ROBUST NONREDUCTIVE MATERIALISM

Derk Pereboom, University of Vermont

Journal of Philosophy XCIX, October 2002, pp. 499-531.

Nonreductive materialism about the mental has been put on the defensive by a series of well-developed arguments against its central claims. Four of these challenges, each of which has been advanced by Jaegwon Kim, are especially prominent: the argument from explanatory exclusion against irreducibly mental causal powers; the contention that the nonreductive view is indistinguishable from the emergentism of Samuel Alexander and C. Lloyd Morgan, a position widely held to be metaphysically extravagant; the claim that the functionalism typically endorsed by nonreductive materialists is incompatible with irreducibly mental causal powers; and the argument that if mental state types are multiply realizable, they cannot be genuinely scientific kinds, for then they will be only as weakly projectible as the wild disjunction of their possible realizations. This last challenge is representative of a growing skepticism about arguments against reductionism from multiple realizability.

I will first examine whether nonreductive materialism can finesse the explanatory exclusion problem. Subsequently I will argue that there are significant differences between the controversial sort of emergentism and nonreductive materialism, and that a nonreductive materialist need not be emergentist in this sense. I will then contend that a position according to which mental states instantiate irreducibly mental causal powers – the key feature of what I will call *robust nonreductive materialism* – indeed cannot be functionalist, but that there is a non-functionalist account of mental states to which the nonreductivist can turn. I will close by examining doubts that have been raised about arguments from multiple realizability against

reductionism, concluding that the nonreductive view can withstand these doubts.

* Thanks to David Christensen, Hilary Kornblith, and Mark Moyer for valuable comments and discussion.

I. AVOIDING EXPLANATORY EXCLUSION

Robust nonreductive materialism, as I conceive it, is a view about specifically psychological explanations, states, and causal powers, although it easily generalizes to other levels of explanation. In this view, an event such as Mary's buying ice cream (M2) will have a psychological explanation in terms of a complex of mental states -- beliefs and desires she has (M1). Each of M1 and M2 will be wholly constituted of microphysical events (P1 and P2 respectively), and there will be a microphysical explanation of P2 in terms of P1. The explanation of M2 by M1 will not reduce to the explanation of P2 by P1 (and, likewise, *mutatis mutandis* for events at various other levels of description). Underlying the irreducibility of this explanation is the fact that M1 is not type-identical with P1, and that M2 is not type-identical with P2. More fundamentally yet, the psychological explanation appeals to the irreducibly mental causal powers of M1 to account for M2, while the microphysical explanation appeals to microphysical causal powers of P1 to account for P2. Accordingly, the causal powers of M1 will not be type-identical with those of P1, and those of M2 will not be type-identical with those of P2. But neither will a corresponding token-identity thesis for these causal powers hold. For if it did, then the causal powers to which the psychological explanation refers would in the last analysis in fact be microphysical. Psychological explanations might then presume a

classification that clusters microphysical causal powers in a way distinct from how microphysics sorts them, but this would not compromise the microphysical status of those causal powers. Hence, robust nonreductive materialism affirms various token-diversity claims for mental causal powers. The token mental causal powers of M1 and M2 will not be identical with the token microphysical causal powers of P1 and P2, nor with the token neural causal powers of the neural states N1 and N2 that constitute M1 and M2, nor with token causal powers at any other level of description more basic than the neural.[1]

Furthermore, in my version of this robust nonreductive conception there will be a microphysical explanation for P2 that appeals to the microphysical causal powers of P1, and at the same time P2 (together with any requisite relational features) will be sufficient for M2. Consequently, there will be a microphysical causal explanation for M2 that appeals to the microphysical causal powers of P1 (given a constitutional account of M2 in terms of P2). For since one standard way of explaining an event causally is to cite a causal power whose activation is a sufficient cause of this event, citing the causal powers of P1 yields a causal explanation of M2. In another prominent nonreductive model, deriving from Hilary Putnam, there exists no genuine microphysical explanation for the action, only a psychological one.[2] I am strongly persuaded by the line of reasoning just presented against this view. Accordingly, I will set the alternative model aside in this discussion.

Familiarly, the position under scrutiny gives rise to a pressing question: what is the relationship between the microphysical and psychological explanations for M2? In particular, given that both sorts of explanation refer to causal powers, what is the relationship between the causal powers to which the microphysical explanation appeals and those to which the

psychological explanation appeals? Here is where Kim's challenge from causal or explanatory exclusion enters in.[3] As we have seen, if a microphysical account yields a causal explanation of the microphysical constitution of M2, then it will provide a causal explanation of M2 itself. What room is then left for a distinct psychological causal explanation of this action? Kim argues that it is implausible that the psychological explanation appeals to causal powers whose activation is sufficient for the event to occur, and at the same time the microphysical explanation appeals to distinct causal powers also sufficient for the event to occur, and that as a result the event is overdetermined. But it is also implausible that each of these distinct sets of causal powers yields a partial cause of the event, and that each by itself would be insufficient for the event to occur.

By the solution to this problem that Kim develops, real causal powers exist at the microphysical level, and so the microphysical explanations refer to real microphysical causal powers. Only if psychological explanations in some sense reduce to microphysical explanations does it turn out that the psychological explanations also appeal to real causal powers – these causal powers will then ultimately be microphysical. Psychological explanations that do not reduce to microphysical explanations will fail to refer to causal powers, and thus will have some lesser status – such explanations might express regularities without at the same time referring to causal powers. This move solves the exclusion problem because if the causal powers to which the psychological explanation appeals are identical with those to which the microphysical explanation appeals then there will be no genuine competition between explanations, and if the psychological explanations do not refer to causal powers at all, there will be no competition either. But this solution, which Kim believes is the only possible solution to the problem he

raises, would rule out any nonreductive view about mental causal powers.

In Kim's conception, any token causal powers of a higher-level property at a time will be identical with some token (micro)physical causal powers. He applies this view to mental properties as:

> [The Causal Inheritance Principle] If mental property M is realized in a system at t in virtue of physical realization base P, the causal powers of *this instance of M* are identical with the causal powers of P.[4]

Kim contends that rejecting this principle would be tantamount to accepting "causal powers that magically emerge at a higher level and of which there is no accounting in terms of lower-level causal powers and nomic connections."[5] By the causal inheritance principle, there would be no token causal powers distinct from token microphysical causal powers, which would preclude any robust nonreductive materialism. Higher-level kinds and explanations would at best group token microphysical causal powers in a way that does not correspond to the classifications of microphysics itself.[6] Such a classification might be of value for prediction. But there would remain no sense in which there exist causal powers that are not microphysical.

Let us first examine the causal inheritance principle, and then return to the explanatory exclusion issue. Is the causal inheritance principle true? And if it is false, is Kim right to suppose that magical emergentism follows? I think that the answer to both of these questions is "no." First, a respectable case can be made that higher-level token entities are typically not identical with their realization bases. The ship of Theseus is not identical with its current token microphysical realization base, for it would have been the same token ship had the token microphysical realization been slightly different, and it will be the same ship when this

microphysical realization if fact changes – the ship is in this sense *token multiply realizable*. True, there is some notion of sameness by which the ship is the same thing as its current microphysical realization – a notion that abstracts from any temporally extrinsic or modal properties.[7] But it is sufficient for absolute non-identity that A and B differ in their temporally extrinsic or modal properties. The same sort of argument can be run for token mental entities.[8] Is token mental state M identical with P, its actual token microphysical realization base? Suppose that M is realized by a complex neural state. It is possible for M to be realized differently only in that a few neural pathways are used that are token-distinct from those actually engaged. We need not rule at this point on whether the actual neural realization is token-identical with this alternative – it might well be. (Just as the Ship of Theseus would retain its identity supposing the replacement of a few of its planks, so it would seem that a token neural state would retain its identity given the replacement of a few of its neural pathways – more on this later.) But it is evident that this alternative neural realization is itself realized by a microphysical state P* that is token-distinct from P. It is therefore possible for M to be realized by a microphysical state not identical with P, and thus M is not identical with P. But furthermore, this reflection would also undermine a token-identity claim for mental causal powers – should they exist – and their underlying microphysical causal powers. For if the token microphysical realization of M had been different, its token microphysical causal powers would also have been different. We therefore have good reason to suppose that any token mental causal powers of M would not be identical with the token microphysical causal powers of its realization.

Still, there would be a sense in which the token causal powers of M would be "nothing over and above" the token causal powers of P – M's causal powers would nevertheless be

"absorbed" or "swallowed up" by" P's causal powers.[9] But there are two importantly distinct modes of this sort of absorption: identity and constitution without identity. And if there were essentially mental causal powers (that are physically realized), the relation of any one such token to its microphysical realization base would be the second and not the first.

On this picture, a token mental state would have the mental causal powers it does ultimately by virtue of the token microphysical states of which it is constituted (setting aside any fundamentally relational causal powers). For this reason it makes sense to say that token mental causal powers are wholly constituted by token microphysical causal powers. More generally, as Hilary Kornblith and I have contended,

> [Token causal power constitution] The causal powers of a token of kind F are constituted of the causal powers of a token of kind G just in case the token of kind F has the causal powers it does in virtue of its being constituted of a token of kind G.[10]

This nonreductive view endorses a weaker but nevertheless plausible version of the causal inheritance principle:

> [The Weaker Causal Inheritance Principle] If mental property M is realized in a system at t in virtue of physical realization base P, the causal powers of this instance of M are wholly constituted by the causal powers of P.

