Mental Causation
Proceedings of the Third International Colloquium in Philosophy of Mind
André Leclerc, Giovanni de Queiroz and Michael B. Wrigley (eds.)

CDD: 128.2

SUPERVENIENCE AND THE PROBLEM OF DOWNWARD CAUSATION

WILSON MENDONÇA

Rua Vitório da Costa 84A/301
22261-060 RIO DE JANEIRO, RJ
BRAZIL

mendonca@ifcs.ufrj.br

Abstract: *It seems that higher-level, nonbasic properties can only manifest their causal powers by exerting causal influence on lower-level, physically basic phenomena in the first place. A very influential line of reasoning conceives of this form of downward causation as either reducible to causation by physical properties or as ultimately untenable, because incompatible with the causal closure of physical reality. The paper argues that this is not so. It examines, first, why it is that a recent attempt by Noordhof to substantiate the notion of supervenient causation in a nonreductive framework fails. The upshot of this examination is the claim that any attempted specification of the most basic causal factors which supposedly underlie a causal transaction cannot account for the counterfactually necessary connections with the effect in question. By contrast, the specification of these factors at a higher level would allow establishing such connections. The paper closes with a discussion of how this view of autonomous causation at the higher-level can coexist with the notion of a complete specification of the causes of any physical effect exclusively in physical terms.*

Key-words: *physicalism; mental causation; causal closure; causal relevance of properties.*

1. SUPERVENIENCE

For a some time, supervenience seemed to be exactly what philosophers of mind needed in order to reconcile materialism with the

view of mental properties as dependent on, but also essentially distinct from physical properties. To be sure, philosophers soon realized that the notion of supervenience could be spelled out in many different ways and that there were, of course, different reasons for favoring one approach over the others. Nevertheless this controversy over the specification of supervenience did not alter the conviction that a suitably defined notion of supervenience would make it possible for philosophers of mind to preserve their materialism while holding on to the "autonomy of psychology" as an irreducible account of the causal relations between mental properties. Accordingly, Jerry Fodor once wrote that "if mind/brain supervenience goes, the intelligibility of mental causation goes with it" (Fodor (1987), p. 42).

At least since the appearance of "The Myth of Nonreductive Materialism" (Kim (1989)) and other seminal papers by Jaegwon Kim (e.g. Kim (1993a) and (1993b)), this consensus has been shattered. Kim's argument challenges the very possibility of mental causation, i.e. the objective relation that supervenience was supposed to secure. It seeks to show that given the causal efficacy of physical properties, which virtually everyone accepts, there is no causal role left for supervenient properties to play. According to Kim, if mental properties merely supervene on physical properties, then it is unlikely that philosophers will be able to find a proper place for them in the causally structured world.

A remarkable feature of this argument is that it is widely indifferent as to how one should conceive the notion of supervenience. The only aspect of supervenience that is relevant for running the argument is the necessary determination of supervenient properties by the properties which constitute their supervenience-base. Kim maintains first that supervenientists, as we may call them, are committed to the idea of downward causation, i.e. causation of lower-level phenomena by supervenient, higher-level phenomena. Moreover, they are committed to downward causation as a relation irreducible to lower-level causation. Kim suggests further that no account of downward causation is possible

that does not flout the materialist assumption of a causally closed physical world. In an important sense, the conclusion is that maintaining supervenience entails doing away with the intelligibility of mental causation.

The present paper takes issue with this view. My first main thesis is that Kim's argument is not as conclusive as it seems. The crucial difficulties concerning downward causation, so I will argue further, find solution in a metaphysical framework that draws on the notion of supervenience and accepts the causal closure of the physical world, upon providing an independently justified interpretation of the latter.

2. DOWNWARD CAUSATION

Consider the case, where an instantiation of a supervenient property *M* causes the instantiation of another supervenient property *M**. An illustration of this would be a mental state causing another mental state. As a materialist, the supervenientist assumes that the appearance of supervenient properties depends on the presence of appropriate basal conditions. The supervenientist is a physicalist in the sense that for her physical conditions ultimately determine the instantiation of all the properties there are. So we have for the higher-level, supervenient property *M** a lower-level, determining physical property *P**. The counterfactual implication of *M*'s claim to being a cause of *M** says that *M** could not have been instantiated, if *M* had not been present on this occasion. The determination relation between *P** and *M**, on the other hand, implies that unless *P** were present on this occasion, *M** could not have been instantiated. A plausibly coherent description of the situation seems to be: the instantiation of *M* causes the instantiation of *M* by causing* the instantiation of *P** in the first place; the later instantiation determines then the instantiation of *M**.

