PHYSICALISM, OR SOMETHING NEAR ENOUGH

Jaegwon Kim

PRINCETON UNIVERSITY PRESS

PRINCETON AND OXFORD

Published by Princeton University Press,
41 William Street, Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
3 Market Place, Woodstock, Oxfordshire OX20 1SY

Third printing, and first paperback printing, 2008
Paperback ISBN: 978-0-691-13385-0

The Library of Congress has cataloged the cloth edition of this
book as follows

Kim, Jaegwon.
Physicalism, or something near enough / Jaegwon Kim.
p. cm.—(Princeton monographs in philosophy)
Includes bibliographical references and index.
ISBN 0-691-11375-0 (hardcover: alk. paper)
1. Philosophy of mind. 2. Mind and body. I. Title. II. Series.

BD418.3.K55 2005
128′.2—dc22
2004053451

British Library Cataloging-in-Publication Data is available

This book has been composed in Janson
Printed on acid-free paper. ∞
press.princeton.edu

Printed in the United States of America

3 5 7 9 10 8 6 4

2

The Supervenience Argument Motivated, Clarified, and Defended

An argument was presented in the preceding chapter to show that, on an influential position on the mind-body problem, mental properties turn out to be without causal efficacy. This is what I have called the supervenience argument, also called the exclusion argument in the literature. The argument has drawn comments, criticisms, and objections from a wide range of philosophers, but mostly from those who want to defend orthodox nonreductive physicalism and other forms of mind-body property dualism. Critics of the argument have raised some significant issues, both about the specifics of the argument and, more interestingly, about the broader philosophical issues involved. In this chapter, I would like to address two of the more pressing problems. One is that of "overdetermination," brought up by a number of philosophers; the second is the problem of "causal drainage," forcefully developed by Ned Block in his "Do Causal Powers Drain Away?"[1] Before we get to these and other issues, I want to set out the leading idea that motivates the supervenience argument and then offer what

1. Ned Block, "Do Causal Powers Drain Away?" *Philosophy and Phenomenological Research* 67 (2003): 133–150.

I hope will be a clearer statement of the argument, along with explanatory comments that some may find useful. But first we need a brief description of the philosophical position that is the target of the supervenience argument.

Nonreductive Physicalism

There is no consensus on exactly how nonreductive physicalism is to be formulated, for the simple reason that there is no consensus about either how physicalism is to be formulated or how we should understand reduction. For present purposes, however, no precise formulation is needed; a broad-brush characterization will be sufficient. Moreover, there need not be a single "correct" or "right" formulation of physicalism; there probably are a number of claims, not strictly equivalent, about the fundamentally physical character of the world, each of which can reasonably be considered a statement of physicalism. The strengths and weaknesses, merits and demerits, of these different physicalisms could be examined and debated, and reasonable people could come to different conclusions about them. In any case, most will agree that the following three doctrines are central to nonreductive physicalism: mind-body supervenience, the physical irreducibility of the mental, and the causal efficaciousness of the mental. Mind-body supervenience, the claim that makes the position a form of physicalism, can be stated as follows:

> *Supervenience.* Mental properties strongly supervene on physical/biological properties. That is, if any system *s* instantiates a mental property M at *t*, there necessarily exists a physical property P such that *s* instantiates P at *t*, and necessarily anything instantiating P at any time instantiates M at that time.[2]

2. There are alternative, not quite equivalent, ways of stating mind-body supervenience; one could get a good idea of what these might be from Brian McLaughlin, "Varieties of Supervenience," in *Supervenience: New Essays*, ed. Elias

I take supervenience as an ontological thesis involving the idea of dependence—a sense of dependence that justifies saying that a mental property is instantiated in a given organism at a time *because*, or *in virtue of* the fact that, one of its physical "base" properties is instantiated by the organism at that time. *Supervenience*, therefore, is not a mere claim of covariation between mental and physical properties; it includes a claim of existential dependence of the mental on the physical. I am assuming that a serious physicalist will accept this interpretation of supervenience. Mind-body supervenience as a bare claim about how mental and physical properties covary will be accepted by the double-aspect theorist, the neutral monist, the emergentist, and the epiphenomenalist; it can be accepted even by the substance dualist.

The second component of nonreductive physicalism reflects the "nonreductive" character of this form of physicalism:

> *Irreducibility.* Mental properties are not reducible to, and are not identical with, physical properties.

