Non-Reductive Physicalism and the Limits of the Exclusion Principle*

Christian List (London School of Economics) and Peter Menzies (Macquarie University)

(forthcoming in the *Journal of Philosophy*)

ABSTRACT

It is often argued that higher-level special-science properties cannot be causally efficacious since the lower-level physical properties on which they supervene are doing all the causal work. This claim is usually derived from an exclusion principle stating that if a higher-level property *F* supervenes on a physical property *F** that is causally sufficient for a property *G*, then *F* cannot cause *G*. We employ an account of causation as difference-making to show that the truth or falsity of this principle is a contingent matter and derive necessary and sufficient conditions under which a version of it holds. We argue that one important instance of the principle, far from undermining non-reductive physicalism, actually supports the causal autonomy of certain higher-level properties.

Non-reductive physicalism is very popular in the philosophy of the special sciences. It consists of three theses. First, the properties studied in the special sciences are not identical to physical properties, since they are multiply realized by them. Secondly, special-science properties nevertheless supervene on physical properties in the sense that there cannot be a difference with respect to these properties without a further difference with respect to physical properties. Thirdly, these higher-level properties are causes and effects of other properties. Like many others, we consider this an attractive package of theses.

But several philosophers, most notably Jaegwon Kim, have argued that non-reductive physicalism is untenable since its first two theses contradict the third.[1] Focusing on the example of how mental properties relate to their underlying physical, neural properties, Kim claims that if mental properties supervene on neural properties without being identical to them, then mental properties cannot be causes of other properties. His argument invokes what he calls the *exclusion principle*: if a property *F* is causally sufficient for some effect *G*, then no distinct property *F** that supervenes on *F* can be a cause of the effect *G*.[2] To sketch the argument, suppose, for a contradiction, that a mental property *M* causes a physical behavioural property *B*. Assuming the causal closure of the physical world, there must exist a physical property *P* that is causally sufficient for the behavioural property *B*. But *P* is plausibly the physical property on which the mental property *M* supervenes. Hence, by the exclusion principle, *P* excludes *M* as cause of *B*, a contradiction.[3]

This argument against non-reductive physicalism is thought to be so forceful because the exclusion principle is taken to be an analytic truth.[4] Even its critics generally suppose that the truth or falsity of the principle can be established *a priori*. In this paper, we challenge this supposition and reach two surprising conclusions. First, we show that the truth or falsity of the principle is in fact a contingent matter, and derive necessary and sufficient conditions for its truth. Second, we establish that, when the principle is true, it can actually support, rather than undermine, the causal autonomy of special-science properties.

Our argument proceeds as follows. In section I, we introduce the exclusion principle, following Kim's original formulation, and illustrate its implications. Although both proponents and critics of the principle usually assume that its truth or falsity can be settled by an investigation of the concept of causation, the discussion seldom employs a well-grounded theory of causation. In section II, we therefore introduce the conception of causation as difference-making. Since a conception of this kind is common to several different theories of causation – e.g., counterfactual, interventionist and contrastive ones – our use of it in investigating the exclusion principle should be congenial to a broad range of such theories. In section III, we show that when causation is understood in this way, Kim's formulation of the exclusion principle is false. In section IV, however, we consider an alternative and more plausible version of the principle not vulnerable to the counterexamples we present against Kim's version. It carefully distinguishes genuine difference-making causes from causally sufficient conditions by requiring that the same effect can never simultaneously have a lower-level difference-making cause and a higher-level one. But we find that, despite being immune to the earlier counterexamples, the truth or falsity of the new principle is still contingent on the causal system in question, and we identify the conditions under which it is violated. In section V, we turn to the conditions under which the principle holds. The principle can apply in two non-trivial ways. The first – the case of *upwards exclusion* – is familiar from the argument against non-reductive physicalism: here a lower-level cause excludes a higher-level one. But the second – the case of *downwards exclusion* – is often overlooked: here a higher-level cause excludes a lower-level one. For example, a mental property might cause a behavioural one while no underlying neural property does so too. Cases of downwards exclusion are particularly interesting, as they support the causal autonomy of higher-level properties. They occur whenever higher-level causal relations are what we call *realization-insensitive*: the presence or absence of some effect does not hinge on the actual

lower-level realization of the relevant cause, and the same effect would have been consistent with other lower-level realizations of the cause. In contrast, whenever there are *realization-sensitive* causal relations at the higher level – i.e., the actual realization of the cause matters – the exclusion principle is false.[5] In section VI, we draw some conclusions.