The first part of this description entails, of course, downward causation, a relation objectively connecting a higher-level phenomenon

(as the cause) to a lower-level phenomenon (as the effect). To this the supervenientist is committed. For, as Kim argues, if supervenient properties are really new and distinct from physical properties, then the causal powers associated with them must be irreducibly distinct from the causal powers of the properties defining the conditions out of which they emerge. This means that the causal role of *M* in the process by which *P** is brought about cannot be "preempted" by any physical property.

However, as we have good reasons to assume that the instantiation of *P** has as its cause an instantiation of another physical property – these reasons being derived from the assumption that the physical world is causally closed – the purported distinctness of supervenient causal powers results in the uncomfortable supposition that physical phenomena underlying supervenient phenomena are systematically overdetermined. They are overdetermined in the sense that they have two distinct causes, a physical cause and a supervenient one. What causes discomfort is the fact that the joint operation of two causes, each one being sufficient to bring about the effect, should manifest itself not occasionally, but *whenever there is causation by supervenient properties.*

Moreover, if we decide to apply to the relation between the supervenient property *M* and its supervenience-base *P* the same reasoning applied to the relation between *M** and *P**, we arrive at the conclusion that it is ultimately *in virtue of* some necessarily co-instantiated physical property *P* that the instantiation of *M* causes the instantiation of *P** (and, therefore, also the instantiation of *M**). As a result, causal facts involving supervenient properties could tentatively be seen as determined by causal facts on the more fundamental physical level. While this move could solve the problem of massive overdetermination of physical phenomena, it would also make higher-level causal relations ultimately dependent on, derivative from the causal processes at the physical level: all irreducible causal powers would turn out to be the ones associated to physical properties. Kim points out at this juncture that the

supervenientist, committed as she is to downward causation, could hardly accept this strategy for avoiding overdetermination. But the only other available alternative seems to be the abandonment of physical causal closure, which is not really open for the supervenientist, insofar as she stays committed to physicalism.

If we now use "property-causation" to refer to the relation by which the instantiation of a property X causes an event of type Y in virtue of being an instantiation of X (and not in virtue of being an instance of some other co-instantiated property Z), the main steps of the argument can be summarized as follows:

(i) M property-causes M^*. [higher-level causation]
(ii) The instantiation of P^* determines the instantiation of M^*. [supervenience]
(iii) M property-causes M^* by property-causing P^*. [downward causation]
(iv) The instantiation of P determines M. [supervenience]
(v) P property-causes P^*. [causal closure of the physical world]
(vi) The instantiation of P^* is simultaneously caused by the instantiation of M and the instantiation of P. [over-determination]

From this point the argument proceeds along well-known *reductio* lines. As massive overdetermination cannot be the rule, M's claim as the property in virtue of which the instantiation of P^* is caused cannot be maintained: the only efficacious property in the process by which P^* is instantiated seems to be P – unless we accept a breach of the principle of the causal closure of the physical world. But if the instantiation of the supervenient property M (*qua* instantiation of M) has no causal power to bring about an instantiation of the physical property P, it is hard to understand how it could exert any influence on higher-level phenomena as well: no higher-level causation without downward causation.

It is important to notice that the argument does not directly concern mental properties. It focuses instead on the relationship between higher-level properties in general and those properties defining their supervenience-base. As the latter can eventually supervene on more fundamental properties, we may assume that there is a last level made up of absolutely basic properties on which all other properties ultimately depend. Basic properties in this sense are conceived of as *physical* properties to be identified by the future development of fundamental physics. Basic properties are also assumed to be the only properties connected by genuine laws of nature, so that the principle of causal closure applies to the physical world as defined by those basic properties. Macrophysical properties, according to our current standards (the standards of current physics), and especially functional properties are of course nonbasic: they supervene, as we may suppose, on the properties discoverable by future physics. This implies that their claim as real causal factors is also challenged by Kim's argument. Specifically mental causation is, therefore, not the main target of the argument. The power of any supervenient property, whether mental or physical by the current standards, to exert causal influence on basic phenomena is what is at stake.