Moreover, on this view there will be a significant degree to which causal powers of higher-level tokens could be explained in terms of the causal powers of their microphysical constituents – limited by the extent to which the causal powers of the higher-level tokens are relational.[11] Furthermore, correlated with the possibility of this sort of constitutional explanation is the fact that the existence and nature of token higher-level causal powers would be predictable in

principle from their microphysical constituents together with the laws governing them. This predictability would again be limited by the extent to which the higher-level powers are relational, but, as I shall argue in the next section, nothing else functions to impede it.

Just as Kim claims that no competition between explanations arises in the case of reduction and identity, I propose that no competition arises in the case of mere constitution either.[12] For if the token of a higher-level causal power is currently wholly constituted by a complex of microphysical causal powers, there are two sets of causal powers at work that are constituted from precisely the same stuff (supposing that the most basic microphysical entities are constituted of themselves), and in this sense we might say that they *coincide constitutionally*.[13] The fact that they now coincide in this way might tempt one to suppose that these causal powers are token-identical, but, as we have just seen, there is a good argument that they are not. And because it is possible for there to be wholly constitutionally coinciding causal powers that are not even token-identical, it is possible that there be two causal explanations for one event that do not exclude each other and at the same time do not reduce to a single explanation.[14]

If identity and not just constitutional coincidence were necessary for explanatory non-competition, then there would be features required for non-competition that identity has and current constitutional coincidence lacks. The candidate features would arguably be constitutional coincidence at all other times, and constitutional coincidence at all other possible worlds, even now. But it is difficult to see how the token causal powers' constitutional non-coincidence at some past time, or at some future time, or their merely possible constitutional non-coincidence even now would introduce explanatory competition, while actual current constitutional

coincidence in absence of any features of this sort (i.e. identity) would guarantee non-competition. Suppose that my current token mental state M actually constitutionally coincides with token microphysical state P. Let us agree with Kim that if M were identical with P, and if their causal powers were identical, there would be no explanatory competition. If mere constitutional coincidence without identity resulted in explanatory competition, that would have to be because at some time in the past or in the future, or at some other possible world even now, M and P and their causal powers are constitutionally non-coincident. Suppose that M would still exist even if a few neural pathways in its neural realization were token-distinct from what they actually are. These neural changes would render M's microphysical realization base distinct from P, and thus M and P would be constitutionally non-coincident in some other possible world, and, similarly, *mutatis mutandis* for their causal powers. How could a possibility of this sort introduce explanatory competition? It would appear that actual current constitutional coincidence alone is relevant to securing non-competition, and thus for this purpose constitutional coincidence without identity would serve as well as identity.[15] Consequently, it would appear that available to the nonreductivist is a solution to the exclusion problem no less adequate than Kim's own.

II. DISTINGUISHING THE NONREDUCTIVE VIEW FROM EMERGENTISM

Kim contends that nonreductive materialism is committed to emergentism: "The fading away of reductionism and the enthronement of nonreductive materialism as the new orthodoxy simply amount to the resurgence of emergentism – not all of its sometimes quaint and quirky ideas but its core ontological and methodological doctrines."[16] In his analysis, emergentism claims a distinction between two sorts of higher-level properties, *resultant* and *emergent,* that

arise from the basal conditions of physical systems.[17] The basal conditions of a physical system comprise (i) the basic particles that constitute the physical system, (ii) all the intrinsic properties of these particles, and (iii) the relations that configure these particles into a structure. The higher-level properties that are merely resultant are simply and straightforwardly calculated and theoretically predictable from the facts about its basal conditions -- which presumably include the laws that govern the basal conditions -- while those that are emergent cannot be calculated and predicted. (Note that the variety of "predictability" at issue is not the possibility of discerning the future given current conditions, but rather the possibility of determining from an entity's realization-base what its concurrent higher-level properties are – it is in this sense synchronic and not diachronic predictability.) Theoretical predictability contrasts with inductive predictability. Having regularly witnessed that an emergent property is realized by particular basal conditions, we would be able to predict this relationship, but this sort of inductive predictability is not at issue. Rather emergentists maintain that knowledge of the basal conditions alone, no matter how complete, will not suffice to yield a prediction of an emergent property.[18] Emergentism of this sort (sometimes called *strong* emergentism) is often regarded as having next to no scientific credibility.[19] I have claimed, however, that on the nonreductive materialist view higher-level properties are in fact predictable from basal conditions (except insofar as higher-order properties are relational[20]). If this is correct, the nonreductive view is not committed to (this often-dismissed variety of) emergentism. We shall revisit the predictability issue shortly.

In Kim's analysis, a further characteristic of emergent properties is that they cannot be reductively explained in terms of the physical basal conditions. Kim rejects Ernest Nagel's "bridge law" conception of reductive explanation in favor of a model that first functionalizes the

higher-level property to be reduced, and then identifies that property with the realizer of that functionalization in the physical base. Here one might be interested in finding a particular realizer for a particular instance of the higher-level property, or one might be interested in finding the general realizer for some species or structure type. Finally, one finds a theory at the level of the physical base that explains how the realizers of the higher-level property can instantiate its functional specification.[21] In this conception, a property turns out to be emergent if it cannot be functionalized, or else if no realizer in the physical base can be found for specific instances of the property or for species or structure types -- no entity in the physical base for which there is a theory that can explain how it can realize that property's functional specification. One should note here that if there is to be a debate about whether non-reductivists might avoid emergentism, this irreducibility condition should at least initially be viewed as necessary but not sufficient for an emergent property.

Emergentism also endorses downward causation – it claims that higher-level states can have lower-level effects. (Kim raises a serious difficulty for the synchronic reflexive version of downward causation; here we will assume the diachronic variety.[22]) As applied to the topic at hand, emergentism asserts that a mental event can cause a microphysical event. According to Kim, the problem with this claim again derives from causal exclusionary considerations. Suppose mental event M1 causes microphysical event P2. M1 will be constituted of some microphysical event, P1, and M1 and P1 will compete as the cause of P2, and P1 will ultimately win out. Only by identifying M1 and P1 can M1's status as cause be salvaged.

To my mind, nonreductive materialism indeed countenances downward causation of this sort, and it can legitimately do so because one can reasonably hold that if M1 is wholly

constituted of P1, M1 and P1 will not compete as causes of P2. But its allowing for downward causation is not by itself sufficient to render the nonreductive position in any sense magical or radical. It is, for instance, wholly compatible with the theoretical predictability of higher-level properties from basal conditions. In my view, an emergentist's endorsement of downward causation would be radical if it specified that emergent properties could effect changes in the laws that govern the microphysical level independently of any emergent properties (henceforth *ordinary* microphysical laws). Supposing that M1 were such an emergent property, M1 could then cause P2 in such a way that P2 is no longer governed by the ordinary microphysical laws, but instead by laws that take into account the special characteristics of the emergent properties (or no laws at all). Elsewhere I have argued that given a materialist metaphysics, agent-causal libertarianism would be committed to the position that the agent-causal power is law-altering in this sense.[23] But nothing essential to nonreductive materialism entails this radical variety of downward causation.

As Randolph Clarke explains, this proposed feature of emergent properties would preclude their theoretical predictability.[24] For an the emergent property's capacity for altering the ordinary microphysical laws would not be predictable from a microphysical base given knowledge of only these ordinary laws. And supposing that the capacity for altering these laws is at least part of what provides the emergent property with its special nature, the property itself would then not be predictable from the microphysical base given the knowledge of only these ordinary laws. But as Clarke also points out, no feature of the nonreductive model *per se* renders higher-level properties any less theoretically predictable than they would be on a reductive model. True, according to the nonreductive view knowledge of the basal conditions of some

entity may not facilitate full predictability of its higher-level relational features, but this is also the case for reductionism. In each model, holding higher-level relational conditions fixed, a particular set of basal conditions will necessitate the same unique higher-level properties. An emergentist might also accept this, but what is crucial is that the nonreductivist is no more beholden to some factor that threatens to inhibit theoretical predictability, such as the power of higher-level properties to alter the ordinary microphysical laws, than is the reductionist. There is a difference in that the nonreductive view claims that there are higher-level causal powers that are multiply realizable, and knowledge of the higher-level features of causal powers alone would in such cases not allow the prediction of actual basal conditions. But this fact does not imperil the predictability of higher-level causal powers from basal conditions. Consequently, knowledge of the basal conditions can provide for exactly the same capacity for predicting higher-level properties in each model. And thus, what according to Kim is the defining feature of emergentism turns out not to be an essential characteristic of nonreductive materialism.

Kim further argues that if the physical base is nomologically sufficient for mental properties, then explanations in terms of mental states will be dispensible. In his example, emergent mental state M causes M*, M's base is physical state P and M*'s base is physical state P*. The point he is making works as well if M is irreducible rather than emergent.