Our discussion has three restrictions, which mostly involve convenient simplifications rather than a loss of generality. First, we formulate simplified versions of the exclusion principle that do not mention overdetermination, in contrast to some versions formulated with appropriate exception clauses.[6] Second, we concentrate on deterministic causation, setting aside probabilistic generalizations,[7] and consider relatively simple causal systems to which the account of causation as difference-making applies most readily.[8] Third, we discuss causal relations involving properties. Causation is best understood, we believe, as a relation between variables. So causation involving properties is a special case in which the variables are binary. A more general treatment would handle causation involving many-valued variables. Throughout the paper, we follow Kim in focusing on the relationship between mental and neural properties, though our conclusions apply more generally.

I. THE EXCLUSION PRINCIPLE AND PROPORTIONAL CAUSATION

The most convenient formulation of the exclusion principle for our purposes in this:

> *Exclusion principle*: If a property *F* is causally sufficient for a property *G*, then no distinct property *F** that supervenes on *F* causes *G*.[9]

It is useful to give a real-life example. The example was first discussed by James Woodward.[10] It concerns the research by Richard Andersen and colleagues at Caltech on the neural encoding of intentions to act.[11] (The ultimate goal of Andersen's work is to develop neural prosthetics for paralysed subjects that decode their intentions to reach for specific targets from neural signals and use these to control external devices.) Andersen and his colleagues made recordings from individual neurons in the parietal reach region (PRR) of the motor cortex of monkeys. This region is known to encode intentions or higher-order plans to reach for specific targets, say a piece of fruit in a particular location.[12] Andersen developed a program that correlated the monkeys' intentions to reach for specific goals, as revealed in their movements, with certain patterns in the recorded firings of neurons in their PPR. Using neural recordings, the program

was able to predict with 67.5% accuracy the reaching behaviour of the monkeys towards eight targets.

The neural signals that encode the monkeys' intentions to reach for certain targets were recorded as averages of the firing rates (spikes per second) of individual neurons. But clearly the same aggregate firing rate in a group of neurons is consistent with a lot of variation in the behaviour of individual neurons. For example, very different temporal sequences of neural firings can give rise to the same firing rate. So an intention to reach for a certain target can be realized in many different ways at the level of individual neurons. Nonetheless, each intention is associated with a distinctive aggregate pattern of firing rates. It is useful to introduce some simple notation. Suppose that the monkeys can have intentions to reach for certain targets, I_1, I_2, I_3 etc., and can perform the corresponding actions A_1, A_2, A_3, etc. Suppose, further, that each intention I_i can be realized at the level of individual neurons in different token patterns of neural firing, N_{i1}, N_{i2}, N_{i3} etc. Suppose that on some specific occasion a monkey forms the intention I_1 to reach for a particular object and performs the corresponding action A_1. Suppose further that N_{11} is the particular token pattern of neural firing that realizes or encodes the intention I_1 on this occasion. The central question is: What was the cause of the monkey's action A_1? Was it the intention I_1, or its particular neural realization N_{11}? The exclusion principle dictates that the cause of the monkey's action is the neural realization N_{11}, not the intention I_1. The reason is that N_{11} is causally sufficient for the action A_1, and by hypothesis I_1 supervenes on N_{11}, so the principle excludes I_1 from being a cause, leaving N_{11} as the only possible cause.

The principle may appear plausible when applied to this example. But we think it is only superficially plausible. As several philosophers have noted, the neural state N_{11} does not satisfactorily fit the role of a cause because it is overly specific and involves extraneous detail.[13] In Stephen Yablo's terminology, it is not *proportional* or *commensurate* with the effect. Although N_{11} is causally sufficient for the effect, causal sufficiency is not the same thing as proportional causation. To illustrate the difference, Yablo asks us to consider a pigeon that has been trained to peck at all and only red objects.[14] The pigeon is presented with a red target and she pecks at it. As it happens, the target is a specific shade of crimson. What caused the pigeon to peck? Was it the fact that the target was red or the fact that it was crimson? The exclusion principle would say that since being red supervenes on being crimson and being crimson is causally sufficient for the pigeon's pecking, the redness of the target is not the cause. But this

seems wrong, as Yablo points out. The target's being red is of the right degree of specificity to count as a cause of the pigeon's action. In contrast, the target's being crimson is too specific to count as the cause: citing it as the cause of the pecking might give the erroneous impression that the pigeon would not peck at anything non-crimson.

