VARIETIES OF EXCLUSION

Marcelo H. SABATES*

* 204 Kedzie Hall, Philosophy Department, Kansas State University, Manhattan, KS 66506, USA. E-mail: sabates@ksu.edu

BIBLID [0495-4548 (2001) 16: 40; p. 13-42]

ABSTRACT: The problem of exclusion threatens non-reductive physicalist theories of the mind by implying that they cannot account for mental causation. This paper attempts to clarify what exactly the exclusion problem is, and, given the problem, to survey the theoretical options open. First I reconstruct the problem from its most influential sources (Malcolm and Kim), showing that it should be understood as an ontological rather than an explanatory problem. I then distinguish the problem from some consequences that seem to follow from it. Finally I sketch a map of possible answers to exclusion.

Keywords: mental causation, exclusion, non-reductive physicalism, psychological explanation, epiphenomenon, Malcolm, Kim.

CONTENTS

When Thomas Reid became aware of the apparently inevitable consequences that Hume extracted from the theory of ideas, he wrote:

> The theory of ideas, like the Trojan horse, had a specious appearance both of innocence and beauty; but if those philosophers had known that it carried in its belly death and destruction to all science and common sense, they would not have broken down their walls to give it admittance.[1]

The theory of ideas was the epistemological orthodoxy of part of the seventeenth and the eighteenth centuries. In the second half of the twentieth century there is an orthodoxy regarding the nature of human beings: non-reductive physicalism. The view, carrying a specious appearance of innocence and beauty, promises loyalty to a broadly naturalist or physicalist metaphysics appropriate to the times, while keeping what seems an ineliminable part of our dignity as human beings: the autonomy of our minds. Already by the early 1970's philosophers of mind had broken down their walls and gave admittance, with very few exceptions, to the Trojan horse.[2]

In the early 1980's philosophers began to inspect the horse's belly. Some of them smelt death and thus different versions of non-reductive physicalism came under attack, for very different reasons. The consequences, however, typically involved the claim that given one or another version of non-reductive physicalism, and for one or another kind of mental state, the mental was left without causal efficacy. It is not surprising that the debate generated by these attacks populated the landscape in analytical philosophy of mind. In the 1980's and early 1990's the dominant discussions were about content externalism and psychophysical laws, including Davidson's anomalism.[3] In the mid and late 1990's the prevalent debates within the field have been about consciousness and exclusion. It can be argued that the problem of exclusion, as presented by Norman Malcolm and Jaegwon Kim, is the more general of all the attacks since it targets every type of non-reductive physicalism and every kind of mental state. The Trojan horse of non-reductive physicalism seems to carry with it the inability of our minds to make any causal difference in the world, including our own actions and even our own thoughts, bringing destruction to all science (of the mind, at least) and common sense.[4]

We need to understand what exactly can make the horse be so deadly. In recent times it has been widely accepted that there are several different problems under the label of "the problem of consciousness" and, arguably, the discussion has gained from a careful distinction between such problems and their relations.[5] Surprisingly, no parallel work has been done regarding the problem of exclusion. In what follows I shall attempt a survey of several exclusion problems and their interconnections. In section 1, I distinguish between causal (metaphysical) and explanatory (epistemological) problems of exclusion and argue that the metaphysical problem is the "hard" one. In section 2, I differentiate between a causal and a general dependence version of exclusion and argue that the full metaphysical problem requires the general dependence formulation. In section 3, I distinguish the metaphysical problem of exclusion from some potential consequences of the problem. In section 4, I sketch a map of

answers to the metaphysical problem of exclusion including some reference to the explanatory problem. At no point I recommend what to do with the horse, but some sympathies towards some attitudes will be apparent in sections 3 and 4.

1. Explanatory Exclusion and Causal Exclusion

The general format of all exclusion problems includes the claim that the mental is excluded by the neural or the physical.[6] It is said that the mental loses its efficacy or role since it is "preempted" or "ruled out" or "screened off" by the neural. The problem, thus, suggests that there is no way of finding a place for the mental in a physicalist or naturalist ontology. Since broadly naturalist ontologies are orthodoxy nowadays, this problem is of utmost importance for contemporary philosophy. Within that general format, however, the exclusion problem has taken different, sometimes non-equivalent forms throughout the years; sometimes appealing to explanatory considerations, sometimes to causal considerations and sometimes to both.[7] I shall review what I consider some of the most influential explanatory versions of the exclusion problem and argue that they ultimately rely on causal considerations. Moreover, I shall show that the only independent argument for explanatory exclusion is much weaker than the causal arguments.

Let us say without much qualification that an argument against the efficacy of the mental belongs to the explanatory exclusion family iff it uses (as an assumption or as a partial conclusion) the claim or principle

> (EE): There cannot be more than one complete and independent explanation of the same phenomenon.

or some related claim.[8] When we add that there is (whether we have it or not) a complete neural/chemical explanation of, say, a particular behavior, we conclude that no mental/psychological explanation can be relevant.

On the other hand, let's say that an argument against the efficacy of the mental belongs to the causal exclusion family iff it uses (as an assumption or as a partial conclusion) the claim or principle

> (CE): There cannot be more than one complete and independent cause of the same phenomenon.

When we add that there is a complete neural/chemical cause of, say, a particular behavior, we conclude that no mental/psychological state can be causally efficacious.

1.1. Malcolm's Argument: Mechanism and Purpose

What is perhaps the first articulated version of the exclusion problem in contemporary philosophy,[9] namely Norman Malcolm's opposition between mechanism and purpose, belongs to the explanatory exclusion family.[10] And this is so because Malcolm explicitly presents the tension between a physical and a mental account of behaviors in terms of explanation and prediction. He assumes that a neurophysiological theory will be adequate to *fully* explain and predict *all* movements of human bodies not due to external physical causes, "the movements that occur when a person signals a taxi, plays chess, writes an essay, or walks to the store" (Malcolm 1968, p. 127). Such a physiological theory will be thus systematic and complete. Systematic since given a realm of phenomena, the theory will explain *all* phenomena within its domain; complete since the explanations given by the theory are entirely *sufficient.* Besides, it will be non-purposive in the sense that it will make "no provision for desires, aims, goals, purposes, motives or intentions" (Malcolm 1968, p. 128).

On the other hand, everyday explanations of behavior refer to purposes, desires and intentions. Behaviors are claimed to occur because we have the purpose of bringing about (or avoiding) some state of affairs. Suppose we want to explain why a man is climbing a ladder. We can say that he is climbing the ladder in order to retrieve his hat from the roof. Malcolm says:

> This explanation relates his climbing to his intention. A neurophysiological explanation of his climbing would say nothing about his intention but would connect his movements on the ladder with chemical changes in body tissues or with the firing of neurons

and he asks whether the two accounts interfere with each other. His reply to this question leads us to explanatory exclusion:

> I believe there would be a collision between the two accounts if they were offered as explanations of one and the same occurrence of a man's climbing a ladder. We will recall that the envisaged neurophysiological theory was supposed to provide sufficient causal explanations of behaviour. Thus, the movements of the man on the ladder would be completely accounted for in terms of electrical, chemical and mechanical processes in his body. This would surely imply that his desire or intention to retrieve his hat had nothing to do with his movement up the ladder (Malcolm 1968, p. 133).

Thus, the ingredients of the argument seem to be that the physical or neurophysiological explanatory system is complete in the sense of being sufficient to explain a behavior, and that once we have such a sufficient explanation any other explanation will be excluded or preempted. And the argument belongs to the explanatory exclusion family since it appeals to an explanatory formu-

lation of the ingredients, and particularly to the idea of a complete explanation as ruling out other explanation-candidates.

But is (EE) so obvious to make the exclusion problem a compelling one? Malcolm does not seem to rely on the intuitiveness of the explanatory formulation. He sees the need to explain why (EE) has to be accepted. And when he has to demonstrate why there is exclusion, the argument runs in causal terms. First, he reintroduces the completeness of neurophysiology in causal terminology: he rapidly shifts from his original formulation in terms of explanation to one using "causal explanation" instead. And later, when he develops the justification for the exclusion principle, he drops completely his reference to explanations. What he assumes is that "the neurophysiological theory would provide sufficient *causal conditions* for all movements" and that neurophysiology "is a closed system in the sense that it does not admit, as antecedent conditions, anything other than neurophysiological *states and processes*" (Malcolm 1968, p. 136, my italics). The following passage, which I take to be the main and more detailed presentation of the argument in Malcolm's paper, is revealing: he not only gives causal reasons for the exclusion principle itself, but also couches the whole argument in causal terminology:

> But if we bear in mind the comprehensive aspects of the neurophysiological theory -that is, the fact that it provides sufficient causal conditions for all movements- we shall see that desires and intentions could not be causes of movements. It has often been noted that to say B causes C does not mean merely that whenever B occurs, C occurs. Causation also has subjunctive and counterfactual implications: if B were to occur, C would occur; and if B had not occurred, C would not have occurred. But the neurophysiological theory would provide sufficient causal conditions for every human movement, and so there would be no cases at all in which certain movement would not have occurred if the person had not had this desire or intention. Since the counterfactual would be false in all cases, desires and intentions would not be causes of human movements (Malcolm 1968, p. 136).