> Now we are faced with P's threat to preempt M's status as a cause of P* (and hence of M*). For if causation is understood as nomological (law-based) sufficiency, P, as M's emergence base, is nomologically sufficient for it, and M, as P*'s cause, is nomologically sufficient for P*. Hence P is nomologically sufficient for P* and hence qualifies as its cause. The same conclusion follows if causation is understood in terms of

> counterfactuals – roughly, as a condition without which the effect would not have occurred... This appears to make the emergent property M otiose and dispensable as a cause of P*; it seems that we can explain the occurrence of P* simply in terms of P, without invoking M at all.[25]

On my nonreductive proposal it is indeed true that P is nomologically sufficient for P*, and that P* might well be asymmetrically counterfactually dependent on P, and that given these analyses of causation, P causes P*. In fact, it seems reasonable to demand that any analysis of causation should count P as causing P*. It also seems right to agree with Kim that we can therefore explain the occurrence of P* simply in terms of P, without invoking M as a cause at all. Indeed, we can also explain the occurrence of M* simply in terms of P, without invoking M as a cause at all. But does any of this make M dispensable as a cause of P*, and can we do away with an explanation of P* in terms of M as a cause?

I don't believe so. First of all, I have argued that token mental causal powers would not be identical with token microphysical causal powers. Suppose that M instantiates an irreducibly mental causal power, and that the activation of this irreducibly mental causal power brings about P*. Accordingly, M will cause P*, and M's causing P* will not be not identical with P's causing P*, and thus there is a significant sense in which M is indispensable as a cause of P*. Given a conception according to which explanations track and express the causal relations in which causal powers participate, the explanation of P* in terms of M will be *bona fide* and indispensible. If one were to claim instead that genuine explanations account for phenomena only in terms of the most fundamental conditions that cause them, then explanations in terms of mental causal powers might well be dispensable, while only microphysical explanations would

survive. But this is not a notion of explanation that we are constrained to accept.

Contrary to Kim's claims, the nonreductive materialist is not forced to affirm higher-level causal powers of the (strongly) emergent type. In short, nonreductive materialism need not agree that mental states have the power to produce deviations from the ordinary microphysical laws, and hence it can avoid this potential hindrance to theoretical predictability of mental states from basal conditions. And no other impediment to this sort of predictability that is not also a feature of reductionism appears on the horizon.

III. A NONFUNCTIONALIST ACCOUNT

By way of protest against Kim's species- or structure-specific reductionism, Ned Block once asked: "What is common to the pains of dogs and people (and all other species) in virtue of which they are pains?"[26] And as Kornblith and I argued,

> ...even if there is a single type of physical state that normally realizes pain in each type of organism, or in each structure type, this does not show that pain, *as a type of mental state,* is reducible to physical states. Reduction, in the present debate, must be understood as reduction of types, since the primary object of reductive strategies are explanations and theories, and explanations and theories quantify over types...[27]

In reply to this objection, Kim points out that nonreductive materialists typically argue from a functionalist perspective, and that functionalists characterize mental states solely in terms of purely relational or extrinsic features of those states. Indeed, functionalism identifies mental state types with type-level dispositions to cause mental states and behavioral outputs given perceptual inputs and mental states -- with the understanding that these dispositions are purely relational, that they are to be analyzed in terms of causal relations to perceptual inputs,

behavioral outputs, and other mental states, and no intrinsic mental components. Functionalists claim that what all pains would have in common, by virtue of which they are all pains, is a pattern of such relations described by some functional specification – call it H. Kim then argues that in providing an answer to Block's question, the local reductionist – the one who opts for species- or structure-specific reductionism – is no worse off than the functionalist. Both are committed to the claim that there is no non-relational or intrinsic property of pain that all pains have in common, and both can specify only shared relational properties:

> The local reductionist must grant that on his view there is nothing intrinsic that all pains have in common in virtue of which they are all pains (assuming that N_h v N_r v N_m [i.e. various neural realizations of pain] have nothing in common). But that is also precisely the consequence of the functionalist view. That, one might say, is the whole point of functionalism: the functionalist, especially one who believes in MR [multiple realizability], would not, and should not, look for something common to all pains over and above H (the heart of functionalism, one might say, is the belief that mental states have no "intrinsic essence").[28]

Kim implies that a functional specification does not provide a genuinely satisfactory answer to Block's question.

The reason he gives is this. On the nonreductive view, if E is a mental property and B is its neural or microphysical base, then realizers for E can be found in B (at the level of B). This position can also allow that nondisjunctive actual realizing properties might be found in B for individual species- or structure-types, as long as there is no property in B that is not wildly disjunctive that realizes E generally, i.e. that realizes every possible instance of E. The

nonreductive materialist claims that none of this entails a genuine reduction of E to properties in B. The standard strategy for preserving E as unreduced is indeed to retain E as a functional mental property. But Kim advances an objection to this functionalist move as a way of preserving nonreductive materialism about the mental:

> How should we counter this line of argument? I think it will be helpful to consider the causal picture, and ask: What are the *causal powers* of *this instance of E,* namely [a system] s's having E on this occasion? If s has E in virtue of E's realizer Q, it is difficult to see how we could avoid saying this: the causal powers of this instance of E are exactly the causal powers of this instance of Q.[29]

Here Kim cites what we might call the Causal Inheritance Principle for functional properties:

> (CIP - FP) If a functional property E is instantiated on a given occasion in virtue of one of its realizers, Q, being instantiated, then the causal powers of this instance of E are identical with the causal powers of this instance of Q.[30]

In Kim's view, the problem with the functionalist picture is that the causal powers of any instance of E will be causal powers in the physical base – they will not, at the token level, be irreducibly mental causal powers. Hence functionalism cannot preserve the view that there exist causal powers that are in the last analysis irreducibly mental, and it is thus incompatible with a robust nonreductive materialism about the mental. Furthermore, Kim points out that given the genuine multiple realizability of the property E, the causal powers of the realizers of E in B will exhibit significant causal and nomological diversity, and for this reason the causal powers of E will exhibit such diversity. Thus, in his estimation, E will be "unfit to figure in laws, and is thereby disqualified as a useful scientific property." He concludes that the functionalist model

cannot protect E as a property with a role in scientific laws and explanations.[31]

I am sympathetic to this general line of argument against a functionalist nonreductive view of higher-level causal powers.[32] However, I have argued that there is available a nonfunctionalist account of these higher-level powers that nevertheless remains nonreductive.[33] First of all, a line of thought with anti-functionalist implications can be found in the deepest criticism of behaviorism that Putnam advances in his classic essay, "Brains and Behavior."[34] There he argues that we should characterize mental states in a way that conforms to our characterization of kinds in the natural sciences. In the case of *polio*, for instance, we have found a biological explanation for the dispositional features of this disease (e.g. its symptoms), and we identify the disease with the underlying biological properties that provide this explanation. By contrast, behaviorists identify mental states with dispositions to behave in certain ways given particular stimuli, and not with underlying mental properties that explain these dispositions. Putnam recommends that we abandon behaviorism in favor of a conception that would characterize mental states in accord with the biological example.

Soon thereafter, however, Putnam developed and endorsed functionalism, and indeed he came to expect that functional properties would yield explanations for the dispositional features of mental states, much in the way that a viral infection provides an explanation for the symptoms of polio:

> My own view is that psychological predicates correspond to *functional* properties of human beings and other sentient beings. The presence of these properties *explains* the clustering of what some have called the 'symptoms' and 'criteria' of the various psychological states and conditions.[35]

This claim, however, amounts to the view that dispositions of types of mental states can be explained by dispositions of the same or other mental state types. As I have argued, this would amount to what is at best a weak sort of explanation, and not an explanation in terms of mental causal powers that is adequate by Putnam's own scientific realist standards.[36] What might it be for a type of dispositional property at a particular level to provide an explanation for another type of dispositional property at that level? To oversimplify, *pain* might be functionally characterized as follows:

> the state caused by pinpricks and pinches, that causes the thought "I should avoid those stimuli from now on," and given the belief that it's O.K. to express one's pain, causes winces and utterances of "ouch."

Now consider the proposal to explain a dispositional feature of *pain*, such as the tendency to cause winces, by means of the essential features of *pain*. Given functionalism, one would then be explaining *pain*'s tending to cause winces by a set of dispositions, and indeed, by a set that includes the very disposition to be explained as one if its components. Such an "explanation" is at best weak, for its crucial feature is a redundancy that undermines its strength.[37] In this respect it differs markedly from explanations that are adequate by scientific realist standards.[38]

However, the model inherited from other sciences, to which Putnam appeals in "Brains and Behavior," is best interpreted as explaining dispositional properties of kinds not simply by dispositional properties, but rather in part by properties intrinsic to those kinds – indeed, intrinsic properties proper to the same level as the kinds themselves. (A property is intrinsic to a kind just in case it is an intrinsic property of every possible instance of that kind.) In chemistry, the dispositional features of compounds are explained in part by their intrinsic molecular structure –

chemical properties intrinsic to those kinds of compounds. In biology, polio symptoms, such as its contagiousness – a dispositional feature — are explained partly by an intrinsic biological property of that kind of disease, a particular viral infection. By analogy, the nonreductivist might consider the possibility that there are properties intrinsic to mental state types that play a part in explaining their dispositional features.