How can we capture the requirement that causes must be proportional to their effects? Yablo formulates a proportionality constraint, which he suggests is implicit in our concept of causation.[15] But his constraint is based on a particular account of the supervenience relation between mental and neural properties that is not shared by all non-reductive physicalists. According to it, supervenient mental properties are related to their subvenient neural properties as determinables like red are related to their determinates like crimson; thus supervenience, like determination, is an unconditional, logical or metaphysical necessitation relation.[16]

Contrary to this account, the physicalist hypothesis that the mental supervenes on the physical is often presented as a contingent claim about the actual world, which means that the supervenience relation is restricted to possible worlds that are like the actual one in certain important respects.[17] Rather than adopting Yablo's own formulation of the proportionality constraint, we therefore take a more general approach, presented in the next section. It should, however, still be in the spirit of Yablo's analysis of the pigeon example.

II. CAUSATION AS DIFFERENCE-MAKING

Yablo says the motivation for imposing a proportionality constraint on causes is the dictum that causes make a difference to their effects. This dictum underlies many different theories of causation: counterfactual, probabilistic, interventionist and contrastive ones. How can we spell out this dictum? We agree with those philosophers who interpret causal claims as claims about relationships between variables and thus interpret the dictum, quite literally, as requiring that changing the value of the cause variable changes the value of the effect variable.[18] Applied to binary variables representing the presence or absence of some property, the dictum says that changing the causal property from being absent to being present (or vice versa) changes the effect property from being absent to being present (or vice versa). Formally, we suggest that the truth conditions for one property to make a difference to another are the following:

> *Truth conditions for making a difference*: The presence of *F* makes a difference to the presence of *G* in the actual situation just in case (i) if any relevantly similar possible situation instantiates *F*, it instantiates *G*; and (ii) if any relevantly similar possible situation instantiates ¬*F*, it instantiates ¬*G*.

For example, the target's being red makes a difference to the pigeon's pecking because in any relevantly similar situation in which the pigeon is presented with a red target, it pecks and in any relevantly similar situation in which it is not presented with a red target, it does not. Various specifications of relevantly similar situations might be given. In the example, they could be situations in which the pigeon has received the same training, the targets are presented to the pigeon in the same experimental setting, there are no confounding influences on the pigeon and so on. But under the same construal of the relevantly similar situations, the target's being crimson does not make a difference to the pigeon's pecking. Condition (ii) is not met: in a relevantly similar situation in which the pigeon is presented with a non-crimson but red target, it still pecks. These observations confirm Yablo's conjecture that the proportionality of causation can be captured by requiring that causes make a difference to their effects.

Further confirmation of this conjecture comes from examining how the suggested truth conditions constrain the specificity of causes: satisfaction of these conditions ensures that causes are specific enough for their effects, but no more specific than needed. This is revealed most clearly in the case of many-valued causal variables. Suppose, for example, there is a drug that causes patients to recover from an illness. The effect variable is a binary variable whose values are recovery or non-recovery. But the cause variable is many-valued, with possible values 0mg, 50mg, 100mg, 150 mg, and 200mg. Suppose that any regular dose at or above 150mg cures a patient, but any lower dose does not. Suppose a patient has taken a regular dose of 150mg and has recovered from the illness. What made the difference to the patient's recovery? According to the truth conditions above, the answer is "Giving the patient a dose of at least 150mg". It satisfies both conditions (i) and (ii): all relevantly similar patients who take a regular dose at or above 150mg recover and all those who take a lower dose don't. Other answers are either too specific, or not specific enough. For example, the cause cannot be "Giving the patient a dose above 50mg" because that does not meet condition (i): some relevantly similar patients who are given a dose above 50mg, say 100mg, do not recover. Similarly, it cannot be "Giving the patient a dose of exactly 150mg" because that does not meet condition (ii): some relevantly similar

patients who are not given a dose of exactly 150mg – say they are given 200mg – nonetheless recover. In this way, condition (i) rules out causes that are not specific enough to account for the change in the effect variable, while condition (ii) rules out causes that are too specific to account for it.