There is no single well-defined sense, or model, of reduction shared by all disputants in this debate, but this will not matter for us in the context of the supervenience argument; all we need to assume here is that physically irreducible properties remain outside the physical domain—that is, if anything is physically reduced, it must be identical with some physical item. The root meaning of reduction was given, I believe, by J.J.C. Smart when he said that sensations are nothing "over and above" brain processes.[3] If Xs are reduced to Ys, then Xs are nothing over and above the Ys.

Savellos and Ümit Yalçin (Cambridge: Cambridge University Press, 1995). In some contexts the interpretation of "necessarily" as it occurs in the last clause can be crucial; for our purposes, there is no need to opt for any special specification.

3. J.J.C. Smart, "Sensations and Brain Processes," in *The Nature of Mind*, ed. David M. Rosenthal (New York and Oxford: Oxford University Press, 1991), p. 170. Originally published in *Philosophical Review* 68 (1959): 141–56.

We now come to the third doctrine, concerning the causal status of these irreducible mental properties.

> *Causal efficacy.* Mental properties have causal efficacy—that is, their instantiations can, and do, cause other properties, both mental and physical, to be instantiated.

This last thesis is important to the many friends of the position I am describing. The irreducibility claim is often motivated by a desire to save mental properties as something special and distinctive, but if these properties turn out to be causally impotent and explanatorily useless, that would rob them of any real interest or significance, rendering the issue of their reducibility largely moot. Or one could argue that since physical properties are assumed to be causally efficacious, causally inert mental properties obviously cannot be physically reduced. This means that the rejection of mental causal efficacy would make the irreducibility claim true but trivial. In these ways, therefore, the doctrines of irreducibility and causal efficacy go hand in hand.

It can be debated whether these three doctrines constitute a robust enough physicalism. The issue obviously turns on the question whether mind-body supervenience as stated is sufficient for physicalism, since the irreducibility and mental causal efficacy have nothing specifically to do with physicalism; Descartes endorsed both. Moreover, classic emergentism, not usually considered a form of physicalism, endorsed all three, making it a target of the supervenience argument.[4] However, this issue will not affect the discussions to follow. My claims and arguments are intended to apply to any position that accepts the three propositions; what else it accepts makes no difference.

4. See my "Being Realistic about Emergence" in *The Emergence of Emergence*, ed. Paul Davies and Philip Clayton (Oxford: Oxford University Press, forthcoming). The three doctrines, however, can be thought of as capturing the physicalist core of emergentism. On supervenience and physicalism, see Jessica Wilson, "Supervenience-Based Formulations of Physicalism," forthcoming in *Noûs*.

The Fundamental Idea

The idea that drives the supervenience argument can be expressed in the following proposition, which I name after the great eighteenth-century American theologian-philosopher Jonathan Edwards:

> *Edwards's dictum*. There is a tension between "vertical" determination and "horizontal" causation. In fact, vertical determination excludes horizontal causation.

What do I mean by "vertical" determination? Consider an object, say this lump of bronze. At any given time it has a variety of intrinsic properties, like color, shape, texture, density, hardness, electrical conductivity, and so on. Most of us would accept the proposition that the bronze has these properties at this time in virtue of the fact that it has, at this time, a certain microstructure—that is, it is composed of molecules of certain kinds (copper and tin) in a certain specific structural configuration. I describe this situation by saying that the macroproperties of the bronze are vertically determined by its synchronous microstructure. The term "vertical" is meant to reflect the usual practice of picturing micro-macro levels in a vertical array, with the micro underpinning the macro. In contrast, we usually represent diachronic causal relations on a horizontal line, from past (left) to future (right)—"time's arrow" seems always to fly from left to right. From the causal point of view, the piece of bronze has the properties it has at t because it had the properties it had at $t - \Delta t$ (and certain boundary conditions obtained during this period). The past determines the future and the future depends on the past. That is what I mean by "horizontal" causation. So we have here two purported determinative relationships orthogonal to each other: vertical micro-macro mereological determination and horizontal past-to-future causal determination.

The lump of bronze has the color yellow at time t. Why is it yellow at t? There are two presumptive answers: (1) because its

surface has microstructural property M at t; (2) because it was yellow at $t - \Delta t$. To appreciate the force of the supervenience argument it is essential to see a prima facie tension between these two explanations. As long as the lump has microproperty M at t, it's going to be yellow at t, *no matter what happened before t*. Moreover, unless the lump has M, or another appropriate microproperty (with the right reflectance characteristic), at t, it cannot be yellow at t. Anything that happened before t seems irrelevant to the lump's being yellow at t; its having M at t is fully sufficient in itself to make it yellow at t.