3. SUPERVENIENT CAUSATION

As remarked above, the supervenientist may seek to avoid the problem of overdetermination by making whole higher-level causal relations dependent on the causal processes at the basic level. This account, of which there are many variations, came to be known under the title of supervenient causation. An examination of the general reasons why the approach of supervenient causation fails will be useful in the identification of the conditions that must be fulfilled by a more satisfactory account.

The supervenient causation account says that A superveniently causes B if A supervenes on A' and B supervenes on B' and A' causes B'. This applies also to the "degenerate" case of downward causation, where B and its supervenience-base B' coincide, that is, where the causal chain starts with a higher-level phenomenon and ends in a basic phenomenon. But if downward causation can be seen as a case of supervenient causation, there can be no competition between it and properly physical, basic causation. The instantiation of P^* – to revert to the scenario of the last section – would be superveniently caused by M and simply caused by P. It is not impossible to see some overdetermination in this, but it seems to be harmless. Arguably, overdetermination is harmful only if two independent causal chains lead to the same effect. However, the move we are now considering implies that it is in virtue of causal processes on the basic level that causal relations on the higher-level are fixed. Causal facts involving basic properties determine, in other words, the facts about the causation of phenomena by supervenient properties, including the causation of basic phenomena by nonbasic properties. The chain of supervenient causation leading to P^* is not independent from the chain of basic causation. Rather, the idea is that the former depends, for its very existence, on the latter.

Kim points out that this notion of supervenient causation is irremediably at odds with the nonreductivist approach characteristic of supervenientism (e.g. Kim (1998), p. 232). For it follows from supervenient causation that the causal powers of M are exactly the same powers already possessed by its supervenience-base P. The property M superveniently causes only those things that are simply caused by P. There may be no harmful overdetermination in the process by which P^* is simply caused by P *and* superveniently caused by M. But there is also no autonomous contribution by M to the causal process in question beyond and above the contribution by P. The property M is in a sense not deprived of causal efficacy, although it is very difficult to distinguish

between the efficacy manifested in supervenient causation, completely derived as it is from basic, physical efficacy, and the null-efficacy of "epiphenomenal" events, that is, events that are effects of physical processes, but have no power to cause the instantiation of any other property (cf. Kim (1998), p. 151). Moreover, whatever causes *P* to be instantiated thereby also causes *M* to be instantiated. The upshot of all this is that there is no distinctive causal power, over and above physical powers, to be associated with the higher-level property *M*. But on what other grounds can the supervenientist still claim distinctness for *M*?

4. NOORDHOF ON PARTIAL ACTIVITY OF THE SUPERVENIENCE-BASE

In an attempt to block the suggestion that supervenient properties, whatever causal powers they happen to have, are deprived of just those autonomous powers that would give them a proper place in the "material world," Paul Noordhof (1999) has made a proposal based on the idea of partial supervenience. Noordhof's main question is this: How can nonbasic, supervenient properties be efficacious if causal circumstances constituted from instantiations of basic properties are causally sufficient for whatever effect is attributed to supervenient causes? He develops an argument to the effect that there is no reason why the instantiation of a (strongly) supervenient property should not be counted as having the efficacy that the causal relations on the level of its supervenience-base imply. We can recognize here the position of supervenient causation. Indeed, Noordhof believes that, if we allow for instantiations of supervenient properties to *exist* in virtue of being constituted from instantiations of basic properties, we should not deny that nonbasic causal relations exist by way of being constituted from basic causal relations (Noordhof (1999), p. 300).

According to Noordhof, higher-level properties in general, and mental properties in particular, are efficacious because (i) the instantiation of a part of one of their minimal supervenience-bases is a

cause of an event of type *E* and (ii) each minimal supervenience-base of a higher-level property is such that all its instantiations would cause events of type *E* in some causal circumstances *C*.

If we allow ourselves some simplifying assumptions, Noordhof's proposal can be presented as follows. Let *F* and *G* be nonbasic properties each with a certain number of minimal supervenience-bases. The minimal supervenience-base of the property *F* is a set of atomic physical properties such that

(i) It is metaphysically necessary that if all members of this set are coinstantiated, then *F* is instantiated; [strong supervenience]
(ii) If not all the members of this set are instantiated, then it is not metaphysically necessary that *F* is instantiated.