The truth conditions for making a difference can be expressed more formally using counterfactuals, as understood in a possible-world semantics. Specifically, we replace the notion of a relevantly similar situation with that of a relevantly similar possible world, identifying a situation in which a property *F* is instantiated (or not instantiated) by the proposition "*F* is present" (or "*F* is absent"), and thus rewrite the conditionals in the truth conditions above as counterfactuals:

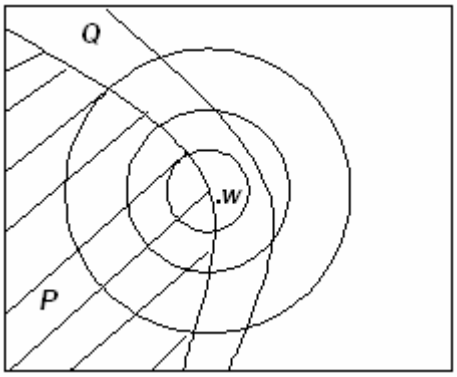
> *Truth conditions for making a difference*: The presence of *F* makes a difference to the presence of *G* in the actual world if and only if it is true in the actual world that (i) *F* is present $\Box\!\!\rightarrow$ *G* is present; and (ii) *F* is absent $\Box\!\!\rightarrow$ *G* is absent.

It is important to be precise about the semantics of these counterfactuals. We use the standard possible-worlds semantics of David Lewis, which provides truth conditions for counterfactuals in terms of a similarity relation between possible worlds.[19] The similarity relation, which may vary with context, is represented by an assignment to each possible world *w* of a system of spheres of worlds centred on *w*. The system of spheres is required to meet certain formal constraints. The spheres are *nested* in the sense that, for any two spheres *S* and *T*, either *S* is included in *T* or *T* is included in *S*. They are *weakly centred* on *w* in the sense that *w* is contained in every sphere.[20] They are *exhaustive* in the sense that there exists a largest sphere containing all relevant possible worlds. They satisfy the *limit assumption* that, for any world *w* and any non-contradictory proposition *P*, there is a smallest sphere containing some world in which *P* is true. Call this sphere the *smallest P-permitting sphere* around *w*.[21] The system of spheres conveys information about the similarity of worlds to the world *w* at the centre. The smaller a sphere, the more similar to *w* are the worlds in it. So whenever one world lies in some sphere around *w* and another lies outside it, the first world is more similar to *w* than the second.[22]

Given these assumptions, we can now state the truth conditions for counterfactuals as follows: $P \Box\!\!\rightarrow Q$ is true in world *w* if and only if *Q* is true in all the *P*-worlds within the smallest *P*-permitting sphere around *w*. (Interpretationally, those worlds are the closest *P*-worlds to *w*.) Figure 1 illustrates a situation in which the counterfactual $P \Box\!\!\rightarrow Q$ is true in the world *w*

at the centre of the system of spheres. The set of *P*-worlds is represented by the region with diagonal lines and the set of *Q*-worlds by the larger convex region that includes the set of *P*-worlds. The smallest *P*-permitting sphere is the innermost sphere.

Figure 1



By adopting this semantic framework, we follow Lewis rather than Stalnaker, in allowing that there may be more than one closest *P*-world to *w*. Although there may sometimes be just one such world, this is not the general rule. However, we diverge from Lewis in imposing only a weak centring requirement on the systems of spheres. We allow the smallest sphere around *w* to contain more than one world. Lewis imposes the stronger requirement that the smallest sphere around *w* contains only *w*. This corresponds to a constraint on the similarity relation whereby no world is as similar to *w* as *w* itself. It also corresponds to the inference rule from the premise *P* & *Q* to the conclusion $P \mathrel{\square\!\!\rightarrow} Q$. In other words, if *P* and *Q* are true in some world so is $P \mathrel{\square\!\!\rightarrow} Q$. Lewis's strong centring requirement, the corresponding constraint on similarity and the corresponding inference rule may appear plausible. But we cannot accept them. If the counterfactual formulation of the truth conditions for difference-making is to match the earlier formulation, clause (i) of the counterfactual formulation must capture the idea that every relevantly similar situation that instantiates *F* also instantiates *G*. In the original formulation, this condition is non-trivial: it rules out insufficiently specific causes, provided the set of relevantly similar situations instantiating *F* includes more than one such situation. To match this condition, the counterfactual formulation must require that even if *F* and *G* are both instantiated in the actual world, the smallest sphere around it also contains some other worlds instantiating *F*.