Richard Boyd develops and defends a theory of kinds inspired by Putnam's claims. He notes that on various anti-realist views, natural kinds are characterized in terms of observable features and dispositional properties. Boyd then argues that a position of this sort fails to account for successful inductions based on natural kinds -- it does not allow for an explanation for the high degree of projectibility of these kinds. The remedy is to characterize natural kinds in terms of underlying causal powers that serve to explain their observable features and dispositional properties, and thereby to build into the nature of these kinds grounds for the success of inductions that appeal to them. Boyd has argued that in mature sciences, natural kinds are in fact typically characterized by such *explanatory essences*:

> Kinds characterized by "explanatory essences" are also kinds from the point of view of inductive generalization: indeed, in mature sciences, kinds which are explicitly characterized in terms of explanatory essences are the overwhelmingly typical cases of inductively natural kinds. Kinds natural from the point of view of successful induction need not always be explanatorily natural kinds, but they must correspond in relevant respects to the (perhaps unobservable) properties and mechanisms which causally determine the observable properties of the subjects of empirical generalizations.[39]

Given the models we have for kinds in the natural sciences, it is reasonable to suppose that these

essences would include properties intrinsic to the kinds.[40]

Now I suspect that most functionalists maintain that the causal powers that have a role in explaining the dispositional features of mental states are nondispositional properties of their realization bases. For example, many suppose that nondispositional neural properties, which instantiate neural causal powers, would serve to explain why being pinched causes wincing behavior. But if these causal powers are all non-mental, a robust sort of nonreductive materialist account of the mental is precluded, for then none of the causal powers would be essentially mental themselves.

By contrast, the nonreductivist might endorse intrinsic mental properties that instantiate specifically mental causal powers.[41] Such a view would be incompatible with functionalism. It need not deny that there exist functional mental properties, or, more generally, relational properties of mental states, but it would endorse nonfunctional mental properties that, by virtue of the causal powers they instantiate, play an important part in explaining dispositional features of mental state types.[42] What would such mental properties be like? First of all, despite the prevalence of functionalism, it is quite natural to suppose that the phenomenal content of a sensation-type is intrinsic to it, and that even if belief-contents are partially extrinsically individuated, it remains natural to suppose that the content of a belief-type is at least partially intrinsic to it. Moreover, we readily assume that behavior is causally explainable by way of intrinsic features of mental states, and this is at least consistent with the claim that these mental states have intrinsic properties that instantiate mental causal powers.

But how might mental states of this description be multiply realizable? Consider the analogy of a ball piston engine, the most recent version of the rotary internal combustion engine,

which has a specific internal structural configuration.[43] Characteristic of this engine is its having parts with particular shapes and rigidities, and these parts must be arranged in a particular way. These features are manifestly not functional relations that such an engine stands in; rather, they constitute intrinsic characteristics of this type of engine. At the same time, these characteristics are multiply realizable. The parts of the engine can be made of material of different sorts -- as long as the material can yield, for example, the required shapes and rigidities. The ball piston engine, then, has a nonfunctionalist intrinsic structure that instantiates its causal powers, but it nevertheless admits of distinct realizations.[44]

By analogy, it might be that the heterogeneous physical realizations of the dog's and the human's belief *that there is danger nearby* exhibit a structure of a single type that is intrinsic to this kind of mental state, a structure that instantiates the causal powers of this belief. This structure may be more abstract than any specific sort of neural structure, given that it can be realized in distinct sorts of neural systems.[45] Perhaps this same structure can be realized in a silicon-based electronic system, and such a system could then also have the belief. Suppose one built a silicon system that replicated the capacities of and interconnections among neurons in a human brain as much as is physically possible, and then excited the system to mimic as closely as possible what happens when a human being has this belief about danger. Is it not a serious empirical possibility that this silicon state would realize the same belief, and have a structure that, conceived at a certain level of abstraction, is similar enough to the structure of the ordinary neural system for both to count as instantiations of the same structure-type? It would certainly seem far from likely that nothing of relevance would be alike in these systems other than relations to perceptual inputs, behavioral outputs, and other mental states. At very least, in this

case and more generally, it makes little sense to retreat to mere functional resemblance prior to investigating whether the relevant similarities extend to intrinsic properties.

To what degree does the position I have been developing differ from Kim's view? The common ground includes the notion that if mental states are causal powers then they cannot be functionally defined. But in addition, Kim allows for neural structure-specific reduction. His idea is that there may, for instance, be neural structures common to more than one species to which some class of a mental-state type can be reduced. I suggest that we might indeed identify a mental state-type with a structure, but a structure more abstract than any specific neural structure, and one that can potentially be realized by a silicon-based system. Kim envisions the reducing structure to be neural, or physical at a lower level yet. My proposal is that there are structure-types that cannot be classified as specifically neural, but which must rather be categorized as mental, and which would be intrinsic properties of mental states.

An important challenge to this proposal is that such structures may not exist, that in general, no significantly homogeneous structure-types correlate with what are intuitively the tokens of mental state types. For instance, it may be that any higher-level structures that instantiate the belief *that there is danger nearby* differ on the order of the way in which a cat and an ordinary mousetrap differ as instantiations of the kind *mouse-catcher*.[46] This would constitute a serious challenge to a robust nonreductive materialist conception, but the view would not yet be decisively undermined. For it may be that these structure-types, although they fail to correlate neatly with our ordinary mental state categories, are not specifically neural structure-types either. In that case, one might take advantage of what room there is for altering the ordinary system for classifying mental states, at least for the purpose of scientific psychology, and identify the

distinct structures with distinct mental state types. Scientific reclassifications relative to ordinary categories are, after all, not unusual. Still, it may also turn out that in general, the only significantly homogeneous structure-types to be found are essentially neural structure-types. In that case, it is hard to see how there could still be irreducibly mental causal powers, and in my view Kim would then be right. But that result does not appear very likely to me, given the thought that the structure of a state of a silicon system, conceived at some level of abstraction, could be similar enough to the structure of a neural realization of a mental state for both to count as features of the same kind.

It seems odd that realists about mental states have so firmly endorsed functionalism, a model for the nature of those states that is so closely tied to a general anti-realist point of view. Positivist anti-realists advocated an operationalist characterization of natural kinds, defining them in terms of the causes and effects of their instances. Logical behaviorism provides a good example of this practice. However, scientific realism rejects this conception, for on this view the instances of natural kinds enter into causal relations without those kinds being defined by those relations. Although functionalist characterizations of mental states are more sophisticated than those of its behaviorist progenitor, functionalism nevertheless fits squarely within this anti-realist tradition. My alternative proposal is not novel in spirit. It simply recommends for mental states what realists typically advocate for natural kinds generally.

IV. MULTIPLE REALIZABILITY AND PROJECTIBILITY

It was once commonly supposed (by me, among others) that nonreductive views about the special sciences are grounded most fundamentally in the phenomenon of multiple realizability by way of a formal sort of argument. Kinds in the special sciences can be realized in different ways

from the perspective of lower level sciences, and thus an attempt to reduce higher-level kinds, laws, and explanations to those at a lower-level will involve replacement by disjunctive properties -- properties that are perhaps even wildly disjunctive in the sense that the disjuncts have at best little in common. Moreover, the disjunctions that these properties feature might even be open-ended or infinite. The received wisdom was that such disjunctive properties are not kinds, for the reason that statements of regularities involving such disjunctive properties fail to be laws, and perhaps most fundamentally, because "explanations" involving such disjunctive properties are not genuine explanations. This standard argument for nonreductive materialism appears to rely on a certain formal prescription for laws and explanations, that they cannot contain disjunctive properties, or at least not wildly disjunctive properties.[47]

Kim argues, however, that a higher-level property is precisely as projectible as the disjunction that expresses its multiply realizable character at a more basic level, and thus a generalization involving such disjunctive properties is just as lawlike as the higher-level generalization that it was meant to reduce.[48] The reason is that a higher-level property is nomically equivalent to such a disjunctive property. Nomic equivalence might be defined in this way: Properties F and G are nomically equivalent if they are coextensive in all possible worlds compatible with the laws of nature.[49] If Kim is right, then the formal argument does not appear to go through, for it relies on the possibility that generalizations involving a higher-level property be lawlike while those involving the corresponding disjunctive property are not. But furthermore, Kim contends that wildly disjunctive properties are not projectible, and hence higher-level properties that are nomically equivalent to such properties are not projectible either. As a result, such higher-level properties cannot figure into laws, and they are not genuinely

scientific kinds.