To put the argument in a few words: once we have sufficient causal conditions, nothing else can have causal efficacy since it doesn't have the counterfactual force required by causation.[11] It is interesting to note that the conclusion itself is presented in causal idioms: desires and intentions cannot be causes of bodily movements.

Explanations are in conflict because there is a competition for causal role between mental and neurophysiological states or properties[12] and we assume that both neurophysiological and psychological explanations are causal explanations. And physical (or neurological) explanations preempt psychological explanations because we have ontological reasons to believe that physical states or properties preempt the role of mental properties as causes of behavior, and

we assume that such ontological reasons are relevant also at the epistemological level. Malcolm's argument, in spite of being presented generally in explanatory terms and thus apparently falling under the explanatory exclusion principle, relies on causal considerations and in particular on the causal principle (CE).

1.2. Kim's Argument: Overdetermination or Preemption

What has been perhaps the most influential treatment of the exclusion problem, namely Jaegwon Kim's papers 'Explanatory Realism, Causal Realism and Explanatory Exclusion' and 'Mechanism, Purpose and Explanatory Exclusion', presents the problem, as it is indicated in the titles, primarily in an explanatory way. Through his several papers on this issue, Kim fluctuates between causal and explanatory formulations of the exclusion problem and of the exclusion principle that grounds it, so it would be inaccurate to present him as favoring the explanatory version.[13] However, it is interesting to see how the explanatory principle is argued for and used in the case of these seminal papers.

Kim affirms that Malcolm is fundamentally right when he argues that rationalizing explanations and physiological explanations exclude each other. The reason is the plausibility of (EE), in Kim's words: "there can be no more than a single *complete* and *independent* explanation of any one event" (1988, p. 233). The burden of Kim's papers is to show that when we plausibly have more than one competing explanation, either they are not complete or one is dependent on the other. And his defense of (EE) in these two papers is mainly "topic-neutral", i.e., largely independent of the subject-matter of the explanations' referring to mental and physical states.[14]

Kim sees, as Malcolm did, the need to justify explanatory exclusion. When (EE) is defended, its justification invariably appeals to causal considerations.[15] There are, in Kim's formulation (as well as in Malcolm's), two shifts that make this clear. First, when support is sought for (EE) the formulation shifts to encompass just causal explanations. Kim says: "It seems to me that the case for explanatory exclusion is most persuasively made for causal explanations (...)"[16] and proceeds to make his case accordingly. Second, and more importantly, when it is argued that causal explanations exclude each other, reasons are given in terms of "sufficient causes", "causal links" and "causal overdetermination". This is particularly important since exclusion for the case of the mental is defended by showing the implausibility of alternative possibilities, and such possibilities are all causally formulated. Thus, it is argued that when competing causes cannot be overdetermining causes (or partial causes or two

causes belonging to the same causal chain), one of them has to "screen off" the other.[17] Now, strictly speaking, causal argumentation only supports the causal principle of exclusion (CE).

Kim is well aware of this and this is why the transition from causal argumentation to (EE) depends (and this is explicitly stated in his (1988)) on the view that explanations are grounded on objective relations such as causation (this view might be called explanatory realism). But the extension from (CE) to (EE) is not obvious. I shall address this extension below, but we can safely say that Kim's classical formulation of the exclusion problem is mainly causal and only derivatively explanatory.

1.3. Preferring the Causal Formulation

By analyzing Malcolm's and Kim's expositions of the exclusion problem, I tried to show that causal considerations were in fact playing the crucial role in explanatory formulations. This is no accident: an explanatory principle is not as easy to argue for as a causal principle. I shall mention some of the reasons and leave some for section 3.

Even if we conceded that the results of causal argumentation are automatically relevant for explanation, they would be so only for *causal* explanations. Given (CE) all we might infer (if we think that what counts for causation counts for explanations) is a weaker exclusion principle for explanations, namely:

> Causal Explanation Exclusion (CEE): There cannot be more than one complete and independent *causal* explanation of the same phenomenon.

It is clear that (CEE) does not yield (EE). We want to remain neutral on whether there can be explanations other than causal explanations (in fact, Kim himself thinks that since there are important "world-cementing" non-causal relations there might well be non-causal explanations). Thus, to use (CE) to exclude more than rival *causal* explanations would be a *non sequitur.*

But even the extension from (CE) to (CEE) may be challenged. For causation is an ontological relation and explanation is an epistemological one.[18] And it may well be that what applies at the ontological level (the exclusion principle between competing causes) does not apply at the epistemological level (the exclusion principle between competing explanations). What I take to be the best account of why an explanation explains, namely, explanatory realism, is the natural candidate for closing this gap, but it is not obvious that epistemological reasons wouldn't prevent us from adopting an exclusion principle at the explanatory level. The previous two reasons may lead us to think

that (EE) is an unjustified extension of (CE). And as I shall argue in section 3 there is a stronger reason for thinking this: the best arguments we have to support (CE) do not support (EE).

There is, however, an epistemological argument for (EE) that is totally independent from causal considerations. It says that if we accept more than one explanation of the same event this would result in some sort of explanatory overpopulation. An account of explanation should take into consideration the unifying and simplifying character of explanatory activity. Kim puts it this way: "when two distinct explanations are produced to account for a single phenomenon, we seem to be headed in a direction opposite to the maxim of explanatory simplification: 'explain as much as you can with the fewest explanatory premises'" (1989a, p. 254). Still, this falls short of establishing (EE). Even an explanatory realist should accept that explaining is an epistemological activity and what counts as a good explanation depends (within the limits fixed by objective relations) on the epistemic situation of those in need of understanding (cf. Kim 1988, p. 225-6). On one hand, in different epistemic contexts, each of the "competing" explanations might enhance our understanding of the explanandum much more than the other. On the other hand, we might be able to provide one of the "competing" explanations but not the other.[19] It seems clear that these are *epistemic* gains that result from allowing two explanations of the same phenomenon. And these gains are to be weighed against the lack of simplicity.

Thus, there is no straightforward argument from a maxim of simplicity to explanatory exclusion. With no decisive argument for explanatory exclusion and with the explanatory principle needing support from causal arguments, it seems reasonable to conclude that causal exclusion is the "hard" problem of exclusion. This also means that challenges to exclusion that stress the implausibility of (EE), sometimes with considerations similar to the ones presented above regarding different "competing" explanations enhancing our understanding of the explanandum more than the other depending on the epistemic context, may not touch the hard problem.

2. Causal Exclusion and Ontological Exclusion

What makes the causal exclusion problem a hard problem? For some, (CE) may be intuitive enough to be used as an assumption and thus generate the exclusion problem. But I think that what makes causal exclusion particularly compelling is that (CE) does not need to be an assumption. It is a consequence of a family of views about the mental that has been the overwhelming orthodoxy for at least thirty years. This family of views defends substance monism

(no souls or non-natural substances) plus mental realism (mental properties are real properties, they are not identical to neural/physical ones) plus the primacy of the physical (physical properties are basic or fundamental with respect to mental properties, which in turn depend on them). Many versions of functionalism, emergentism, supervenientism and even epiphenomenalism share these three claims. I shall present a version of exclusion that is generated by this family of views and argue that (CE) is only half of what is needed to challenge the causal efficacy of the mental.

Substance monism and mental realism are, at least for our purposes, self-explanatory.[20] How do we articulate the basicness of the physical? Most physicalists agree that the following two claims express the minimal commitment a physicalist has to accept.