Typically, there will be more than one minimal supervenience-base of a property *F*. This is tantamount to saying that the property *F* is multiply realizable. Any member of such a base is a part of a minimal supervenience-base of *F*. The same applies of course to the property *G*.

For *F* to be causally efficacious in the process by which *G* is instantiated, two conditions must be satisfied. The first condition says that a minimal supervenience-base of the instantiation of *F* causes a minimal supervenience-base of *G* in a given circumstance. Note that this is a case of causation by instantiation of properties, not property-causation. The operative idea is to offer an analysis of the latter in terms of the former. The second condition states that "each minimal supervenience-base of *F* ... is such that all its instantiations would cause ... an instantiation of one of the minimal supervenience-bases of *G* ... , if they were in some causal circumstances *C* – where *C* may vary for each instantiation of *F*" (Noordhof (1999), p. 307).

For the instantiation of a minimal supervenience-base to cause anything, only a part of it, not all of it, may do the causal work. This is perhaps the most distinctive claim in Noordhof's attempt to flesh out

the approach of supervenient causation. Accordingly, a supervenient property is efficacious if its instantiation in the given causal circumstance is a cause. And the instantiation of the supervenient property is a cause if part of its actual minimal supervenience-base is involved in the process by which the effect is brought about (the first condition). The counterfactual content of the second condition is subjected to the same constraint. If the causal circumstance were different from the actual one, the instantiation of *F* would supervene on another minimal supervenience-base; and only a part of the latter–which does not necessarily coincide with the part playing the active role in the actual case – would be doing the causal job.

It should be noticed that Noordhof's general proposal, if it proves sound, can offer a very elegant solution for the problem of mental causation of behavior by intentional content, even if intentional properties, according to externalism, should not be individuated exclusively by reference to internal states. For, if intentional properties supervene at least partly on internal factors, this alone could make them efficacious. Two conditions must of course be satisfied for the instantiation of an intentional property to be the cause of an instantiation of a behavioral property. First, part of an instantiation of its minimal supervenience-base, namely, an instantiation of an internal physical property, must cause the behavior. Second, each minimal supervenience-base of the intentional property must be such that all its instantiations would cause the behavior in some causal circumstance *C*. This application to the particular case of intentional causation of the general account of supervenient causation developed by Noordhof would be in order *if* the general account is.

5. MINIMAL ACTIVITY?

But does Noordhof's approach succeed as an account of the possibility of higher-level causation in general? My answer to this

question is no. Again I assume that scrutiny of a wrong way can help us locate the right one.

Noordhof's account maintains commitment to the physicalist requirement that in a causal chain leading up to a given effect, causal factors constituted from instantiations of physically basic properties should be causally sufficient for the effect at stake. Sufficiency of physical causation, however, is not the only aspect to be secured by a physicalist account. The idea that genuine causal processes are ultimately fixed by causal connections among basic properties must also honor the requirement that causes are, in the actual circumstance, counterfactually necessary for their effects. Causal statements are not made true simply by the actual occurrence of the cause and the actual occurrence of the effect. Causal connections imply, therefore, certain counterfactual links. Thus in stating that *A* causes *B*, we assume that the effect would not have occurred if the cause had not occurred. The truth of causal statements involves, therefore, counterfactual conditionals. This means, in a physicalist framework, that causal factors constituted from instantiations of physically basic properties must be shown to be causally sufficient *and* counterfactually necessary for any effect that gets produced. As I will presently argue, this is the place where Noordhof's account founders. The closer to the supposedly basic causal factors it gets, the less able it is to formulate counterfactually necessary conditions for the effects in question.

The reasons for my verdict have to do with what I called the most distinctive claim in Noordhof's account: for a supervenient property to be efficacious, only part of its actual minimal supervenience-base must manifest efficacy in the process by which the effect is brought about. Suppose, for the sake of the argument, that in an attempt to find the ultimate conditions for the efficacy of a supervenient property *M*, we reach the actual minimal supervenience-base, i.e. the actual physical realization, in the given circumstance, of the property *M*. What we have at this level is "only" a set of basic properties such that every member of

The argument developed in the preceding section follows this common practice in philosophy of mind. It takes the instantiation of a property admitting of variation in its parameter to be a cause of an event of type *E*. The immediate consequence of showing that this cause is not physical (in the physicalist's sense) is the denial of the causal closure of the physical.