There are also independent reasons for weakening the centring requirement in this way. First, Lewis's strong centring requirement is justified on the seemingly reasonable grounds that no world can be as similar to world *w* as world *w* itself. For example, the actual world is more

similar to itself than any other world. But this presupposes very fine-grained standards of similarity and difference. It is only if we assume such fine-grained standards that we can exclude worlds that are nearly identical to the actual world from the smallest sphere around it. But Lewis himself warns us that not every similarity or difference between worlds should get counted in the overall similarity relation for counterfactuals.[23] As he points out, if every similarity or difference were counted, this would refute his analysis of counterfactuals. Sometimes the following counterfactuals seem true: "If *P*, the world would be very different; but if *P* and *R*, the world would not be very different". But these counterfactuals can be true only if the formal similarity relation disagrees with explicit judgments about what is "very different". So some obvious respects of similarity and difference count for nothing in the overall similarity relation. But this also suggests that worlds that differ from the actual world only in respects that do not count should be included in the innermost sphere around it.[24]

Another reason for weakening the centring requirement is that the strong version introduces an unjustified asymmetry into our reasoning about counterfactuals. Consider the inference pattern *strengthening the antecedent*, which goes from the premise $P \mathrel{\Box\!\!\rightarrow} Q$ to the conclusion $P \,\&\, R \mathrel{\Box\!\!\rightarrow} Q$. (The same point can be made in terms of *contraposition* and *transitivity*.) When the antecedent *P* is false, this inference is generally invalid. For instance, the counterfactual "If I were to stir sugar into tea, it would taste sweet" is true, but the counterfactual "If I were to stir sugar and castor oil into my tea, it would taste sweet" is false. However, under Lewis's strong centring requirement, this inference pattern is valid when the antecedents of the counterfactuals are true. But the inference seems equally bad when the antecedent is true as when it is false. This point is especially significant in the case of future-tense counterfactuals whose antecedents are not known to be true or false. It is crucial to the use of such counterfactuals for prediction and decision-making that inferences be valid or invalid regardless of the truth of their antecedents.[25]

Before we apply the difference-making account of causation to the exclusion principle, we note an implication of the account. Several philosophers have observed that causal statements are contrastive in character.[26] They have pointed out that descriptions of both cause and effect seem to involve reference to a contrast situation, or set of contrast situations. Sometimes the contrasts are made obvious by the use of contrastive focus. For example, asserting a sentence such as "Giving the patient a *150mg dose* of the drug caused his recovery" highlights the fact that the

150mg dose was one in a range of doses and not all doses within this range cause recovery. But often the contrast situations are left implicit. The rule for reconstructing the contrast situations is straightforward in the case of causal claims involving binary variables. Here the contrast situation is simply the opposite value to the actual one. So the causal claim "The presence of *F* caused *G* to be present" is to be understood as "*F* being present rather than absent caused *G* to be present rather than absent". All these observations are predictable based on the account of causation as difference making. If causal statements convey information about how variation in one variable is associated with variation in another, as explicated by a pair of counterfactuals, it is no surprise that they can be expressed contrastively.[27]

III. APPLICATION TO THE EXCLUSION PRINCIPLE

Both examples discussed above – Andersen's monkey and Yablo's pigeon – can be seen as counterexamples to the exclusion principle. In each case, the exclusion principle leads us to identify the intuitively wrong property as the cause of the given effect. In Yablo's example, the proportional difference-making cause of the pigeon's pecking is not the crimson, but the redness of the target, contrary to what the exclusion principle implies. This is supported by the truth of the counterfactuals:[28]

Target is red $\Box\!\!\rightarrow$ pigeon pecks.

Target is not red $\Box\!\!\rightarrow$ pigeon does not peck.

In contrast, the following counterfactuals are not both true:[29]

Target is crimson $\Box\!\!\rightarrow$ pigeon pecks.

Target is not crimson $\Box\!\!\rightarrow$ pigeon does not peck.

It is natural to interpret these counterfactuals in terms of a similarity relation that makes the closest-worlds in which the target is not crimson ones where it is some other shade of red. (We discuss this assumption in the final section.) Given this assumption, the second counterfactual is false: in the closest worlds in which the target is not crimson it is some other shade of red, in which case the pigeon would still peck.