> (SS) Mental properties depend on physical properties in the sense that they strongly supervene upon physical properties. This means that necessarily, for any property M belonging to the family of mental properties, if something has M, then there is a property P belonging to the family of physical properties, such that that thing has P, and necessarily, anything that has P has M.[21]
>
> (WC) Physical properties are self-sufficient in the sense that the physical is closed under causation. This means that every instantiation of a physical property has a complete generating causal chain entirely composed of instantiations of physical properties.[22]

2.1. Mental-to-Mental Causation

Now, when is a mental property causally efficacious? When it has the potential to cause the instantiation of other properties.[23] There are three candidates for the role of effects caused by mental properties: other mental properties, physical properties, and higher-than-mental properties. So we have to consider these three possibilities.

Suppose that a mental property M is said to cause another mental property M'. Let's say that my belief that Maura is in Kansas City for a short visit (M) causes my desire to see Maura (M'). But M' is, by (SS), dependent on a physical property P', an appropriate neural condition. And given this dependence relation, P' is alone sufficient to bring about M' (recall that necessarily, every individual possessing P' possess M'). So unless we think that my desire to see Maura (M') is overdetermined, we have to conclude that the only way in which my belief that Maura is in Kansas City (M) can be causally efficacious for the

occurrence of my desire M' is being in some way causally responsible for the occurrence of its neural base P'.

It is important to note that this will be the case for every mental property. So the alternative is that each time in which a mental property causes another we have a case of overdetermination. Pervasive overdetermination gives us an implausible and inelegant picture of reality, and many are reluctant to countenance this. Stephen Schiffer, for instance, says: "This causal superfluousness is hard to believe in; it is hard to believe that God is such a bad engineer." (1987, p. 148). However, the main reason to reject overdetermination is, I think, that it threatens physicalism itself. For if there were a complete non-physical cause for my desire M' that cause would be sufficient for M', and so it could be the case that M' existed without any physical base. But a physical base for every mental property is required by (SS). Therefore, the overdetermination of the mental violates physicalism.

Why is it that when we are confronted with competing properties overdetermination is the only real alternative to one property screening off the other? Perhaps our competing properties are not in fact independent from each other, or perhaps none of them is a complete cause of the effect. These are real alternatives in many cases in which there are two competing causes (an improvement in competence in a foreign language caused by continuous exposure to native speakers and by hard study, for instance). But we are entitled to rule out the following possibilities in the case of mental-to-mental causation.[24] M cannot be a partial cause of M', since if it were, P' wouldn't be sufficient for M' thus violating (SS). Moreover, it wouldn't make sense to say that M is not independent from P' because of M being identical to P' (if P' is identical to a mental property, it would be identical to M' -a possibility that will be ruled out below, anyway). There is, finally, another way in which M and P' can be non-independent: by being different links of the same causal chain. And it is because of this option that the conclusion of the mental-to-mental causation case should be hypothetical: if there is mental-to-mental causation, it must be through mental-to-physical causation.

Before discussing mental-to physical causation, we can easily rule out that a mental property M causes higher-than-mental properties, let's say a "social" property. Can M be the cause of S', a higher-than-mental property that depends on another mental property M'? By an argument analogous to one considered above, we can show that mental to higher-than-mental causation presupposes mental-to-mental causation, which in turn presupposes mental-to-physical causation, which is what needs to be addressed.

2.2. Mental-to-Physical Causation

So my belief M must be able to cause P', the physical base of my desire M'. Note that this case is similar for our purposes to the case in which a desire (my desire to listen to Hindelmith's *Der Schwanender*, for instance) causes an action (my turning on the CD player, for instance). Now, according to (WC), P' must have a complete generating causal chain composed of physical properties. And it seems entirely plausible that a property P, the property constituting the physical base of my belief M, be taken as the cause of P', the physical base of my desire M'. But if P is sufficient for the instantiation of P', my belief M has no causal role in this picture unless we claim that P' is overdetermined. Here again, overdetermination should be ruled out not only because of elegance and plausibility considerations but also because it violates a physicalist commitment. For if there were an overdetermining non-physical cause of P', P' could have been instantiated without any physical cause in its ancestry, thus violating claim (WC). It could be replied that since, by (SS), M has to have a dependence base, the situation we are entertaining here is not possible. However, M can be realized in different physical properties (Pi, Pj, Pk, etc.), and it seems plausible to think that not all of these alternative bases would cause P'. So the situation in which P' is caused only by a non-physical cause is possible and the argument against overdetermination stands. Thus, my belief M cannot be causally efficacious for P' (and *mutatis mutandi* for any physical property); and in this way it cannot be responsible for my desire M' (and *mutatis mutandi* for any mental property): M seems causally impotent, and this generalizes over every mental property.

Again, in the case of mental-to-physical causation, we are entitled to rule out alternatives other than overdetermination. Can the competing causes M (my belief that Maura is in Kansas City for a short visit) and P (M's neural base) be partial causes? This would again violate thesis (WC) since a non-physical property would be necessary for the instantiation of a physical property. Can M and P be the same property? We cannot rule out this possibility *in abstracto*, but it should be clear that in the context of a mental realist theory the proposed identity can hardly be accepted.[25] Can M and P be links belonging to the same causal chain, a causal chain that leads to P'? It seems obvious that M cannot be a cause of P. For it doesn't make sense to claim that a supervenient property can be the cause of its base.[26] So the only remaining option is that P causes M. However, non-causal dependence relations are different from causal relations so this answer is implausible.[27] Moreover, this possibility would make M part of the causal chain leading to P', plausibly violating (WC).

Our two-step conclusion is, then, that for a mental property M to be able to cause another mental property M', M has to cause a physical property P' (surely the physical base of M'). But M happens to be unable to cause P' either. So M is causally inefficacious. Giving admittance to the horse has proven destructive.

2.3. Ontological Exclusion

Interestingly enough, we didn't use the principle (CE) as an additional assumption since from the very claims (SS) and (WC) we were able to discard all the options leaving us with exclusion. In the case in which M competes with P as a cause of P' (mental-to-physical causation) once we eliminate the possibility of the overdetermination of P' by M and P (and the other possibilities), one of the two purported causes has to exclude or preempt the other. And as physicalists we say that P preempts the causal role of M, the putative mental cause. However, in the case in which my belief M competes with the neural property P' for the "authorship" of my desire M' (mental-to-mental causation), it is not the case that one of two purported *causes* excludes or preempts the other. For in this case no one is claiming that P' is a cause of M'. What is at stake here is a more general principle that comprises the causal exclusion principle as a special case. This general principle, which is inevitable for a full formulation of the argument, might be dubbed the ontological exclusion principle:

(OE): there cannot be more than one independent and complete "necessitator" of the same phenomenon.

This means that there cannot be two independent and complete (sets of) properties each of which fully necessitates some single property. So we can say that once we discard the possibility of the overdetermination of M' by M and P', one of the two purported "necessitators" has to exclude the other. And as physicalists we say that P' preempts the causal role of M, the putative mental cause.

A principle like (OE) inevitably results when we formulate the exclusion problem as covering the mental-to-mental causation case. Since most of the formulations are interested mainly in mental-to-physical causation (belief-desire causing action), it is not surprising that this fact has not been noticed.[28] But once we obtained this general principle from the very physicalist assumptions we can provide a perspicuous summary of the exclusion problem: given the ontological exclusion principle (OE), in the case of mental-to-mental causation, the physical base of the mental effect preempts (via supervenience) the causal role of the putative mental cause; and in the case of mental-to-physical

causation, the physical base of the putative mental cause preempts (via causality) its causal role. Therefore, the mental is causally inefficacious.

One of the conclusions of section 1 was that concentrating on the causal formulation might close the door to some objections to exclusion that are purely epistemological. But the results of this section seem to point to the opposite direction. We know now that in order to declare the mental fully inefficacious we need more than causal exclusion. And what we should add is, taken with no qualification, controversial. Is it plausible to claim that any pair of relations of necessitation compete with (and preempt) each other? The evening star's being visible today necessitates Venus' being visible today, but the evening star's being visible today does not seem to prevent the evening being clear in Kansas today from causing Venus' being visible today. If (OE) has any chance to be plausible, it needs to be qualified: necessitation relations exclude each other as producers of an effect (or consequence) only if they are both conceptually/metaphysically contingent. If one of the putative competitors in question is non-contingently related to the effect (because it is just a synonym, or a determinable of it), exclusion does not occur.[29] But now the burden is on the defender of exclusion to show that the supervenience relation between the mental and the physical is conceptually/metaphysically contingent. There is no agreement about this (in part because there has been almost no explicit discussion of this), but physicalists seem to be moving closer to mind/body supervenience being a conceptually/metaphysically necessary relation.[30]

If the proposed qualification does not help, mental-to-mental causation still stands: there is no reason to think that M and P' will compete for the production of M'. True, this only saves half of mental causation, but it may open an interesting strategy for recovering the other half. If we consider actions to be themselves mental states supervening upon bodily movements, we save causation of intentional behavior by saving mental-to-mental causation. What has been called "dual explanandum" strategy may gain new life in this way.