The example of a disjunctive property Kim adduces to make his point is *jade*. 'Jade' is a category that comprises two mineralogical kinds, *jadeite* and *nephrite*, and hence *jade* is the same property as *jadeite or nephrite.* As a result, *jade* will not be projectible. For suppose that we're trying to confirm the generalization 'jade is green'. We might check many instances of jade and find that they are all green. But it could be that the entire sample consists of jadeite, and no nephrite. We must conclude that the generalization is not confirmed, and thus *jade* is not projectible.[50]

To clarify his claim, Kim considers the objection that we can think of genuinely projectible kinds as disjunctive properties. *Emerald*, for example, can be thought of as *African emerald or non-African emerald*. But, he says, this possibility fails to undermine the projectibility of *emerald* – for example, it doesn't show that there is anything wrong with the lawlikeness of "All emeralds are green." However, this analogy does not serve to reinstate the projectibility of *jade*, for, by contrast with "jadeite or nephrite,"

> the disjunction, "being an African emerald or non-African emerald," does not denote some heterogeneously disjunctive, nonnomic kind; it denotes a perfectly well-behaved nomic kind, that of being an emerald! There is nothing wrong with disjunctive predicates as such; the trouble arises when the kinds denoted by the disjoined predicates are heterogeneous, "wildly disjunctive", so that instances falling under them do not show the kind of "similarity", or unity, that we expect of instances falling under a single kind.[51]

But given this analysis, even *jade* might turn out to be a kind after all. As Block points out, all samples of jade share certain appearance properties, similarities that give rise to a certain degree

of projectibility.[52] In Block's view, more generally, properties that are multiply realizable can yet be projectible with respect to "properties of channeled selection, learning, and design."[53] Because there are typically only a few ways in which entities of a particular higher-level type can be designed and produced, we can expect relatively broad similarities among these things that would render corresponding higher-level properties significantly projectible.[54]

The point I want to extract from this debate is that the heterogeneity of the possible realizations of a property is compatible with their having significant features in common, features that will undergird the projectibility of the property to some degree or other.[55] This point is consistent with Kim's claim that a higher-level property is precisely as projectible as the disjunctive property that comprises all of its possible realizations. One should not conclude from the heterogeneity of the possible realizations of a higher-level property that there is no feature that can sustain its projectibility – in fact, of both the higher-level property and of the disjunctive property that comprises all of its possible realizations. Indeed, the projectibility-sustaining feature of a kind could be a structure that is significantly homogeneous across its heterogeneous realizations, a structure that might instantiate a unitary causal power at the level of description of the kind. Note that disjunctive terms will typically fail to express or will at least mask any such homogeneous structural features and unitary causal powers to which they might correspond. In the case of the kind *ball piston engine*, for example, a disjunctive term that details its possible realizations would fail to express or would at least mask the characteristic structural features on which the projectibility of this kind is based. By contrast, the term 'ball piston engine' itself can serve to express these structural features without obscuring them.

But note that one cannot conclude merely from the fact that a property is projectible that

it is an intrinsic structural feature that instantiates a unitary causal power at the level of description of the property. Functional properties, for example, may be projectible while they comprise neither intrinsic features nor unitary causal powers. *Being soluble* is projectible, yet although for any instance of this property there will be a causal power that will explain its projectiblity, it does not itself instantiate a unitary causal power. The tie between projectibility and unitary causal powers is therefore looser than entailment. Whether a property is strongly or weakly projectible, it may turn out to be a functional property and have no intrinsic features that can instantiate a unitary causal power, let alone a unitary causal power at the level of description of the property.

Consider an instructive example of Kornblith's. In 1869 the term 'neuraesthenia' was introduced to designate a nervous disease that results in severe fatigue – a characterization that is at least fairly functional. The term was soon established worldwide, but "like most descriptive terms, where basic organic or psychological understanding was lacking, it tended to be overinclusive and a receptacle for many diverse conditions."[56] But when cures for neuraesthenia were sought, it was found that different sorts of causes had to be treated. Several distinct sorts of underlying causes were discovered for the dispositional features of this purported natural kind. As a result, the term 'neuraesthenia' became obsolete by about 1930.[57]

What makes us think that *neuraesthenia* fails to instantiate a unitary causal power at the level of description of this property? First, there is the evidence that it is not projectible to a high degree. In addition, researchers discovered a disjunction of properties coextensive with *neuraesthenia* each of which is more strongly projectible. Explanations involving these properties effectively replaced those involving *neuraesthenia*. Moreover, the characterization of

neuraesthenia was forced to remain fairly functional because no homogeneous underlying intrinsic features were discovered across its instances that could explain its dispositions and surface features. Accordingly, I would suggest that whether there is good evidence that mental states instantiate unitary and specifically mental causal powers depends on whether mental state types are projectible to a high degree, on the failure of a search for coextensive sets of properties that are more strongly projectible, and on whether intrinsic and specifically mental explanatory essences can be found. In short, whether there exists good evidence of this sort depends on whether there are powerful, resilient, and thoroughly realist psychological explanations in which mental state types play a part.

V. THE SIGNIFICANCE OF MULTIPLE REALIZABILITY

If the multiple realizability of a mental property fails to support the claim that its realization base is at best only weakly projectible while the mental property itself is strongly projectible, does there remain a role for multiple realizability in the argument for nonreductive materialism? Let us first consider whether multiple realizability of mental state types might still sustain the nonreductive view, whereupon we will turn to the multiple realizability of mental tokens.

(i) *Types.* Lawrence Shapiro is skeptical about any such significance for multiple realizability of mental state types:

> Take what appears to be a legitimate case of multiple realization... Either the realizing kinds truly differ in their causally relevant properties, or they do not. If they do not, then we do not have a legitimate case of multiple realizability... If the realizing kinds do genuinely differ in their causally relevant properties, then they are different kinds... and so

we do not have a case in which a single kind has multiple realizations.[58]

To illustrate the notion of a causally irrelevant property, Shapiro points to the color of a corkscrew. Corkscrews can be grey or black, for example, but the color of a corkscrew is causally irrelevant to its nature -- in this case, to what it does. He argues on this basis that differences in color among corkscrews do not amount to a legitimate case of multiple realization.

Shapiro extends the claim of causal irrelevance to an example involving neural and silicon realizations of a mind. "If each neuron's contribution to psychological capacities is solely its transmission of an electrical signal, and if silicon chips contribute to psychological capacities in precisely the same way, then the silicon brain and the neural brain are not distinct realizations of a mind."[59] Here, he thinks, the sameness in contribution to psychological capacities "screens off" the difference between neurons and silicon chips, and makes it the case that they are not distinct realizations of these capacities. Legitimate cases of multiple realization of E would have to feature realizations of E that differ in their causally relevant properties, and realizations differ in their causally relevant properties only when they make distinct contributions to the nature of E.

But first of all, would it be inconsistent to claim that the identical contributions to psychological capacities made by neural and silicon systems do amount to a case of multiple realizability while denying this of the grey and the black corkscrews? No. The color of a corkscrew is causally irrelevant in a starker sense than the one Shapiro has in mind, for the color makes *no positive causal contribution whatever* to its nature – in this case, to what it does, never mind different colors not making *distinct causal contributions* to what it does. Shapiro's characterization extends causal irrelevance to pairs of realizers, each of which in fact makes a causal contribution to the nature of the thing, whenever each makes the same causal contribution.

This characterization fails to count as different realizations pairs of distinct realizers each of which does in fact make a causal contribution to the nature of the thing that has it. So he says: "Steel and aluminum are *not* different realizations of the waiter's corkscrew because, relative to the properties that make them suitable for removing corks, they are identical."[60] But unlike color, being made of steel or of aluminum plausibly does make a causal contribution to what a corkscrew does – in this respect these properties are causally relevant in a way in which colors are not. Suppose that an effective corkscrew can only be made of steel and of aluminum, for only these materials have the right kind of rigidity, but that they make exactly the same contribution to what it does. The fact remains that these materials, as opposed to any others, have the right kind of rigidity. Accordingly, *making a causal contribution to the nature of the thing that has it* might be the notion of causal relevance that is pertinent to a condition on multiple realization. This alternative conception would license steel and aluminum but not distinct colors as multiple realizations of a corkscrew, and for silicon and neural systems to count as multiple realizations of psychological features.

Why adopt this alternative conception of multiple realizability? Perhaps it is enough to point out that distinct realizers can make the same positive causal contribution to the nature of a thing, and this is just what we mean when we talk about multiple realizability. But in addition, the fact that the neural system and the silicon system make identical contributions to psychological capacities seems to force us to say that the features thus contributed are neither essentially neural nor essentially silicon-structural. Here there is significant work for multiple realizability to do: because some one type of thing can have realizations of distinct types F and G, it can be characterized neither as essentially F nor as essentially G. Shapiro is in no position

to deny any of this. But given his conception, the realizations in this example will not really be multiple, and thus his conception fails to allow multiple realizability to do this work -- which it can in fact do. For this reason, his conception of multiple realizability is best rejected in favor of the proposed alternative.

What then is the legitimate role for multiple realizability in supporting nonreductive materialism? The answer is implicit in the above discussion. Whether or not a property is multiply realizable can indicate the level at which it should be classified. Is the kind *corkscrew* a kind of steel thing? No, for it also has a possible aluminum realization. Is the kind *mind* a neural kind of thing? If mental states are also realizable in silicon, then no. Multiple realizability might then provide the key to precluding classification of mental states as essentially neural, or as essentially classified at some lower level yet.