A similar treatment can be given for the example of the monkey. The proportional difference-making cause of the monkey's reaching action A_1 is not its particular neural state N_{11}, but its intention I_1. The following counterfactuals are true:[30]

Monkey has intention I_1 $\Box\!\!\rightarrow$ monkey performs A_1.

Monkey doesn't have intention I_1 $\Box\!\!\rightarrow$ monkey doesn't perform A_1.

Whereas the following counterfactuals are not both true:[31]

Monkey has neural property N_{11} $\Box\!\!\rightarrow$ monkey performs A_1.

Monkey doesn't have neural property N_{11} $\Box\!\!\rightarrow$ monkey doesn't perform A_1.

Assuming that the closest worlds in which the monkey doesn't have neural property N_{11} are ones in which it has another neural property realizing the intention I_1, one can see that the second counterfactual is false: in any such world, the monkey has another neural property that realizes I_1, and so performs A_1.

Yablo's insight about the exclusion principle has thus been vindicated. Understanding causes as proportional difference-makers, one can see that the exclusion principle is false. Even when some property *F* is causally sufficient for some effect *G*, a property *F** that supervenes on *F* can nonetheless be a cause of *G*. The monkey's intention I_1 to reach for a specific target is the cause of its reaching action A_1 even though it supervenes on the neural property N_{11}, which is causally sufficient for the action.

To be fair, Kim would not regard this as a refutation of the exclusion principle in his intended sense. He states that his version of the exclusion principle is to be understood in terms of a conception of causation as generation or production rather than a counterfactual one as advanced here. He writes, for example: "Causation as generation, or effective production and determination, is in many ways a stronger relation than mere counterfactual dependence, and it is causation in this sense that is fundamentally involved in the problem of mental causation."[32] Our response is that the concept of generation or effective production requires clarification. In the sentence before the one quoted, Kim says that the conception he has in mind is the one described by Elizabeth Anscombe.[33] But Anscombe treats productive causation as primitive; and neither she nor Kim elucidates the concept. Woodward proposes that production or necessitation can be understood in counterfactual terms:[34] an event *c* produces another event *e* just in case the counterfactual "*c* occurs $\Box\!\!\rightarrow$ *e* occurs" is robustly true under variation of the background conditions. If this is the interpretation of production or generation, however, then this concept is not a genuinely causal one at all. There are well known counterexamples to this view of causation (Salmon 1984). For example, a man's taking a contraceptive pill is sufficient for, or necessitates, his not getting pregnant. But there is no causal connection or relevance here, as a

man's taking a contraceptive pill makes no difference to his not getting pregnant. Even if he had not taken the pill, he wouldn't have got pregnant.

These points suggest that unless a better explication can be given of causation as production, this notion can hardly play a significant role in the debate about mental causation. In contrast, the notion of difference-making is clearly a causal one. This is evident from several facts. First, it is often used as the central motivating idea behind various theories of causation. Second, the epistemology of causation, especially in the form of randomized controlled experiments, is closely tied to establishing properties as difference-makers for other properties.[35] Third, the notion of causal relevance or difference-making plays a central role in theories of causal explanation.[36] Fourth, difference-makers are ideally suited for the purposes of manipulation and control.[37] And finally, the clear truth conditions we have offered for claims about difference-making stand in stark contrast to the opacity of the notion of productive or generative causation.

IV. A REVISED EXCLUSION PRINCIPLE

We have seen that within a difference-making account of causation there are some persuasive counterexamples to the exclusion principle. A central feature of this principle is that it is couched in terms of causal sufficiency: it states that a property that is *causally sufficient* for some effect excludes certain other properties from being causes of that effect. But one might ask: "Why talk of causal sufficiency rather than causation?" The reference to causal sufficiency harks back to older empiricist accounts of causation that take causation simply to be some form of subsumption under laws. Indeed, several features of Kim's formulation of the exclusion argument depend on somewhat outmoded assumptions from deductive-nomological accounts of causation and causal explanation.[38]

Naturally, this raises the question of what happens if we reformulate the exclusion principle, replacing the reference to causal sufficiency with one to causation in a more adequate sense, understood as difference-making. So let us consider the following revised principle:

> *Revised exclusion principle*: For all distinct properties *F* and *F** such that *F** supervenes on *F*, *F* and *F** do not both cause a property *G*.