3. Exclusion and its Consequences

The exclusion argument shows that under the orthodox non-reductive physicalism the mental is causally inefficacious. The exclusion debate, however, sometimes includes discussion about potential consequences of the causal inefficacy of the mental, consequences that would certainly make the destruction to all science of the mind and common sense even worse. There are, I think, at least three of these consequences that deserve attention. While these are all plausible consequences of exclusion they are not part of the problem itself, and distinguishing the consequences from the problem can be helpful when it comes to

assess the different approaches that can be taken vis-à-vis exclusion. Let us, then, discuss the consequences in turn.

3.1. Mental Irrealism

Under apparently plausible assumptions, the causal inefficacy of the mental amounts to mental irrealism. Realism, in the sense I am using the concept, makes an existential claim about a type of entities. Mental irrealism is the view defending that there is no entity (object, property or whatever) that counts as mental. Since we are using property terminology, we can say that mental irrealism claims that there are no genuine mental properties.

A majority (but not unanimous) view about the nature of properties argues that the criterion to determine when a property (or a kind of properties) is real is a causal power criterion:

> (CPC) A property is real only if it contributes to the active causal powers of the object that has it.[31]

Or, in our terminology, a property is real only if it is causally efficacious. Given (CPC) the consequence we are forced to draw from the exclusion argument is that mental properties cannot be real, since they are causally inefficacious. Kim himself, by endorsing what he calls the "Alexander Dictum" ("to be real is to have causal powers"), has championed irrealism as an inevitable result of exclusion. The mental being causally inefficacious was an already damaging consequence of non-reductive physicalism. But mental irrealism is even more devastating, since, as it should be obvious, it contradicts a constitutive claim of non-reductive physicalism, namely, that mental properties conform a realm of real properties that cannot be eliminated in favor of physical properties: we cannot have both the physicalist and the realist components of the most popular family of theories of the mind.

At this point, with no trace of the mental, we can either reject non-reductive physicalism or be convinced that something went wrong along the exclusion reasoning. But I think it would be unjustified to accept the irrealist consequence with no fight. On the one hand we can resist any sort of causal criterion for property reality altogether. The force of criterion (CPC) relies mostly on epistemological arguments (Shoemaker 1980) that are not conclusive. On the other hand, the strongest of those arguments do not rule out the reality of causal effects that are not themselves efficacious (Armstrong 1978) (i.e. properties with "passive" causal powers) and thus they only favor a criterion such as:

> (CRC) A property is real only if it contributes to the *active or passive* causal powers of the object that has it.

And this criterion, which counts epiphenomenal properties as real properties, would be enough to make mental realism and causal inefficacy compatible.[32] Consequently, it would be a mistake to assume that exclusion entails mental irrealism, and it would be a mistake to base our reaction to the exclusion argument on our reaction to mental irrealism. A detailed discussion of the metaphysics of epiphenomenal properties is lacking in the literature and thus I don't think we have to accept the *inevitability* of mental irrealism for epiphenomenalism.

3.2. Explanatory Irrelevance

A second damaging consequence of the exclusion problem is the potential irrelevance of psychological explanations. In section 1 I concluded that explanatory exclusion is not the hard problem. However, once we establish causal inefficacy, there are at least two ways of linking this result with explanatory irrelevance without appealing to any sort of explanatory exclusion principle. One argument is quite straightforward. All we have to add is the thesis that all explanation is a causal explanation. Or, in other words, that we can explain an event or a property only if we show the cause (or at least a cause) of that event or property. The most visible form of this view defends that every explanation has to "track" an *objective* relation between the events described by the explanans and the explanandum, and this objective relation has to be *causation.*[33] If the mental is causally inert, mental terminology cannot explain.

There is another argument from causal inefficacy to explanatory irrelevance. Until 30 years ago, many philosophers defended the view that rationalizing or reason-giving explanations are not causal explanations. But a highly influential paper by Donald Davidson (Davidson 1963) reversed this orthodoxy: he convincingly argued that the only way a rationalizing explanation can *explain* an action is if it invokes the action's cause. There are many cases in which we can have competing rationalizations for the same action, so if we do not include a causal element:

> (...) something essential has certainly been left out, for a person may have a reason for an action and perform that action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that the agent performed the action because he had the reason (Davidson 1963, p. 9).

If Davidson is right, we cannot explain an action if there isn't a causal relation between our mental states and it. The apparently inescapable conclusion for our purposes is that if the mental is causally impotent then intentional or rationalizing explanations are in fact irrelevant.

No matter which of the two routes we follow, mental inefficacy seems to yield explanatory irrelevance. The question here is, of course, not whether we *use* psychological explanations; it is rather whether these explanations can be *justified* by grounding them in *bona fide* objective relations. Here is a representative (yet hyperbolic) opinion about that result:

> (...) if commonsense intentional psychology really were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species; if we're that wrong about the mind, then that's the wrongest we've ever been about anything (Fodor 1987, p. xii).

This sort of reaction, again, can push us to reject non-reductive physicalism or be convinced that the exclusion reasoning has to be mistaken.

There is, however, a plausible answer to both arguments. The first one assumes that every explanation has to "track" a causal relation. That view is far too narrow. It would even rule out explanations of supervenient properties in terms of their supervenient bases. A pluralism that allows explanations to track different dependence relations seems a more natural option.[34] However, it would be too quick to claim that every dependence relation or "path" yields a *bona fide* explanation. In our particular case, the problem seems to be that the "arrow of dependence" that goes from a supervenient mental state to its base (and then to the effect of the base, i.e., the explanandum) has the wrong direction. Should this bother us? Consider these cases involving dependence relations. A tornado causes both extensive damage to crops and the loss of many lives. Yet we don't explain the crop damage in terms of the deaths (or vice versa). The smoothness of Claudia's skin supervenes on the molecular structure of her skin. Yet we don't explain the molecular structure of the skin in terms of its smoothness. Many sorts of dependence relations can be explanatory. Nonetheless, the *converses* of dependence relations are typically non-explanatory. If the epiphenomenalist wants to keep psychological explanations, she has to show *why* the converse of supervenience can explain an effect of a supervenience base while it cannot explain that base. And they have to show *why* the converse of supervenience can be explanatory while the converse of other dependence relations (causation, mere conceptual dependence) cannot explain.[35] But even if work remains to be done to articulate the explanatoriness of supervenient properties, it would be premature to rule out *bona fide* psychological explanations given the causal inefficacy of mental properties.

There is also ammunition against the Davidsonian argument. True, the only reason that can explain the action is the reason *why* the individual did it. And since "cause" is what is behind "because", the only apparent way a rationalizing explanation can explain an action is if it invokes the action's cause. If exclusion

is right, this has to be a neurophysiological cause of the behavior. But this neurophysiological cause necessitates (being its supervenience base) the intentional states that are cited in the rationalizing explanation. Of course, that neurophysiological cause does not necessitate *other* potential reasons the agent might have for performing the action. So within the model we are singling out the reason *why* the agent acted as he did and discarding the other reasons *for* the action that are not related to the actual cause of the action. The fact that we are (or may be) ignorant of what is going on at the physiological level should not bother us, for the intentional states give us enough ground for believing that an appropriate neurophysiological property is doing the causal job.[36]

We have to conclude again that it would be a mistake to assume that exclusion entails the irrelevance of psychological explanations, and it would be a mistake to base our reaction to the exclusion argument on our convictions about the indispensable character of intentional explanations. Recapitulating, then, what the Trojan horse brings, given exclusion, is destruction from the causal efficacy of the mental. But we cannot take for granted that it brings mental irrealism or the irrelevance of psychological explanations.

3.3. The Inefficacy of the Functional and Generalization

Another problem for mental causation that has received attention in recent years is the so-called problem of functional properties.[37] I shall argue that the problem of the inefficacy of the functional is just a version of the problem of exclusion. Simply, it is the problem in which the properties whose causal efficacy is preempted are functional properties; but the reasoning that leads to the preemption is exactly the same. This is in general very easy to see: functionalists should accept all the claims I presented as essential to a non-reductive physicalist and in particular the dependence and closure theses. Functionalism is committed to a particular understanding of the dependence relation, the relation of realization, but insofar as it accepts the physicalist commitments it cannot avoid the preemption problem. I shall try to show in some detail that what is perhaps the most carefully presented version of the problem of functional properties does not differ from our exclusion problem.