Note that the realizability of mental states in both neural and silicon systems would not all by itself establish a robust nonreductive materialism. In addition, we would at least need evidence that mental states have intrinsic properties that might then instantiate essentially mental causal powers. For only then could we show from the evidence that mental states are realizable in non-neural systems that there indeed exist mental causal powers that are not essentially neural.

Now I was surprised to discover several reductionist-leaning philosophers contending that physical types common to neural and silicon systems might well be discovered, but that this would lend *support* to reductionism. Oron Shagrir adopts this line of argument,[61] as do William Bechtel and Jennifer Mundale:

> A computer that could exemplify sufficiently similar behavior to biological organisms to justify the imputation of psychological states is likely to be very different from the ones

> we humans have designed to date, and the characterizations we would have to employ of its physical operations might turn out to be far more similar to those we use of brains than we currently expect. Such machines, for example, would likely have areas devoted to processing different sensory inputs and controlling motor outputs; potentially this might provide a basis for a common taxonomy of the physical processing states underlying psychological functioning.[62]

The advocate of a robust nonreductive materialism would welcome such a common taxonomy of these physical processing states, and so if the point that these advocates of "reductionism" are making is representative of their position, perhaps we can all agree! However, I would want to emphasize that this common taxonomy will not be essentially neural, and so neural reductionism would be precluded. True, this taxonomy will be physical, but no more physical than a taxonomy for irreducibly neural states, which the nonreductive materialist also welcomes. The term 'physical' has a narrow and a broad sense: in the narrow sense, it picks out specifically anything over which theories in the science of physics quantifies; in the broad sense it distinguishes anything wholly constituted from entities that are physical in the narrow sense. Nonreductive materialism is, after all, a kind of materialism, and hence does not countenance anything (in whatever realm it is claimed to hold true) that is not physical in the broad sense. If these reductionists are merely contending that for mental states shared by a brain and a computer we may find a common taxonomy that is physical in the broad sense, they are not proposing a hypothesis that is incompatible with nonreductive materialism.[63] Perhaps at this juncture these reductionists and some of their nonreductive opponents might find common ground.

Now Patricia and Paul Churchland have argued that the multiple realizability of

psychological states by type-distinct neural states and by both neural and non-neural states does not undermine reductionism, for the reason that reductionism might be "domain-specific":

> ...visual experience may count as one thing in a mammal, and a slightly different thing in an octopus, and a substantially different thing in some possible metal-and-semiconductor android. But they will all count as visual experiences because they share some set of abstract features at a higher level of description. That neurobiology should prove capable of explaining all psychological phenomena in humans is not threatened by the possibility that some *other* theory, say, semiconductor electronics, should serve to explain psychological phenomena in *robots*. The two reductions would not conflict. They would complement each other.[64]

If indeed visual experience in humans and in mammals had only functional characteristics in common, then the claim that they have distinct reductions would be plausible, for these distinct varieties of visual experience would not share causal powers. But if these varieties of visual experience share intrinsic structures, then they could share causal powers that are essentially neither neural nor electronic, but rather psychological or mental. Visual state-types could then be mental state types that conform to the robust nonreductive materialist conception. A key motivation for the Churchlands' remarks is their sense that it is neuroscientific research that will reveal the nature of human psychology. But there is a natural way of understanding this motivation so that it is consistent with the nonreductivist picture I have been developing. If there exist irreducibly mental intrinsic structures, then a very likely avenue for discovering them would indeed be research in neuroscience. For even if such structures could also be realized in silicon-based electronic devices, it is highly plausible that their neural manifestations would first be

discovered -- precisely by way of neuroscientific research.

Nevertheless, various reductionists have, in my estimation, provided appreciable reason to believe that realizability in different kinds of neural systems alone need not advance the cause of robust nonreductive materialism. Suppose that what would seem to be a single mental state type were realizable only neurally, albeit in neural systems that differ, such as a human's, a dog's, and an octopus's. Imagine first that a single structural property was found that is intrinsic to this mental state type. This property might well count as neural and not as irreducibly mental, for as Bechtel and Mundale point out, within the realm of the neural itself there are possibilities for classification at different levels of abstractness or coarse-grainedness, and the property might well be characterizable as neural at some sufficiently high level of abstractness.[65] Then we would not have multiple realizability at the neural level after all. In addition, as Shagrir argues, if distinct neural mechanisms were found, then it might be that they correspond to distinct types of mental states, although it may initially have appeared that a single type of mental state was at issue.[66]

(ii) *Tokens*. Even if types of mental causal powers are not identical with types of neural causal powers, it still could be that in the case of a normal human being, for example, every token mental power is identical with a neural causal power. (Earlier, I argued that token mental causal powers are not identical with token microphysical causal powers, but I reserved judgment on the token identity thesis for neural causal powers.) This is important, for if our mental causal powers were token-identical with neural causal powers, a robust nonreductive materialism would be precluded, for then in the last analysis there would in fact be no irreducibly mental causal powers. In order to foreclose this sort of token identity, token mental causal powers would have

to be multiply realizable in the right way.

The claim that token mental states and causal powers are identical with token neural states and causal powers is more resilient than is sometimes supposed. Indeed, stable tokens (given a certain level of complexity) often retain their identity over certain changes in their constitutions and configurations, and, significantly, they enjoy a certain resiliency in the production of their characteristic effects under these changes. So, for example, my decision to ring the doorbell can plausibly survive changes in its realizing microphysical state, and nature has likely endowed it with a resiliency for producing its characteristic effects under these small changes. But to establish that a token mental state M is multiply realizable at the microphysical level does not suffice to show that it is not identical with the token neural state N that constitutes it.[67] For N might be similarly multiply realizable at the microphysical level, and thus its being multiply realizable in this way is consistent with its being identical with M.

Whether our mental states are token-identical with neural states, even if they are not type-identical with neural states, is not a thoroughly obvious matter. But perhaps it can be shown that token mental state M, which is in fact realized by token neural state N, could have been realized by a token neural state that is type-distinct from N given the best neural classification. Given that N would not be identical to a token of such a distinct neural type, M would then not be identical N. Or maybe it can be established that M could have been realized by a token state that is not wholly neural but at least in part a silicon structure. Since N would not be identical to a realizing token that is not wholly neural, M would then also not be identical to N. This last scenario does not seem implausible. Suppose that at some time in the future we are capable of fitting brain-damaged patients with silicon-structural prostheses for the damaged parts of their brains. Now

consider my desire to ring the doorbell, a token mental state, which, let us suppose, is wholly realized by a neural token. Would it have been the same mental token had it been realized in a neural and silicon structure rather than in a purely neural structure, holding all else fixed as much as possible? To my mind, it could well have been the same mental token, while at the same time the actual neural realizing token would be distinct from the neural and silicon realizing token. If all of this is indeed plausible, then it might well be that the mental token is distinct from its neural realizing token, and, more generally, that the mental/neural token identity thesis is false.

VI. THE PROSPECTS FOR A ROBUST NONREDUCTIVE MATERIALISM

Kim and others have developed a number of very strong challenges to the nonreductive materialist position. In my estimation, the most daunting of these are the argument from explanatory exclusion and the contention that the functionalism that many nonreductive materialists espouse cannot accommodate irreducibly mental causal powers. I have attempted to answer these challenges, but one aspect of the anti-functionalist response bears highlighting. Common-sense functionalist characterizations of mental states need not await the results of scientific investigation. Hence, if such functionalist characterizations did capture the nature of mental states, and nonreductive materialism could accommodate functionalism, then in an important respect nonreductive materialism need not await the results of scientific investigation for its validation. However, Kim is right to claim that functionalism does not countenance mental causal powers, and for this reason functionalism is incompatible with a robust nonreductive materialism. As I have argued, mental state types would instead have to feature structural properties intrinsic to those types, and those structural properties must be distinct from any non-mental properties. But whether there exist structural properties of this sort is indeed a

matter for empirical investigation, which is currently incomplete, and for this reason one's confidence that a robust nonreductive materialism is true might have to be moderated. If it turns out that there are no intrinsic properties of the right sort to be found, the nonreductive materialist will be forced toward reductionism. In the meantime, if my responses to the counterarguments are plausible, one might be confident that a robust nonreductive materialism about mental states remains a serious option.

DERK PEREBOOM

University of Vermont

Notes

1. Jaegwon Kim discusses several nonreductive views that are not robust in this sense in *Mind in a Physical World* (Cambridge: MIT Press, 1998), pp. 67-87.

2. Hilary Putnam, "Language and Reality," in *Philosophical Papers, Volume 2*, pp. 272-90, at p. 278; (this paper was delivered as a Machette lecture at Princeton, 22 May, 1974). Elliott Sober makes a strong case for doubting the applicability of this alternative in "The Multiple Realizability Argument Against Reductionism," *Philosophy of Science* LXVI *(*1999): 542-64.

3. See, for example, Jaegwon Kim, "The Myth of Nonreductive Materialism," in his *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993), pp. 265-84, at pp. 280-2. (This article was first published in *Proceedings and Addresses of the American Philosophical Association* LXIII (1989): 31-47) Stephen Yablo provides a historical bibliography of the causal/explanatory exclusion argument in "Mental Causation," *The Philosophical Review* CI (1992): 245-80, at p. 247, note 5.

4. Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction," in his *Supervenience and Mind*, pp. 309-335, at p. 326. (This article was first published in *Philosophy and Phenomenological Research* LII (1992): 1-26.)

5. "Multiple Realization and the Metaphysics of Reduction," p. 326-7.

6. Terence Horgan, "Kim on Mental Causation and Causal Exclusion, *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 165-84, at p. 179. See also Kim's discussion of this move in *Mind in a Physical World,* pp. 67-72.

7. See Mark Moyer, *A Semantic Approach to Material Constitution,* unpublished Ph.D. Dissertation, Department of Philosophy, Rutgers University, New Brunswick, New Jersey, 2002.

8. cf. Saul Kripke, *Naming and Necessity* (Cambridge: Harvard University Press, 1980), pp. 144-8; Derk Pereboom and Hilary Kornblith, "The Metaphysics of Irreducibility," *Philosophical Studies* LXIII (1991): 125-145, at pp. 131-2; Lynne R. Baker, *Explaining Attitudes: A Practical Approach to the Mind* (Cambridge: Cambridge University Press, 1995), pp. 9-10.

9. cf. John Heil, "Multiple Realizability," *American Philosophical Quarterly* XXXVI (1999):189-208.

10. "The Metaphysics of Irreducibility," p. 131.

11. Because it allows for this sort of constitutional explanation, nonreductive materialism does not have the consequence that information about the brain is of little or no relevance to understanding psychological processes (cf. Pereboom and Kornblith, "The Metaphysics of Irreducibility," pp. 140-2). William Bechtel and Jennifer Mundale express the concern that nonreductive materialism will have this sort of result, "Multiple Realizability Revisited: Linking Cognitive and Neural States," *Philosophy of Science* LXVI (1999): 175-207, at p. 176.

12. Here I develop an argument I suggested in *Living Without Free Will* (Cambridge: Cambridge University Press, 2001), pp. 77-8.

13. Conceivably, two space ships might be made of such extraordinary material that they can fly through each other, for a moment wholly coinciding spatially (as Ted Sider argues in his review of Lynne Baker's *Persons and Bodies*, this JOURNAL, XCI, 1 (January 2002): 45-8.).

Intuitively, their causal powers might be explanatorily competitive. But it is not conceivable (at least to me) that two entities constituted of exactly the same stuff — two entities that are constitutionally coincident — might in this way be explanatorily competitive. Thanks to Mark Moyer for this point.

14. Barry Loewer argues that Kim's reasons for concern about overdetermination apply only to multiple determination by independent causes (like two assassins acting independently) but not to cases of overdetermination in which the causes are metaphysically connected (in his review of Kim's *Mind in a Physical World*, this JOURNAL XCVIII (2001): 315-324.) I endorse the substance of Loewer's criticism. However, as a terminological matter, it is not obvious to me that cases in which the causes are metaphysically connected in the way that M1 and P1 are should be classified as cases of overdetermination. The paradigm cases of overdetermination are those in which the causes are independent of one another (in a sense that one might want to spell out further) and do not include those in which the causal processes coincide constitutionally. Kim surely had in mind such paradigm cases, and he would be right to suggest that the mental/physical case does not fit this model. On the one hand, one might argue that due to such constitutional coincidence, it is best to say that there is no multiple determination, and thus no overdetermination, of M2 by P1 and M1, but rather a single determination, just as there would be if the mental and microphysical causal powers (and their activations) were identical. Or one might contend that because the causes are not identical, the case is best classified as one of overdetermination. I am neutral on this question — it seems clear that the important philosophical considerations are independent of any resolution of this terminological issue. Thomas Crisp and Ted Warfield also propose that Kim too hastily dismisses the

overdetermination solution in their "Kim's Master Argument," *Nous* XXXV (2001): 304-16.

15. The proposal that the causal powers of M and P do not compete because they coincide constitutionally is one of a broader class of proposals which also includes Stephen Yablo's suggestion that the causal powers of M and P don't compete because they are related as determinable and determinate -- "Mental Causation," and "Wide Causation," *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 251-81. His example of properties that stand in this relation are *red* and *scarlet – red* is a determinable of which *scarlet* is a determinate. It is clear that *red* and *scarlet* will not compete explanatorily. The relation between these two properties is more intimate than constitutional coincidence, and so if the relation between the causal powers of M and P could be shown to fit this model, there may be a respect in which we might then have a more satisfying solution yet to Kim's problem. But still, given that the causal powers of M and those of P coincide constitutionally, even if M's are not a determinable of which P's are the determinate, it is hard to see why we the solution we now have to the explanatory exclusion problem is not just as adequate as Kim's.

16. Jaegwon Kim, "Making Sense of Emergence," *Philosophical Studies* XCV (1999): 3-36, at p. 5.

17. "Making Sense of Emergence," pp. 6-7.

18. "Making Sense of Emergence," p. 8.

19. Brian McLaughlin, "The Rise and Fall of British Emergentism," in *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism,* eds. A. Beckermann, H. Flohr, and J. Kim

(New York: Walter de Gruyter, 1992); Mark Bedau, "Weak Emergence," *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 375-99, at pp. 376-7; Alexander Rueger, "Physical Emergence, Diachronic and Synchronic," *Synthese* CXXIV (2000): 297-322, at pp 317-8. For dissenting voices, see Timothy O'Connor, "Emergent Properties," *American Philosophical Quarterly* XXXI (1994): 91-104; and John Dupré, "The Solution to the Problem of the Freedom of the Will," *Philosophical Perspectives* X (Oxford: Blackwell, 1996), pp. 385-402. Strong emergence is sometimes contrasted with weak emergence (e.g. Bedau, Rueger), a notion we can ignore for present purposes. Randolph Clarke provides a good sketch of emergentism and argues that the nonreductivist can avoid it in "Nonreductive Physicalism and the Causal Powers of the Mental," Erkenntnis LI (1999): 295-322.

20. Whether a large piece of rock is a planet (*being a planet* is a relational property) is not predictable from its basal conditions, at least if the particles that constitute the system are only those that constitute the piece of rock itself. But, intuitively, this does not make *being a planet* an emergent property.

21. "Making Sense of Emergence," pp. 10-11.

22. "Making Sense of Emergence," pp. 28-31.

23. *Living Without Free Will*, pp. 69-74.

24. "Nonreductive Physicalism and the Causal Powers of the Mental," p. 309.

25. "Making Sense of Emergence," p. 32.

26. Ned Block, "Introduction: What is Functionalism?" in *Readings in the Philosophy of*

Psychology (Cambridge: Harvard University Press, 1980), pp. 171-84, at pp. 178-9.

27. "The Metaphysics of Irreducibility," p. 135.

28. "Multiple Realization and the Metaphysics of Reduction," p. 332.

29. "Making Sense of Emergence," p. 16.

30. "Making Sense of Emergence," p. 16. I deny (CIP-FP) for the following reason. Suppose M is a mental functional property realized at the neural level by N, and N is actually realized at the microphysical level by P1, but is multiply realizable at the microphysical level by P1 and P2. (CIP - FP) would have it that the causal powers of this instance of M are identical with the causal powers of N, and also with the causal powers of P1. By transitivity, the causal powers of N would be identical with the causal powers of P1, but by my argument above they will not be.

31. "Making Sense of Emergence," pp. 17-8.

32. In addition, one might note in this context a point made against functionalism by Paul Churchland, that the functionalist version of antireductionism can preserve theories that should be eliminated, such as alchemy and phlogiston theory, as easily as it can preserve the mental states of commonsense psychology ("Eliminative Materialism and the Propositional Attitudes," this JOURNAL LXXVIII: 67-90, at pp. 78-81). But in a nonreductive materialist view according to which mental states are irreducibly mental intrinsic properties of this sort of worry will not arise.

33. Derk Pereboom, "Why a Scientific Realist Cannot Be a Functionalist," *Synthese* LXXXVIII (1991): 341-358.

34. Hilary Putnam, “Brains and Behavior,” *Philosophical Papers, Volume 2* (Cambridge: Cambridge University Press, 1975), pp. 325-41, (first published in *Analytical Philosophy Second Series* (Oxford: Blackwell, 1965), pp. 1-19.)

35. Hilary Putnam, "Language and Reality," p. 278. Putnam endorsed functionalism in "The Nature of Mental States," in *Philosophical Papers, Volume 2*, pp. 429-40, (first published as “Psychological Predicates,” in *Art, Mind, and Religion,* Capitan and Merrill, eds. (Pittsburgh: University of Pittsburgh Press, 1967), pp. 37-48.)

36. "Why a Scientific Realist Cannot Be a Functionalist," pp. 348-50.

37. This argument has the conclusion that the general fact that pain causes winces cannot adequately be explained by the general features of pain supposing that it is a functional kind. But this does not show that there is no context in which a particular wince might be adequately explained by pain, supposing that it is a functional kind.