So far as I know, Jonathan Edwards was the first philosopher who saw a tension of precisely this kind. Edwards' surprising doctrine that there are no temporally persisting objects was based on his belief that the existence of such objects is excluded by the fact that God is the sustaining cause of the created world at every instant of time. There are no persisting things because at every moment God creates, or recreates, the entire world *ex nihilo*—that is what it means to say that God is the sustaining cause of the world. Consider two successive "time slices" of the bronze: each slice is created by God, and there is no causal or other direct existential relationship between them. To illustrate his argument, Edwards offers a marvelously apt analogy:

> The *images* of things in a glass, as we keep our eye upon them, seem to remain precisely the same, with a continuing, perfect identity. But it is known to be otherwise. Philosophers well know that these images are constantly renewed, by the impression and reflection of *new* rays of light; so that the image impressed by the former rays is constantly vanishing, and a *new* image is impressed by *new* rays every moment, both on the glass and on the eye. . . . And the new images being put on *immediately* or *instantly* do not make them the same, any more than if it were done with the intermission of an *hour* or a *day*. The image that exists at this moment is not at all *derived* from the image which existed at the last preceding moment. As may

> be seen, because if the succession of new *rays* be intercepted, by something interposed between the object and the glass, the image immediately ceases; the *past existence* of the image has no influence to uphold it, so much as for a moment.[5]

Successive images are not causally related to each other; they are each caused by something else. If we suppose that the persistence of an object requires causal relations between its earlier and later stages, Edwards is arguing that "horizontal" causation involving created substances is excluded by their "vertical" dependence on God as a sustaining cause of the world at every instant. Remove God as the sustaining cause; the whole world will vanish at that very instant.[6]

It is simple to see how Edwards's dictum applies to the mind-body case, causing trouble for mental causation. Mind-body supervenience, or the idea that the mental is physically "realized"—in fact, any serious doctrine of mind-body dependence will do—plays the role of vertical determination or dependence, and mental causation, or any "higher-level" causation, is the horizontal causation at issue. The tension between vertical determination and horizontal causation, or the former's threat to preempt and void the latter, has been, at least for me, at the heart of the worries about mental causation.

5. Jonathan Edwards, *Doctrines of Original Sin Defended* (1758), Part IV, Chapter II. The quotation is from *Jonathan Edwards*, ed. C. H. Faust and T. H. Johnson (New York: American Book Co., 1935), p. 335. (Italics in the original.) It seems, however, that Edwards's argument may well have been foreshadowed by the occasionalists of the 17th century.

6. Some will argue that these considerations—and some of the crucial steps in the supervenience argument—depend on the use of a robust, "thick" concept of productive or generative causation rather than a "thin" concept based on the idea of counterfactual dependence or simple Humean "constant conjunctions," and that thin causation is all the causation that there is. See Barry Loewer's "Comments on Jaegwon Kim's *Mind in a Physical World*," *Philosophy and Phenomenological Research* 65 (2002): 655–62, and my reply to Loewer, ibid., 674–77.

The Supervenience Argument Refined and Clarified

Let us now turn to a restatement of the supervenience argument in a more explicit and streamlined form. It is useful to divide the argument into two stages; I believe each stage has its own interest, and this will also enable me to present two materially different ways of completing the second stage of the argument.

Stage 1

We begin with the supposition that there are cases of mental-to-mental causation. Let M and M* be mental properties:

(1) M causes M*.

Properties as such don't enter into causal relations; when we say "M causes M*," that is short for "An instance of M causes an instance of M*" or "An instantiation of M causes M* to instantiate on that occasion." Also for brevity we suppress reference to times. From *Supervenience*, we have:

(2) For some physical property P*; M* has P* as its supervenience base.

As earlier noted, (1) and (2) together give rise to a tension when we consider the question "Why is M* instantiated on this occasion? What is responsible for, and explains, the fact that M* occurs on this occasion?" For there are two seemingly exclusionary answers: (a) "Because M caused M* to instantiate on this occasion," and (b) "Because P*, a supervenience base of M*, is instantiated on this occasion." This of course is where Jonathan Edwards's insight, encapsulated in Edwards's dictum, comes into play: Given that P* is present on this occasion, M* would be there no matter what happened before; as M*'s supervenience base, the instantiation of P* at *t* in and of itself

necessitates M*'s occurrence at *t*. This would be true even if M*'s putative cause, M, had not occurred—*unless, that is, the occurrence of M had something to do with the occurrence of P* on this occasion*. This last observation points to a simple and natural way of dissipating the tension created by (a) and (b):

(3) M caused M* *by* causing its supervenience base P*.