Jackson and Pettit require, as we did, that in order for a property to be causally efficacious with regard to an effect it be the property in virtue of whose instantiation, at least in part, the effect occurs. And they say that a property F is not causally efficacious in the production of an effect e if the following three conditions are fulfilled together (cf. 1990, p. 108):

(i) there is a property G different from F such that F is efficacious in the production of e only if G is efficacious in the production of e.

(ii) the F-instance does not help to produce the G-instance: they are not sequential causal factors of e.
(iii) the F-instance does not combine with the G-instance to produce e: they are not coordinate causal factors.

The moral is that all functional properties conform to the role of F in (i)-(iii), where G is the realizing or structural property. And therefore all functional properties, and *a fortiori* all mental properties, are causally inefficacious. Let's see how this works in an example.[38]

I move my hand towards a glass of water. Why? First answer: because of my desire for a drink of water. Second answer: because of the particular neural configuration that preceded the movement of my hand. Jackson and Pettit say that the property of desiring water was efficacious in producing my movement only if my particular neural configuration was efficacious: hence (i). It is clear that in order to obtain (i), part of what Jackson and Pettit need is the claim that mental properties such as my desire of drinking water (and *mutatis mutandi* other functional properties) are dependent on more basic properties such as my neural configuration. They are not explicit about this, but the reason is that given the dependence relation, my desire would be instantiated *only if* a neural property was instantiated. (They also need, of course, to affirm that a neural property has to be causally efficacious for my movement, and this would involve a closure claim).

Furthermore they say that the property of desiring water cannot help to produce my particular neural configuration: hence (ii). It is interesting to see the reason they give for this. Regarding an analogous example also involving a functional property they argue:

> (...) the fragility did not help to produce the molecular structure in the way in which the structure, if it was efficacious, helped to produce the breaking. There was no time-lag between the exercise of the efficacy, if it was efficacious, by the disposition and the exercise of the efficacy, if it was efficacious, by the structure (Jackson and Pettit 1990, p. 109).

So the reason is that the instantiation of both properties is simultaneous (since it is a case of properties involved in realizing or dependence relations), while causation requires a "time-lag". They are, thus, presupposing that supervenience or realization is a non-causal relation, a presupposition that played an important role when we discarded the alternatives to overdetermination in section 2.

Finally, my desire and my neural configuration cannot be coordinate causal factors since "full information about [the neural configuration] and the laws would enable us to predict [the effect] e; [my desire] would not need to be taken into account as a coordinate factor: hence (iii)." To obtain (iii) Jackson

and Pettit need a closure principle for they assume that the physical level must provide a complete account of my movement, which is a physical event.[39] The conclusion is that since my desire to drink water complies with (i)-(iii), it is causally inefficacious.

When Jackson and Pettit discuss the three cases, they use the very same assumptions we used in the formulation of the exclusion problem. In fact, the main difference between their formulation and ours is that they do not offer an argument to rule out overdetermination. Although they have the ingredients to offer such an argument they just discard such an option as unacceptable. They say:

> (...) on any account of efficacy, it is Pickwickean to describe the F-property as efficacious, given that any efficacy it is alleged to have exercised would have been screened off by the influence of the G-property. No conception of efficacy, no matter how debunking, should allow that efficacy can be exercised across such a screen (Jackson and Pettit 1990, p., 110-1).

There is also a minor difference between Jackson and Pettit's formulation and our formulation: the former only covers mental-to-physical causation. This is not surprising since most presentations of the exclusion problem (beginning with Malcolm's seminal one) also concentrate in this case. But it should be clear that Jackson and Pettit have all the elements that would allow them to formulate all the cases and to declare the mental inefficacious in full generality.

There might be, however, another apparent difference between the exclusion problem and the problem of the functional. The latter involves the causal inefficacy of macro properties, biological properties and in general every property functionally or dispositionally characterized (the so-called "second-order" properties), for insofar as these properties are realized in multiple bases or different structural properties, the causal role is in charge of such bases. So isn't it that the problem of the functional is broader in scope than the exclusion problem? The natural, yet sometimes unnoticed, answer is "no".[40] The general ontological view of non-reductive physicalism also considers macrophysical, biological and chemical properties as supervenient properties. This view implies that the exclusion problem reappears for each layer depending on the basic physical layer, for in each case the causally closed physical realm will be in charge of the entire causal responsibility. The only restriction would be regarding those non-basic properties which are susceptible of being type-identified (if this were possible) with the basic properties (for they would be as efficacious as the basic ones). But this same restriction would apply to those functional properties which could be type-identified (if this were possible)

with a structural property.[41] So there is no difference in scope between the two problems. We can conclude that what has been called the problem of the causal efficacy of the functional is just a version (the version that arises for the functionalist kind of non-reductive physicalism) of the exclusion problem.

Our previous discussion is appropriate to clarify the following issue. Isn't the fact that the exclusion problem generalizes over every non-basic property, leaving the biological and the chemical causally inefficacious too, proof that we should either reject non-reductive physicalism or challenge the exclusion reasoning?[42] I don't think that the generalization makes things worse for non-reductive physicalism. Mental inefficacy is bad enough. Perhaps the strongest causal intuitions we have are the ones associated with agency and deliberation, and the original exclusion argument already damaged them. If we are ready to accept epiphenomenalism at that level, accepting it at the other levels does not seem to add much to the problem. Note, also, that the two strategies suggested above for the case of the mental can be applied to every supervenient property, so that those properties can be considered genuine and explanatorily relevant.

4. What to do with Exclusion? A Map of Answers

Once we have determined that the problem of exclusion is primarily ontological, can be generated by the constitutive claims of non-reductive physicalism, and can be separated from the issues of mental realism and psychological explanations, which are the alternatives we have? I shall briefly explore a map of available options. While I shall be pointing out briefly some problems for each of the options, there will be no attempt to argue for any particular view.

Since the problem of exclusion apparently shows that the Trojan horse of non-reductive physicalism entails that the mental is causally inefficacious, there are two incompatibilist reactions: we can reject non-reductive physicalism (to keep, I assume, mental causation) or deny mental causation (to keep, I assume, non-reductive physicalism). There is of course a compatibilist strategy attempting to show that the exclusion problem is the result of some kind of misunderstanding of the basic notions involved and thus that we can keep both non-reductive physicalism and mental causation.

4.1. Anti-Orthodox Incompatibilism

Non-reductive physicalism was characterized in section 2 as a conjunction of substance monism, property dualism (including, of course, mental realism), and the basicness of the physical expressed in terms of supervenience and the

closure of the physical. The exclusion or preemption problem, as we have seen, emerges from these very assumptions. So the first reaction is to close the gates and keep the Trojan horse of non-reductive physicalism out. This can be done in at least the following three ways.

The first way is to return to *Cartesian Dualism* by denying substance monism. This option not only defies the physicalist orthodoxy, but also faces a related problem when dealing with soul-body interaction.

The second is to reject the primacy of physical properties over mental properties without challenging substance monism. This would give us some sort of *strong property dualism*, a position that, as far as I know, has not had major defenders in the recent literature on mental causation.[43] The reasons seem clear: First, the relation between the mental and the physical becomes a complete mystery. Second, what is the advantage of rejecting immaterial, non-physical substances if we think that mental properties are absolutely autonomous with respect to physical properties? What could be the rationale of defending that mental properties are instantiated in *physical objects* if they don't depend in any sense on *their physical properties*? It seems that if we are to protect the intuition that there are no extraphysical or supernatural entities we have to keep, as a minimum, some sort of dependence thesis. And here emergentism enters the scene.

Emergentism, as we will see below, can be interpreted as a compatibilist approach. But it could also be argued that it is a special case of this second way of leaving the horse out -perhaps its most plausible version. Emergentism is committed to some kind of dependence thesis as a way of connecting the emergent and the base properties of different strata: emergent properties have to depend on those properties from which they emerge. However, emergentism makes mental-to-physical (or downward) causation one of its principal claims. If this is so, emergentism rejects only the causal closure of the physical. And by keeping the dependence thesis it may avoid the problem that a stronger property dualist faces: the threat of collapsing into Cartesian dualism.

