevance: R*s having A at t1 is causally relevant to R's having B at t2 iff if R* had not had A at t1, then R might not have had B at t2. (From now on I will take 'A' and 'B' as labels for any intrinsic properties, including both complete physical and macroscopic properties.) This analysis maintains that R*'s having A at t1 could make a difference to whether or not R has B at t2, and thus expresses a minimal notion of causal relevance. But it has the upshot that the instantiation of any property in any region in a light cone is causally relevant to the instantiation of a property by a region at the vertex of that light cone. For no matter what property B is instantiated at the vertex of the light cone at t2, and no matter what property A is instantiated in some region R* in the light cone at t1, an extreme alternative filling in of R* at t1, e.g. the presence of a black hole, could prevent R's having B at t2. This upshot will seem less counterintuitive when one reflects that a hypothetical enquirer who expected R* to contain a black hole at t1 would find it explanatorily relevant in understanding why R had B at t2, to learn that contrary to expectations, R* had in fact the innocuous property A at t1. I therefore think this counterfactual has some claim to providing an analysis of

causal relevance. It cannot, however, be used to set up a problem of mental causation as it applies to instantiations of all properties, including mental properties.

In conclusion, when puzzlement arises as to whether there can be any mental causation in a physical world, that puzzlement concerns an objective two-place notion of causation. Whether or not the puzzlement is motivated by considerations of exclusion, in order for there to be a problem there must be some causal notion for which there are physical causes but no mental causes. In the course of this investigation we have encountered two objective two-place causal notions. But only one of them, a notion of causal sufficiency, holds of physical causes but not mental causes. As the failure of this causal notion to apply to mental causes is not particularly counterintuitive, it is implausible to think a problem of mental causation is thereby raised. But should anyone think a problem does arise, the problem cannot be solved as it is often thought to be by invoking psychophysical identity. If a mental property is plausibly to be identified with a physical property it will be with a functional or neural property rather than a complete physical property. So if psychophysical property identity is to help save mental causation, there must be a notion of causation that gives rise to a problem of mental causation and that takes instantiations of neural or functional properties as causes. We have not encountered a causal notion that does this.

There are many other candidates for objective causal notions that I have not had space to consider here.[13] But from this investigation I think we can at least say that the burden of proof lies with Kim and anyone who shares his worry to present a causal notion which generates the problem.

[1] It should also be acknowledged that under local reduction or identity the mental properties that would inherit causal relevance would no longer be purely mental properties.

[2] See, e.g. *Philosophy of Mind* (Boulder: Westview, 1996) pp 152 and 234.

[3] For example Tim Crane "The Mental Causation Debate" *Proceedings of the Aristotelian Society Supplementary Volume* LXIX 1995.

[4] *Mind in a Physical World* (Cambridge: MIT Press, 1998), hereafter: MPW, p 32.

[5] MPW p 65.

[6] MPW p 43.

[7] MPW p 37.

[8] The preferred choice of Bennett in his discussion of property instantiations in *Events and their Names* (Hackett 1988) is instantiation of properties by spatiotemporal regions. I believe results establishable for instantaneous property instantiations can be straightforwardly adapted to temporally extended property instantiations.

[9] Some may argue that there is really only one central causal notion and it is a three-place notion, with the relativity capturing all the variety there appears to be in causal notions. However, even if a single three-place notion is central, one may always derive two-place notions from such a three-place notion. One way of getting a subjective two-place causal notion is by taking a three-place notion, dropping the explicit relativity to enquirer, interest, or contrast class, and replacing it by an implicit relativity to an appropriate or normal enquirer, interest, or contrast class. The subjectivity enters in judging what is normal or appropriate. And one can always derive an objective two-place notion from an objective three-place notion by filling in the third place, e.g. replacing the three-place ' x caused y from the standpoint of enquirer z' with 'x caused y from the standpoint of an omniscient enquirer'.

[10] In *Philosophy of Mind* pp 139-144, Kim argues that these counterfactuals are ultimately analysable in terms of rough laws, and rough laws cannot be assumed to be genuine causal laws rather than spurious laws. Spurious laws like those relating movements of barometer needles to rainfall will display exceptions under the possible circumstances in which instruments malfunction. In MPW pp 50 and 71, Kim offers slightly different reasons for thinking that counterfactuals and rough laws are inadequate as indicators of genuine causation.

[11] This assumption is challenged on some interpretations of the Einstein-Podolsky-Rosen experiment. But it would not affect my argument significantly if we were to reject the assumption; we would simply have to replace time-slices of the light cone by time-slices of the whole universe.

[12] See his *Cement of the Universe* (Oxford: Clarendon Press, 1974).

[13] Elsewhere I discuss some ways of refining the inus condition considered here and some corresponding changes in the objective counterfactual. I also look at probabilistic analyses, and how analyses are affected by relaxing the assumption of determinacy. And I look at causal notions relating different kinds of particulars. (In "Singular Causal Statements and Strict Deterministic Laws" *Pacific Philosophical Quarterly* 1987, I examine causal relations between concrete particulars.) I argue that it is implausible that there are any notions of causation

that give rise to a problem of mental causation and that take instantiations of neural or functional properties as causes.