Here the truth conditions for causation are those for difference-making introduced above. The principle can also be formulated in two different, but logically equivalent ways. The first is the

counterpart of Kim's original principle, whereas the second is seldom explored in the debate about the exclusion problem:

> *Revised exclusion principle (upwards formulation)*: If a property *F* causes a property *G*, then no distinct property *F** that supervenes on *F* causes *G*.
>
> *Revised exclusion principle (downwards formulation)*: If a property *F* causes a property *G*, then no distinct property *F** that subvenes or realizes *F* causes *G*.

Although logically equivalent, the two formulations draw our attention to two different ways in which the exclusion principle can apply. An instance of *upwards exclusion* occurs when there exists a subvenient difference-making cause that excludes a supervenient one; and an instance of *downwards exclusion*, usually not recognized, occurs when there exists a supervenient difference-making cause that excludes a subvenient one. We turn to such instances in the next section.

Is the revised exclusion principle true or false? Let us focus on the instance of the principle that concerns the causal relationships between a mental property *M*, a neural property *N*, and a behavioural property *B*. Throughout the discussion, we assume that *N* realizes *M* in the actual world. We are interested in the logical relationship between the following two propositions:

(1) The presence of *M* is a difference-making cause of the presence of *B*.

(2) The presence of *N* is a difference-making cause of the presence of *B*.

Using the truth conditions introduced above, each of these propositions is equivalent to a conjunction of counterfactuals:

(1a) *M* is present $\Box\!\!\rightarrow$ *B* is present.

(1b) *M* is absent $\Box\!\!\rightarrow$ *B* is absent.

(2a) *N* is present $\Box\!\!\rightarrow$ *B* is present.

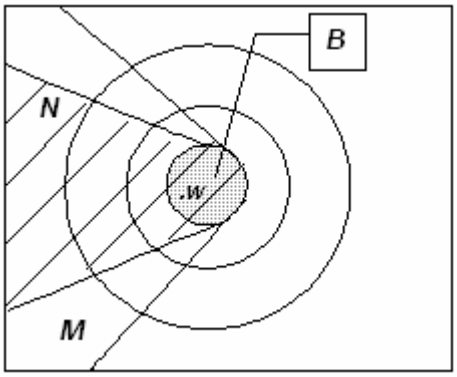
(2b) *N* is absent $\Box\!\!\rightarrow$ *B* is absent.

The revised exclusion principle dictates that propositions (1) and (2), or equivalently (1a), (1b), (2a) and (2b), are never simultaneously true. But is this claim actually correct? Our approach makes this question logically tractable. In the appendix we prove the following result:

> *Compatibility result*: Propositions (1a), (1b), (2a) and (2b) are true together if and only if (i) *B* is present in all closest *M*-worlds; (ii) *B* is absent in all closest $\neg M$-worlds; and (iii) *B* is absent in all closest $\neg N$-worlds that are *M*-worlds.

To show that conditions (i), (ii) and (iii) on the right-hand side of this biconditional can indeed be met, consider the example represented in Figure 2. As before, the concentric spheres represent sets of more and more similar worlds to the actual world; the innermost sphere contains the actual world, labelled *w*, and other worlds deemed maximally similar to it. The set of *N*-worlds is represented by the region with diagonal lines, the set of *M*-worlds by the larger region that includes the set of *N*-worlds. The shaded region represents the set of *B*-worlds.

Figure 2



It is easy to check that conditions (i), (ii) and (iii) are met. Since *M* is present throughout the innermost sphere, that sphere picks out the closest *M*-worlds, and since *B* is also present in it, condition (i) is met. Since *B* is not present in any ¬*M*-worlds, it is also absent in all closest ¬*M*-worlds, and thus condition (ii) is met. Finally, since *B* is not present in any ¬*N*-worlds, it is also absent in all closest ¬*N*-worlds that are *M*-worlds, and so condition (iii) is met. This is, of course, just one of many situations that illustrate the compatibility of (1a), (1b), (2a) and (2b). To understand the conditions in full generality, we recommend working through the proof in the appendix.