The third view within this group is to deny mental realism. *Mental irrealism* may take one of two major forms: a milder form claiming that we have to keep mental expressions for pragmatic reasons (retentive irrealism) or a stronger form recommending the elimination of mental terminology (eliminativism). Their common denominator is that property dualism is false since mental properties are not real or genuine properties. Type identity theories claiming that mental properties are "nothing over and above" physical properties may be included as retentive irrealism. Kim's recent "second-order prop-

erty" account is clearly a version of retentive irrealism too, one which in fact is largely motivated by the exclusion problem.[44] All these positions dissolve the exclusion problem since there is no mental causation to worry about. But then we suffer the destruction to all science (of the mental) and common sense without the benefits of the horse.[45]

4.2. Epiphenomenalist Incompatibilism

Epiphenomenalism gives up mental causation but keeps the three claims defining non-reductive physicalism. The Trojan horse is welcome in spite of the devastating consequences. It has been said that epiphenomenalism is simply unacceptable since it "solves" the problem of exclusion (or, for that matter, any problem for mental causation) by giving up one of our most entrenched intuitions: mental causation. It is not surprising then that for many even considering the view has only a propaedeutic value: it just shows us the urgent need for an account of mental causation.

The widespread resistance to give up mental causation is clear by the fact that not even the two most notorious recent defenses of the position presented themselves as epiphenomenal views: Kim's supervenient causation and Jackson and Pettit program explanation models were originally sold as slightly deflated vindications of mental causation and thus there could be a temptation to (mis)classify them as compatibilist approaches.

I do think, however, that epiphenomenalism deserves to be explored in more detail. In order for the theory to be plausible, the two pressing issues are exactly the two consequences discussed in section 3: the reality and the explanatoriness of epiphenomenal properties. While the first issue remains unexplored,[46] some progress has been made regarding the second one.[47] If we succeed, the view will add to ontological incompatibilism mental realism and a compatibilist stance for the explanatory (epistemological) level.

4.3. Compatibilism

The last type of option is the one most of us would like to be true: *Compatibilism,* in our context, is the claim that something went wrong along the exclusion reasoning and that we can keep both some sort of non-reductive physicalism and mental causation. Despite the prophets of doom, the Trojan horse is not that dangerous after all. This type of strategy needs to explain *what* went wrong and this is usually achieved by claiming that one of the two litigants (mental causation or non-reductive physicalism) has not been understood prop-

erly in setting the exclusion problem. I shall mention here three of the approaches that follow this route, with no intention of being exhaustive.

Emergentism is an obvious candidate, since it would block exclusion by claiming that physicalism needs to be reinterpreted. As we have seen, this means that while some sort of supervenience or dependence is needed, the causal closure of the physical realm should be rejected. Now, the two questions that, in the context of our inquiry, an emergentist has to face are these: is the *partial* abandonment of the basicness of the physical enough to avoid the conclusions of the exclusion argument? And, can a theory that gives up the causal closure of the physical be considered physicalist in any significant sense? I think the answer to the both questions should be a qualified "no". Regarding the first question, recall that in the case of mental-to-mental causation (the possibility of M causing M'), the problem is that the causal role of M with respect to M' is preempted by P', the dependence base of M'. So the closure thesis plays no role when we say that the exclusion problem forbids mental-to-mental causation: the work is done by the dependence thesis, a thesis that emergentism has to accept.

An instructive way of thinking about the second question is to discuss what an emergentist would say about the competition between the relations M-P' (downward causation) and P-P' ("horizontal" causation). Since there is no closure thesis, the emergentist is under no pressure to deny M-P'. So would she say that P' is overdetermined by the two causes, M and P? Overdetermination doesn't violate the emergentist's physicalism since she doesn't defend causal closure. But we still have the "elegance and plausibility" argument: if the emergentist wants to avoid a world in which the instantiation of almost every property[48] has two complete sources she has to deny the overdetermination answer. Now, would she say that M and P are partial causes? Since, again, closure is not a factor, we cannot say that M's being a partial cause violates (the emergentist's) physicalism. However, given the emergentist's adherence to the dependence thesis according to which P guarantees the presence of M, how plausible is it to consider M an independent partial cause of P'? So the only remaining possibility seems to be that P has no causal role in the production of P', and in general, that physical properties have no causal role in the production of those properties of which emergent properties are thought to be causes. This is much stronger result than denying the causal closure of the physical; it amounts to denying *any* causal role for the physical with respect to an important group of other physical properties. So the emergentist's debilitation of physicalism goes beyond the denial of closure. And reinterpreting physicalism

now begins to look similar to denying it, which would be to go back to some sort of strong dualism

The next attempt is to reinterpret mental causation by pointing out that a defective formulation of causation has been used to build the exclusion problem. A way of doing this has been called the *"explanatory primacy" approach.*[49] The mental, according to this view, is causally relevant or efficacious insofar as it figures in successful explanations. Baker says:

> If we put aside the metaphysical picture and begin with explanations that work, causation becomes an explanatory concept. This presents a sharp contrast to the metaphysical picture, which subordinates explanation to causation, where causation, in turn, is conceived as an "objective" relation in nature. (...) if we reverse the priority of explanation and causation that is favoured by the metaphysician, the problem of mental causation just melts away (1993, p. 93).

But it seems clear that once we "reversed the priority", our view will plausibly end up as an irrealist approach about the mind, possibly of a retentive type. Most realists would maintain that within a realist framework the possibility of mental causation is a precondition of the possibility of *causal* psychological explanations. Given the "explanation primacy" approach, mental predicates may well be mere explicative and predictive tools; mental reality doesn't seem to be among the theses to be defended unless we are ready support a claim like this: "A property is real if and only if it figures in successful explanations". But we can plausibly assume that this is not the kind of view on property reality that the mental realist wants. So this supposedly compatibilist strategy might be just an incompatibilist stance in which the victim is mental realism.

The final option consists in claiming that if we reinterpret mental causation in terms of a *counterfactual account* of causation we can avoid the exclusion problem.[50] This option is perhaps the one that has received more attention and thus the one that exhibits more versions. A common problem that all counterfactual versions seem to face (and some attempt to solve) is that the counterfactual test is a poor test to assess causal (and in general dependence) directionality.[51] For suppose we have a property P instantiated at t which causes property Q at t' and (through a rather different causal path) property R at t''. If conformity with counterfactuals is to be accepted as a sufficient condition for causation, we may have to accept that R causes Q. And of course we don't want to explain a property by citing a (possibly temporally posterior) clearly unrelated property. An example by Segal and Sober (1991, p. 5) makes this clear. Suppose that a red piece of coal causes a piece of tissue to smolder. The reason the coal is red is that it is hot. If the coal had not been red it would not have

been hot and if it had not been hot the tissue would not have smoldered. But we don't want to claim that the redness of the coal causes (or explains) the tissue's smoldering.

Compatibilist strategies in philosophy are typically more rewarding but they sometimes fail to address the hard problems. In our case, given how unpalatable the incompatibilist options are, our dissatisfaction with the compatibilist strategies can perhaps coexist with the hope that one of them has to be true. In any case, if we break the walls and give admittance to the horse we rather avoid celebrations and keep its belly under close scrutiny.[52]

Notes

1 Reid (1764, p. 132b). Reid, as it is known, felt compelled to block the entrance of the horse and defended direct realism.

2 I am primarily thinking of functionalism (Putnam 1967), Fodor (1974), Davidson's anomalism (Davidson 1970, 1974) and Kim's early work on supervenience (1978).

3 For the first problem, cf. Stich (1983), Fodor (1987), Kim (1982); for the second, cf. Honderich (1982), Sosa (1984), Heil & Mele (1993) part I.

4 Compare with the often quoted passage by Fodor: "If it isn't literally true that my wanting is causally responsible for my scratching, (...), and my believing causally responsible for my saying, (...), if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world" (1990, p. 156)

5 For instance, Chalmers' (1996) distinction between "easy" problems and the "hard" problem and Tye's (1995) various distinctions.

6 "Physical" is here understood in a broad sense, including what is strictly speaking physical plus what is chemical and biological.

7 An incomplete list of works presenting the exclusion problem in one fashion or the other includes: Malcolm (1968), Kim (1979), (1988), (1989a), (1993b), (1998), Yablo (1992), Baker (1993), Crane (1995), Sabatés (1997), Corbí & Prades (2000).

8 (EE) is not the only possible explanatory claim, but it is probably the standard one. An alternative, non-equivalent explanatory principle is sometimes assumed: If a phenomenon is completely explained in terms of theory X, and theory Y is independent from theory X, theory Y cannot explain it. The alternative claim does not forbid, *prima facie*, that there can be two complete and independent explanations provided they are within the same theory, but seems strong enough to generate the explanatory irrelevance of the mental.