38. Block argues for a stronger anti-functionalist thesis, that functional properties cannot be causally efficacious in standard cases (cases in which no intelligent being recognizes them) and not only, as I have contended, that they cannot adequately explain dispositional features of kinds at the type-level ("Can the Mind Change the World," in *Meaning and Method* (Cambridge: Cambridge University Press, 1990), pp. 137-170). Functional properties, he argues, are second-order properties, and second-order properties cannot be causally relevant in standard cases. Examples help confirm this point. The property *dormitivity* is a classic example of second-order property, defined as follows:

x is dormitive = x has some property that tends to cause sleep.

When a dormitive pill is slipped into your food it is the chemical property of the pill, not dormitivity, that causes sleep. In addition, Block points out that the relation between a second-order property and an effect is logical, and not typically nomological. Consider the proposal: having some property that tends to cause sleep caused him to sleep. This claim would seem to specify a logical relation, and not (necessarily) a causal one. There might be a nomological relation in some such cases, but in most examples of this kind there seems to be no reason to posit one. (cf. Ruth Millikan, "Historical Kinds and the 'Special Sciences'," *Philosophical Studies* XCV (1999): 45-65).

39. Richard Boyd, "Scientific Realism and Naturalistic Epistemology," in *Proceedings of the Philosophy of Science Association* 1980, Vol. 2, pp. 613-62, at p. 642..

40. See Boyd's "Kinds, Complexity, and Multiple Realizations, " *Philosophical Studies* XCV (1999): 67-98 for his view of the prospects for psychological kinds fitting this model. He appears to be less optimistic than I am.

41. Robert Van Gulick's conception of "higher order patterns" that "have a degree of independence from their underlying physical realizations" also inspires the sort of view I am developing here; cf. "Who's in Charge Here? And Who's Doing All the Work?" in J. Heil and A. Mele eds., *Mental Causation* (Oxford: Oxford University Press, 1993), pp. 233-56, at pp. 249-56.

42. "Why a Scientific Realist Cannot Be a Functionalist," pp. 350-1.

43. See the website: http://www.ballpistonengine.com.

44. I suspect that the same point can be made with Putnam's peg-and-board example, in "Language and Reality," pp. 295-8. If Philip Kitcher is right about the irreducibility of certain biological properties, it would appear that analogous claims would hold for them; "1953 and All That: A Tale of Two Sciences," *The Philosophical Review* XCIII (1984): 335-73.

45. On abstractness of this sort, see Richard Boyd, "Kinds, Complexity, and Multiple Realization," pp. 91-6.

46. cf. Richard Boyd, "Kinds, Complexity, and Multiple Realization," pp. 71-2.

47. Jerry Fodor, "Special Sciences," *Synthese* XXVIII (1974): 97-115. Kornblith and I endorsed a version of this argument in "The Metaphysics of Irreducibility," pp. 126-8, against which William Jaworski has advanced an impressive counterargument in "Multiple Realizability, Explanation, and the Disjunctive Move," *Philosophical Studies* CVIII (2002): 298-308, especially pp. 301-3.

48. "Multiple Realization and the Metaphysics of Reduction" pp. 319-325.

49. Ned Block, "Anti-Reductionism Slaps Back," in *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 107-132, at p. 109.

50. Jerry Fodor argues that Kim's example merely involves a sampling error; "Special Sciences: Still Autonomous After All These Years," *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 149-63, at pp. 151-2.

51. "Multiple Realization and the Metaphysics of Reduction," p. 321; cf. Lenny Clapp "Disjunctive Properties: Multiple Realizations," this JOURNAL XCVIII: 111-36, at pp. 120-1.

52. “Anti-Reductionism Slaps Back," pp. 126-7. Louise Antony and Joseph Levine make a similar point in "Reduction with Autonomy," *Philosophical Perspectives* XI (Oxford: Blackwell, 1997), pp. 83-105, at pp. 1997, 90-1.

53. “Anti-Reductionism Slaps Back," pp. 120-9.

54. cf. Antony and Levine, "Reduction with Autonomy," pp. 92-4.

55. cf. Lenny Clapp, “Disjunctive Properties: Multiple Realizations,” pp. 123-32. Clapp provides a fine account as to how a disjunctive predicate can indicate a nondisjuctive property on a causal powers notion of properties.

56. L. Diamond, "Neuraesthenia," in Benjamin Wolman, ed., *International Encyclopaedia of Psychiatry, Psychology, Psychoanalysis, and Neurology* VIII (New York: Aesculapius, 1977), pp. 27-8.

57. Hilary Kornblith, *Knowledge Without Foundations: A Causal Theory*, unpublished Ph.D. dissertation, Department of Philosophy, Cornell University, Ithaca, New York, 1979, pp. 120-3; Chatel and Peele "The Concept of Neuraesthenia," *International Journal of Psychiatry* IX (1970): 36-49.

58. Lawrence Shapiro, “Multiple Realizations,” this JOURNAL XCVII (2000): 635-54, at p. 647.

59. “Multiple Realizations,” p. 645.

60. “Multiple Realizations,” p. 644.

61. Oron Shagrir, “Multiple Realization, Computation, and the Taxonomy of Psychological

States," *Synthese* CXIV (1998)*:* 445-61, at pp. 454-5.

62. "Multiple Realizability Revisited," p. 204.

63. Some might want to contend that these common features would not be psychological because they would have a physical-structural description, and anything that has this sort of description is *per se* not psychological. In this conception, anything that is psychological is describable only in familiarly mentalistic terms. But this view, advocated by some of the logical positivists, mistakenly ties the notion of the psychological too closely to a type of vocabulary. Moreover, the debate between nonreductive and reductive materialists should not be conflated with the debate between reductive materialists and those who hold that the mental is in no sense physical. (For an opposing view see R. Endicott, "On Physical Multiple Realization," *Pacific Philosophical Quarterly* LXX (1989): 212-24.) The claim that mental states are not identical with states that have a physical/structural description akin to descriptions of states in sciences such as biology and chemistry should not be regarded as an essential feature of nonreductive materialism.

64. Paul M. Churchland and Patricia. S. Churchland, "Intertheoretic Reduction," in their *On the Contrary* (Cambridge, MA: MIT Press, 1998), pp. 65-79, at p. 78. They cite temperature as an example of a property that has been domain-specifically reduced – their claim is that temperature is reduced differently in a gas, a solid, and in a vacuum. In my view, the differences among temperature in a gas, a solid, and a vacuum indicate that temperature is multiply realizable, but not that it has distinct reductions. For as Kornblith and I have argued, there is a unitary property that these cases of temperature have in common, with which temperature should be identified ("The Metaphysics of Irreducibility," pp. 138-9).

65. "Multiple Realizability Revisited," pp. 201-4; cf. Jaegwon Kim, "Phenomenal Properties, Psychophysical Laws, and the Identity Theory," *Monist* LVI (1972): 190-1, at p. 190; L. Mucciolo, "The Identity Thesis and Neurophysiology," *Nous* VIII (1974): 327-42.

66. "Multiple Realization, Computation, and the Taxonomy of Psychological States," pp. 451-2.

67. As a case in point, consider Yablo's argument for the primacy of the mental tokens over physical tokens as causes of actions in his "Mental Causation." He argues that when an event, such as *my ringing the doorbell*, has *prima facie* as causes both a token mental event -- a decision --, and a token physical event -- the physical realization of the decision --, it might well be that the mental event is *the* cause, albeit that both events are causally relevant. In his conception, which event is the cause depends on considerations of *proportionality*. Yablo's characterization of proportionality is complex, but in this case the relevant component is that the cause be *required* for the effect. In his definition of this notion, x's are causes, y's are effects, and event x^- < event x just in case x^- is less specific than x; for example *Socrates' drinking the potion* is less specific than *Socrates' guzzling the potion*:

> x is required for y just in case for all x^- < x, if x^- had occurred without x, then y would not have occurred. (p. 276)

In Yablo's estimation, the physical realization of the decision is more specific than the decision itself, for the reason that there are many distinct possible physical realizations of the decision. So the decision – the mental event – is the lesser of the two. Is it true that if the mental event had occurred without its specific physical realization, then my ringing of the doorbell would not have occurred?

Of course the decision had a physical determination p; but most people would also say,

> and I agree again, that it would still have been succeeded by the ringing, if it had occurred in a different physical way, that is, if its physical determination had been not p but some other physical event. And this is just to say that p was not *required* for the effect. (p. 278)

The relevant alternative physical event is the physical determination (realization) of the decision in the nearest world in which it is different from its actual physical determination. The problem is that Yablo's argument all by itself does not establish that the cause won't be, for example, neural rather than mental. Here it is important to distinguish two distinct "physical" realizations of the decision, the neural and the microphysical. I think Yablo would be right to claim that if the decision had occurred in a relevant alternative microphysical way, the effect would still have occurred. So the cause won't be the actual microphysical realization of the decision. But this still leaves both the neural realization and the mental event itself as candidates for the cause. And the neural realization cannot be dismissed on the grounds that the decision would have been succeeded by the ringing if it had occurred in a different microphysical way, but only if it would have been succeeded by the ringing had it occurred in a different neural way. And whether it can be dismissed in this way is not immediately obvious. This worry about Yablo's argument can also be raised against Kripke's concerns about the token-identity thesis (*Naming and Necessity*, pp. 144-8), which Yablo cites (p. 269).