This completes Stage 1. What the argument has shown at this point is that if *Supervenience* is assumed, mental-to-mental causation entails mental-to-physical causation—or, more generally, that "same-level" causation entails "downward" causation. Given *Supervenience*, it is not possible to have causation in the mental realm without causation that crosses into the physical realm. This result is of some significance; if we accept, as most do, some doctrine of macro-micro supervenience, we can no longer isolate causal relations within levels; any causal relation at level L (higher than the bottom level) entails a cross-level, L to L − 1, causal relation. In short, *level-bound causal autonomy is inconsistent with supervenience or dependence between the levels*. Further, an important part of the interest of the supervenience argument is that it shows that, under the physicalist assumptions we are working with, mind-to-mind causation is in trouble just as much as mind-to-body causation. Often the problem of mental causation is presented as that of explaining how the mental can inject causal influences into the causally closed physical domain, that is, the problem of explaining mental-to-physical causation. I wanted to do something more, namely to show that physicalism can put in peril all forms of mental causation, including mental-to-mental causation.[7] This is why the argument begins with line (1). It is at Stage 2 that we take up mental-to-physical causation. It is noteworthy that,

7. As we will see in the next chapter, an interesting parallel holds in the case of substance dualism: under substance dualism, mental-to-mental causation turns out to be as problematic as mental-to-physical causation.

unlike in the second stage below, the argument up to this point makes no explicit appeal to any special metaphysical principles; in particular, no specific assumptions about the physical domain, such as its causal closure or completeness, enter the picture at this stage.[8] Mental-physical supervenience is the only substantive premise that has been in play thus far.

Stage 2

There are two ways of completing the argument, and I believe the second, which is new, is of some interest. I will first present the original version in a somewhat clearer form:

COMPLETION 1

We now turn our attention to M, the supposed mental cause of M*. From *Supervenience*, it follows:

(4) M has a physical supervenience base, P.

There are strong reasons for thinking that P is a cause of P*. I will not rehearse the considerations in support of this idea; let us just note that P is (at least) nomologically sufficient for M, and the occurrence of M on this occasion depends on, and is determined by, the presence of P on this occasion. Since ex hypothesi M is a cause of P*, P would appear amply to qualify as a cause of P* as well. So we have:

(5) M causes P*, and P causes P*.

8. On some occasions I have tried to argue for (3) by invoking an exclusion principle—see, for example, the "principle of determinative/generative exclusion" in chapter 1. I think it preferable not to appeal to any general principle here; I now prefer to rely on the reader's seeing the tension I spoke of in connection with the two answers to the question "Why is M* instantiated on this occasion?" Anyone who understands Jonathan Edwards's argument and his mirror analogy will see it; I don't believe invoking any "principle" will help persuade anyone who is not with me here.

Note that P's causation of P* cannot be thought of as a causal chain with M as an intermediate causal link; one reason is that the P-to-M relation is not a causal relation. Note also that since M supervenes on P, M and P occur precisely at the same time. (Moreover, as we will shortly see, the two principles that will be introduced, *Exclusion* and *Closure*, together disqualify M as a cause of P*, making the idea of a causal chain from P to M to P* a nonstarter.)

To continue, from *Irreducibility*, we have:

(6) M ≠ P.[9]

Again, (5) and (6) present to us a situation with metaphysical tension. For P* is represented here as having two distinct causes, each sufficient for its occurrence. The situation is ripe for the application of the causal exclusion principle, which can be stated as follows:

> *Exclusion*. No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Let us assume that this is not a case of causal overdetermination (we will discuss the overdetermination issue below).

(7) P* is not causally overdetermined by M and P.

By *Exclusion*, therefore, we must eliminate either M or P as P*'s cause. Which one?

9. Note: this only means that this instance of M ≠ this instance of P. Does this mean that a Davidsonian "token identity" suffices here? The answer is no: the relevant sense in which an instance of M = an instance of P requires either property identity M = P or some form of reductive relationship between them. (See *Mind in a Physical World*, ch. 4). The fact that properties M and P must be implicated in the identity, or nonidentity, of M and P instances can be seen from the fact that "An M-instance causes a P-instance" must be understood with the proviso "in virtue of the former being an instance of M and the latter an instance of P."

(8) The putative mental cause, M, is excluded by the physical cause, P. That is, P, not M, is a cause of P*.