9 The issue of whether the exclusion problem has strong connections with the debate surrounding Cartesian interactionism is beyond the scope of this paper. However, it is reasonable to claim that neither the problem that appeals to the "different essences/no common measure" of the soul and the body nor the problem that appeals to the "conservation of motion" are strictly speaking exclusion problems, but general problems of impossibility of the immaterial soul to be efficacious regarding the physical.

10 Cf. Malcolm (1968). It is somewhat paradoxical that the first formulation of the problem which is now seen by many physicalists as the most serious challenge for psychological explanations and for mental causation itself was devised as a way of challenging the idea that human behavior can be explained in physical terms. In Malcolm's terms: "there is a respect

in which mechanism is not conceivable. This is a consequence of the fact that mechanism is incompatible with the existence of any intentional behaviour." (p. 145).

11 The argument can be more carefully reconstructed as follows:

1) Neurophysiological states provide sufficient causal conditions for all bodily movements.

2a) For something to be causally efficacious for an event E, it should have counterfactual force with respect to E.

2b) If there is a sufficient cause for E (say C), nothing different from C can have counterfactual force with respect to an event E.

Therefore, 2) if something is causally sufficient for an event E, nothing else can be causally efficacious for E.

3) Intentions and purposes are not neurophysiological states.

Therefore, 4) Intentions and purposes cannot be causally relevant for any bodily movement.

(A hint of a most important argument against overdetermination is already present in this argument.) For other (slightly different) reconstructions of this argument, see Kim (1989a) and Goldman (1970).

12 Since it is widely acknowledged that the problem of mental causation is the problem of the causal efficacy of mental *properties* I will using property terminology most of the time. State terminology, however, is used in some contexts.

13 I shall not attempt an exegesis of Kim's different versions of the exclusion principle, but there seems to be a reason for such fluctuations: he considers that both the explanatory and the causal principle are roughly equivalent, probably the epistemological and ontological sides of the same coin (this is why he used several times the expression "causal-explanatory exclusion" and also why, when he is using the explanatory principle, he refers in general to causal explanations). Thus, in contexts in which his main worries are related to causation (cf. 1989b, 1991a) he uses the causal formulations, and in contexts in which he deals with explanatory issues (1988, 1989a) he prefers the explanatory one (indeed, in (1989a) to be discussed here, the choice for a explanatory version may be just prompted by the fact that he is re-presenting the debate opened by Malcolm's paper). Moreover, in his latest works Kim began to favor the causal formulation (1993b, 1998).

14 This is particularly so in Kim (1988). See discussion on generalization in section 3.

15 It must be said, however, that Kim hints, in Kim (1989a), at an argument for (EE) which does not rely on causal considerations; I shall consider it below.

16 1989a, p. 250. Cf. also Kim (1988, p. 233).

17 Cf. Kim (1989a, pp. 250-4) and (1988, pp. 233-5). In the next section, I shall use this same strategy to defend a more general version of exclusion.

18 I am assuming, of course, some minimal realism about causation. Strawson says: "We sometimes presume (...) that causality is a natural relation which holds in the natural world (...) just as the relation of temporal succession does or that of spatial proximity. We also, and rightly, associate causality with explanation. But if causality is a relation which holds in the natural world, explanation is a different matter. (...) it is not a natural relation [but] an intellectual or rational or intensional relation." (1985, p. 115).

19 Variations on this point have been made by Horgan (1997) and Burge (1993), among several others. I agree with them that this may provide grounds to block the epistemological argument for (EE), but I am hesitant about its effectiveness against the causal/metaphysical problem.

20 Is mental realism compatible with some version of type-type identity theory? I don't think so, but some "conservative" as opposed to "eliminative" identity theories could be consid-

ered by some as mental realist views. The answer to the question depends on whether strong ontological reduction of the sort type identity provides can leave room for realism about the reduced level.

21 Cf. Kim (1987, 1989b) for arguments showing that no weaker version of dependency or supervenience (i.e., weak and global supervenience) serves the purposes of physicalism. Another intriguing option, a notion of probabilistic supervenience, remains largely unexplored (but see Glymour, Sabatés & Wayne (forthcoming)).

22 This is a relatively weak formulation of closure. Some use a stronger formulation: no instantiation of a physical property has non-physical properties in its generating causal chain. This stronger version would automatically exclude emergentism and perhaps other views from the physicalist family I am characterizing, and would rule out mental-to-physical causation without the need of further argument. But the weaker, more inclusive formulation is enough to generate the exclusion problem.

23 I will simplify the presentation by eliminating in most contexts expressions like "the instantiation of" and "has the potentiality to".

24 For a detailed and complete discussion of these alternatives see Kim (1988, p. 233-5).

25 Kim's most recent views (1998) are an example that identity is, at most, compatible with a retentive approach regarding mental vocabulary, but incompatible with mental realism.

26 Moreover, this would violate again (WC). For P itself, as a physical property, requires a complete causal chain composed of physical properties, a causal chain whose penultimate link will plausibly be a physical property P-.

27 We could appeal to majority and authority regarding this issue. With the exception of John Searle (1992), philosophers of mind defend (or assume) that a relation of synchronic dependence such as the one between P and M is a non-causal relation. Some among countless examples include Segal & Sober (1991), and Jackson and Pettit (1990). The latter explicitly denies that a supervenience relation can be considered causal and this denial plays an important role in their formulation of the problem of functional properties (cf. their 1990, p. 109, and Kim 1998).

28 There are, however, hints at the need of a more general principle in Yablo (1992) and Kim (1993b).

29 This might be a diagnosis of why Yablo's (1992) analogy including determinates and determinables does not in fact dissipate the exclusionary worries.

30 Kim (1998), Chalmers (1995) and Jackson (1995) are examples of this. The modal force in question here is that of the second "necessarily" in (SS).

31 (CPC) is explicitly adopted, among others, by Shoemaker (1980), Kim (1993b) and Fodor (1987). This criterion is sometimes formulated as a biconditional, but it is not relevant for my purposes whether causal potency is a sufficient condition for reality.

32 Even (CRC) is perhaps too strong. Suppose that a property B is causally isolated but non-causally depends or supervenes on a causally efficacious property A. Shouldn't we consider it part of a net of real relations and thus a genuine property while avoiding more pragmatic approaches to what counts as a genuine property? (I discuss this in Sabatés (unpublished)).

33 Cf. Lewis (1986) and Salmon (1984) (Not all explanatory causalist is an explanatory realist, though. For that to be the case, realism about causation has to be added).

34 See Ruben (1990) and Kim (1994) for a defense of this view.

35 I discuss these issues in Sabatés (1999) and, particularly, Sabatés (1997).

36 It seems clear that the problem cannot be that we cannot be giving an explanation unless we are sure of which is the actual physiological cause. A bad intentional explanation would

not point to a dependence path that includes the cause of the action. But, in the same way, a bad physical explanation would not point to a causal path that produces the effect.

37 Cf., for instance, Jackson and Pettit (1990) and Block (1990). I shall focus mainly on Jackson and Pettit's formulation. For a more extended comparison between exclusion and the problem of functional properties see Kim (1998) and Sabatés (1997).

38 I am adapting here Jackson and Pettit's first example (about fragility, p. 109) to a case involving mental properties.

39 Jackson and Pettit (1990, p. 109). Jackson and Pettit are not completely consistent in keeping their discussion in causal terms: they use explanatory expressions in an argument that is supposed to be ontological. But this does not conceal the causal nature of their argument.

40 Kim, Horgan and Yablo are exceptions to this: they are aware that the exclusion or preemption problem reaches every property that depends on the physical ones. Jackson and Pettit, on the other hand, are explicit about the problem of the functional being general in this sense.

41 Does it make sense to say that some functional properties might be "singly realizable"? Perhaps not, but it is not clear that this makes sense for supervenient properties belonging to non-basic layers either.

42 Burge (1993), for instance, takes generalization to be a *reductio* of exclusion.

42 But see Hasker (1999).

44 Cf. Kim (1998), and Sabatés (forthcoming) for an evaluation of how Kim's new view fares vis-à-vis the exclusion problem. See also Pineda (1998) for another recent defense of reductive physicalism in connection with the exclusion problem.

45 The reductive physicalist can point out that since the science of the mental is lost anyway, there is no point in keeping the horse.

46 I elaborate on this in Sabatés (unpublished). Notice that if this first issue cannot be solved, epiphenomenalism would be just a version of mental irrealism.