There is, however, an alternative way to this view, an alternative that could preserve the main point of the argument while still keeping to the causal closure of the physical. This alternative starts with the distinction between events and states, the case for which has been forcefully made by Steward (cf. specially Steward (1997), chapter 7). Accordingly, it is not wrong to insist that a particular cause-event must combine with an independent standing condition to give rise to effects. What is wrong, or at least misleading, is the idea that this is a case of *partial causes* combining in the production of an event-effect. Consider again the example of a match being lighted. For the striking of the match (the event-cause) to produce the desired effect a necessary condition must be satisfied – there must be enough oxygen around. It is therefore only in conjunction with an independent condition that the striking of the match exerts its causal powers in the desired way. In all nomologically possible worlds in which this condition is not satisfied the particular event-cause is not followed by the lightning of the match. It is misleading to conceive of what is lacking in these worlds as another partial cause, as this may suggest the absence of another particular beyond the causal factor referred to by "the striking of the match." Clearly, what has to be given in the actual situation for the particular event-cause in question to bring about the lightning of the match is not a particular entity, which can be represented by a singular term, but a fact, which has to be represented by a sentence.

Being a kind of fact, a standing state bears a relation to the effect it helps to produce which is very different from the relation connecting an event-cause (a particular) and an event-effect (another particular). In

Steward's proposed terminology, the first relation is "the relation of causal relevance." Its expression is a "sentential causal claim." The second is "the relation of causing," which is expressed by a "singular causal claim."

If we take this into account in the interpretation of the argument developed in the previous section, we arrive at a new result concerning the compatibility of autonomous causation at the higher-level and the claim of a complete specification of the causes of any physical effect exclusively in physical terms. The remarks in the previous section draw on the connection between the causal efficacy of properties defining standing conditions for causal processes, on the one hand, and the possibility of establishing counterfactually necessary connections between these properties and the effect in question, on the other hand. In other words, counterfactual relevance for the effect in question is used as a test of the existence of a "relation of causal relevance." Some supervenient properties like "presence of oxygen" pass the test, while the corresponding physically basic properties in their supervenience-bases do not pass the test. Thus some states can be causally related to physical effects without being themselves physical (in the physicalist's sense).

The existence of a relation of causal relevance connecting nonphysical states – which are ontologically kinds of facts, not particulars – with effects implies nothing at all about the possibility of describing the corresponding "relations of causing" exclusively in physical terms. From the point of view of causal relevance of facts or conditions, it is entirely open whether we able to designate the particulars involved in a causal process as event-cause and event-effect in physical terms alone. "Sentential causal claims" as expressions of relations of causal relevance cannot dictate the form of "singular causal claims" which express relations of causing between particular events. In particular, no assertion of a relation between an irreducibly nonphysical causal condition and an effect can show the futility of a purported translation of causal claims relating to particular events into the language of fundamental physics. For all we know, this translation may succeed.

We now have the means to formulate the principle of the causal closure of the physical world in such a way that it is not contradicted by the main argument of this paper. The principle says that, for any particular physical event whatsoever, the chain of events connected to it by the relation of causing contains only particulars which can be completely designated by physical terms alone. That these particulars have sometimes to combine with nonphysical facts or conditions to bring about effects does nothing to change their status as physical entities.

REFERENCES

CORBÍ, J.E. & PRADES, J.L. (2000). *Minds, Causes, and Mechanisms* (Oxford, Blackwell).

HEIL, J. & MELE, A. (eds.) (1993). *Mental Causation* (Oxford, Clarendon Press).

FODOR, J. (1987). *Psychosemantics* (Cambridge, Mass., The MIT Press).

KIM, J. (1989). "The Myth of Nonreductive Materialism." *Proceedings of the American Philosophical Association*, 63, pp. 31-47.

———. (1993a). "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?" In Heil and Mele (1993), pp. 19-26.

———. (1993b). "The Non-Reductivist's Troubles with Mental Causation." In Heil and Mele (1993), pp. 189-210.

———. (1998). *Philosophy of Mind* (Boulder, Westview Press).

MILL, J.S. (1873). *A System of Logic* (London, Longmans).

NOORDHOF, P. (1999). "Causation by Content?" *Mind and Language*, 14, pp. 291-320.

STEWARD, H. (1997). *The Ontology of Mind* (Oxford, Oxford University Press).