We can give relatively informal reasons for choosing P over M as the cause of P*, but for a general theoretical justification we may appeal to the causal closure of the physical domain:

> *Closure*. If a physical event has a cause that occurs at *t*, it has a physical cause that occurs at *t*.[10]

If we were to choose M over P as P*'s cause, *Closure* would kick in again, leading us to posit a physical cause of P*, call it P_1 (what could P_1 be if not P?), and this would again call for the application of *Exclusion*, forcing us to choose between M and P_1 (that is, P). Unless P is chosen and M excluded, we would be off to an unending repetition of the same choice situation; M must be excluded and P retained.

It is worthwhile to reflect on how *Exclusion* and *Closure* work together to yield the epiphenomenalist conclusion (8). *Exclusion* itself is neutral with respect to the mental-physical competition; it says either the mental cause or the physical cause must go, but doesn't favor either over the other. What makes the difference—what introduces an asymmetry into the situation—is *Closure*. It is the causal closure of the physical world that excludes the mental cause, enabling the physical cause to prevail. If the situation with causal closure were the reverse, so that it was the mental domain, not the physical domain, that was causally closed, the mental

10. For discussion of physical causal closure, or "completeness," see, e.g., David Papineau, *Thinking about Consciousness* (Oxford: Clarendon Press, 2002), ch. 1; E. J. Lowe, "Physical Causal Closure and the Invisibility of Mental Causation," in *Physicalism and Mental Causation*, ed. Sven Walter and Heinz-Dieter Heckmann (Exeter, UK: Imprint Academic, 2003). A simpler statement of causal closure in the form "If a physical event has a cause, it has a physical cause" will not do; given the transitivity of causation, the requirement would be met by a causal chain consisting of a physical effect caused by a mental cause which in turn is caused by a physical cause.

cause would have prevailed over its physical competitor. I suppose this could happen under some forms of Idealism; one would then worry about the "problem" of physical causation.

COMPLETION 2

Let us begin with the last line of Stage 1:

(3) M causes M* by causing its physical supervenience base P*.

From which it follows:

(4) M is a cause of P*.

By *Closure* it follows:

(5) P* has a physical cause—call it P—occurring at the time M occurs.

(6) M ≠ P (by *Irreducibility*).

(7) Hence, P* has two distinct causes, M and P, and this is not a case of causal overdetermination.

(8) Hence, by *Exclusion*, either M or P must go.

(9) By *Closure* and *Exclusion*, M must go; P stays.

This is simpler than Completion 1. *Supervenience* is not needed as a premise, and the claim that M's supervenience base P has a valid claim to be a cause of P* has been bypassed, making it unnecessary to devise an argument for it. However, Completion 1, in some ways, is more intuitive; it better captures Jonathan Edwards's fundamental insight and makes it particularly salient how putative higher-level causal relations give way to causal processes at a lower level. Either way, the main significance of Stage 2 lies in what it shows about the possible hazards involved in the idea of "downward" causation, namely that *the assumptions of causal exclusion and lower-level causal closure disallow downward causation*.

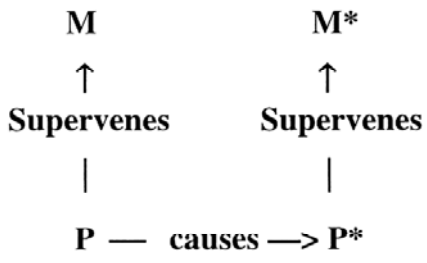

Figure 1.

Figure 1 pictures the outcome of the argument under Completion 1. In this picture, there is but one causal relation, from physical property P to another physical property P*, and the initially posited causal relation from M to M* has been eliminated. An apparent causal relation between the two mental properties is explained away by their respective supervenience on two physical properties that are connected by a genuine causal process. In this picture neither M nor M* is implicated in any causal relations; they play no role in shaping the causal structure—they only supervene on properties that constitute that structure. The supervenience relations together with the causal relation involved can generate counterfactual dependencies between the two mental properties, and between them and the physical properties; but these are no more causal than counterfactual dependencies involving any other supervenient property and its subvenient base (compare the aesthetic properties of a work of art and their base physical properties). Completion 2 presents a picture that is a bit less full: we no longer have the vertical "supervenience" arrow from P to M. M of course must have a physical supervenience base, but the argument, unlike in Completion 1, does not require it to be a cause of P*, although, as Completion 1 suggests, it may well be. The moral, however, is the same: the $M \rightarrow M^*$ and $M \rightarrow P^*$ causal relations have given way to an underlying physical causal process, $P \rightarrow P^*$.