47 Cf. Jackson & Pettit (1990) and Sabatés (1997). Conee (1995) and Bieri (1992) are other examples of epiphenomenalism taken seriously.

48 Plausibly emergentism is a general thesis about non-basic properties, not just about mental properties.

49 Cf., for instance, Baker (1993), Burge (1993) and Van Gulick (1993). Horgan (1997) seems to be close to this view. Some of these views combine explanatory primacy with a counterfactual approach.

50 Cf., for instance, LePore & Lower (1987), Horgan (1989, 1997), Yablo (1992) and Corbí & Prades (2000) (although some of these attempts (Horgan (1997), Corbí & Prades (2000) can also be understood as a denial of non-reductive physicalism).

51 Cf., for instance, Kim (1995, 1998) and Segal and Sober (1991) for this objection in the context of mental causation. For a general discussion see Horwich (1987).

52 I am indebted to Josep Corbí, Terry Horgan, Jaegwon Kim, Diana Pérez, Loretta Torrago and audiences at the University of Utah and SADAF for comments and suggestions.

BIBLIOGRAPHY

Armstrong, D.: 1978, *A Theory of Universals*, Cambridge, Cambridge University Press.
Baker, L.R.: 1993, 'Metaphysics and Mental Causation', in J. Heil & A. Mele (1993).
Bieri, P.: 1992, 'Trying Out Epiphenomenalism', *Erkenntnis* 36.
Block, N.: 1990, 'Can the Mind Change the World?', in G. Boolos (ed.): *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge, Cambridge University Press.

Burge, T.: 1993, 'Mind-Body Causation and Explanatory Practice', in J. Heil & A. Mele (1993).
Chalmers, D.: 1996, *The Conscious Mind*, Oxford, Oxford University Press.
Conee, E.: 1995, 'Supervenience and Intentionality', in E. Savellos & U. Yalcin (eds.): *Supervenience. New Essays*, Cambridge, Cambridge University Press.
Crane, T.: 1995, 'Mental Causation', *Proceedings of the Aristotelian Society*, sup. vol. 69.
Corbí, J. & Prades, J.L.: 2000, *Minds, Causes and Mechanisms*, Oxford, Blackwell.
Davidson, D.: 1963, 'Actions, Reasons, and Causes', reprinted in *Essays on Actions and Events*, Oxford, Oxford University Press, 1980 (references to the reprinted version).
Davidson, D.: 1970, 'Mental Events', in *Essays on Actions and Events*, Oxford, Oxford University Press, 1980.
Davidson, D.: 1974, 'Psychology as Philosophy', in *Essays on Actions and Events*, Oxford, Oxford University Press, 1980.
Fodor, J.: 1974, 'Special Sciences, or the Disunity of Science as a Working Hypothesis', *Synthese* 28.
Fodor, J.: 1987, *Psychosemantics*, Cambridge, MIT Press.
Glymour, B., Sabatés, M. & Wayne, A.: forthcoming, 'Quantum Java: The Upwards Percolation of Quantum Indeterminacy', *Philosophical Studies*.
Goldman, A.: 1970, *A Theory of Human Action*, Englewwod Cliffs, Prentice Hall.
Hasker, W.: 1999, *The Emergent Self*, Ithaca, Cornell University Press.
Heil, J. & Mele, A. (eds.): 1993, *Mental Causation*, Oxford, Oxford University Press.
Honderich, T.: 1982, 'The Argument for Anomalous Monism', *Analysis* 42.
Horgan: 1989, 'Mental Quasation', *Philosophical Perspectives* 3.
Horgan, T.: 1997, 'Kim on Mental Causation and Causal Exclusion', *Philosophical Perspectives* 11.
Horwich, P.: 1987, *Asymmetries in Time*, Cambridge, MIT Press.
Jackson, F. & Pettit, P.: 1990, 'Program Explanation: A General Perspective', *Analysis* 50.
Jackson, F.: 1995, 'Essentialism, Mental Properties and Causation', *Proceedings of the Aristotelian Society* 69.
Kim, J.: 1978, 'Supervenience and Nomological Incommensurables', *American Philosophical Quarterly* 15.
Kim, J.: 1979, 'Causality, Identity and Supervenienceain the Mind-Body Problem', *Midwest Studies in Philosophy* 4.
Kim, J.: 1982, 'Psychophysical Supervenience', *Philosophical Studies* 41.
Kim, J: 1987, "Strong' and 'Global' Supervenience Revisited', *Philosophy and Phenomenological Research* 48.
Kim, J.: 1988, 'Explanatory Realism, Causal Realism and Explanatory Exclusion', *Midwest Studies in Philosophy* 12.
Kim, J.: 1989a, 'Mechanism, Purpose and Explanatory Exclusion', *Philosophical Perspectives* 3.
Kim, J.: 1989b, 'The Myth of Nonreductive Materialism', *Proceedings and Addresses of the American Philosophical Association* 63.
Kim, J.: 1993a, *Supervenience and Mind*, Cambridge, Cambridge University Press.
Kim, J.: 1993b, 'The Nonreductivist's Troubles with Mental Causation', in J. Kim (1993a).
Kim, J.: 1994, 'Explanatory Knowledge and Metaphysical Dependence', *Philosophical Issues* 5.
Kim, J.: 1995, *Philosophy of Mind*, Boulder, Westview.
Kim, J.: 1998, *Mind in a Physical World*, Cambridge, MIT Press.
Lewis, D.: 1986, 'Causal Explanation', in *Philosophical Papers* 2, Oxford, Oxford University Press.
LePore, E. and Loewer, B.: 1987, 'Mind Matters', *Journal of Philosophy* 86.
Malcolm, N.: 1968, 'The Conceivability of Mechanism', *Philosophical Review* 77.

Pineda, D.: 1998, *Problemas y perspectivas del Fisicismo*, Ph.D. Dissertation, University of Barcelona.
Putnam, H.: 1967, 'The Nature of Mental States', in *Mind Language and Reality*, Cambridge, Cambridge University Press.
Reid, T.: 1764, *An Inquiry into the Human Mind on the Principles of Common Sense*, page references to Hamilton (de.): 1863, *The Works of Thomas Reid*, Edinburgh, Maclachan & Steward.
Ruben, D.: 1990, *Explaining Explanation*, London, Routledge.
Sabatés, M.: 1997, 'Should a Cognitive Psychologist Worry About the Causal Inefficacy of the Mental?', in B. Niggemeyer (ed.): *The Cognitive Level*, Duisburg, LAUD.
Sabatés, M.: 1999, 'Consciousness, Emergence and Naturalism', *Teorema* 18.
Sabatés, M.: forthcoming, '*Mind* in a Physical World?', *Philosophy and Phenomenological Research*.
Sabatés, M.: unpublished, 'Epiphenomenalism, Isolationism, and the Causal Criteria for the Reality of Properties'.
Salmon, W.: 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton, Princeton University Press.
Schiffer, S.: 1987, *Remnants of Meaning*, Cambridge, MIT Press.
Searle, J.: 1992, *The Rediscovery of the Mind*, Cambridge, MIT Press.
Segal, G. & Sober, E.: 1991, 'The Causal Efficacy of Content', *Philosophical Studies* 63.
Shoemaker, S.: 1980, 'Causality and Properties', in S. Shoemaker: 1984, *Identity, Cause and Mind*, Cambridge, Cambridge University Press.
Sosa, E.: 1984, 'Mind-Body Interaction and Supervenient Causation', *Midwest Studies in Philosophy* 9.
Stich, S.: 1983, *From Folk Psychology to Cognitive Science*, Cambridge, MIT Press.
Strawson, P.: 1985, 'Causation and Explanation', in B. Vermazen & M. Hintikka (eds.): *Essays on Davidson*, Oxford, Oxford University Press.
Tye, M.: 1995, *Ten Problems of Consciousness*, Cambridge, MIT Press.
Van Gulick, R.: 1993, 'Who is in Charge Here? And Who is Doing All the Work?', in Heil and Mele (1993).
Yablo, S.: 1992, 'Mental Causation', *Philosophical Review* 101.

Marcelo H. Sabatés is Licenciado in Philosophy from the University of Buenos Aires (1988) and holds a Ph.D. from Brown University (1996). He is currently an assistant professor at Kansas State University. He has published articles in the areas of philosophy of mind, metaphysics and epistemology in several journals (*Philosophical Studies*, *Philosophical Issues*, *Análisis Filosófico*, *Teorema*, *Philosophy and Phenomenological Research*, etc.) and books. His principal topics of research are mental causation and the nature of depedence relations.