The upshot of this result is that the revised exclusion principle is not generally true. It is not an *a priori* truth that a lower-level difference-making relation such as that between *N* and *B* excludes a higher-level one such as that between *M* and *B*, or vice versa. These causal relations can coexist when the conditions in the compatibility result are satisfied. The philosophical debate about the validity of the exclusion principle has proceeded as if the principle is an *a priori* truth about all causal systems. However, we can now see that the exclusion principle is true of some systems, and false of others, and we must therefore consider its applicability on a system-by-system basis.

One general point is nicely illustrated by Figure 2, however. Although *M*, and not just *N*, is a difference-making cause of *B* here, the causal system is very sensitive in the sense that small perturbations in the way in which *M* is realized will result in the absence of *B*. Call a causal relation between *M* and *B* *realization-sensitive* if, in all those *M*-worlds that are closest ¬*N*-worlds (such that *M* has a different realizer), *B* is no longer present. Then another way of expressing our compatibility result is to say that the exclusion principle is false whenever some higher-level property stands in a realization-sensitive causal relation to another property.

V. CONDITIONS FOR UPWARDS AND DOWNWARDS EXCLUSION

So far we have emphasized the conditions under which the exclusion principle fails to hold. Let us now focus on those under which it holds. Perhaps it is true of many systems, even if not of all. If so, this has important implications for the autonomy of causal relations at different levels. As an immediate corollary of our compatibility result, we can state the conditions under which the exclusion principle holds:

> *Necessary and sufficient conditions for the exclusion principle*: The revised exclusion principle holds if and only if, for all relevant properties *M*, *N* and *B* (with *N* realizing *M*), either (i) *B* is absent in some closest *M*-worlds, or (ii) *B* is present in some closest ¬*M*-worlds, or (iii) *B* is present in some closest ¬*N*-worlds that are *M*-worlds.

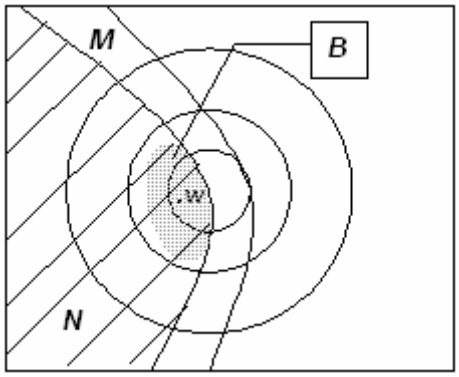
There are two non-trivial ways in which these conditions can be met (the third, trivial one involves the absence of any higher- or lower-level causal relations whatsoever). The first is an instance of upwards exclusion, which occurs when *N* causes *B* but *M* doesn't cause *B* (in which case clauses (i) or (ii) are met). The second is an instance of downwards exclusion, which occurs when *M* causes *B* but *N* doesn't cause *B* (in which case clause (iii) is met). Let us first consider the case of upwards exclusion.

> *Necessary and sufficient conditions for upwards exclusion*: An instance of upwards exclusion occurs if and only if *N* is a difference-making cause of *B* and either (i) *B* is absent in some closest *M*-worlds that are ¬*N*-worlds or (ii) *B* is present in some closest ¬*M*-worlds outside the smallest ¬*N*-permitting sphere.

A proof is given in the appendix. Figure 3 represents an example of upwards exclusion. As before, the region with diagonal lines represents the set of *N*-worlds, the larger convex region the

set of *M*-worlds, the shaded region the set of *B*-worlds. The actual world, *w*, lies in the innermost sphere, and within it *M*, *N*, and *B* are present.

Figure 3



It is easy to check that the truth conditions for counterfactuals (2a) and (2b) are satisfied so that *N* is a difference-making cause of *B*. As further required, clause (i) of the necessary and sufficient conditions for upwards exclusion is met, because *B* is absent in the closest *M*-worlds that are ¬*N*-worlds, namely those worlds in the non-shaded part of the innermost sphere. Indeed, one can easily see that the counterfactual "*M* is present $\Box\!\!\rightarrow$ *B* is present" is false, implying that *M* cannot be a difference-making cause of *B*. (In this example, we assume that the innermost sphere of worlds contains worlds besides the actual world.)

To give a less formal illustration of upwards exclusion, let us go back to Yablo's example, but suppose now that the pigeon has been trained to peck at all and only crimson objects. In this case, a target's being crimson is a difference-making cause of the pigeon's pecking, whereas the target's being red is not. The target's being red does not satisfy the conditions for making a difference precisely because the counterfactual "The target is red $\Box\!\!\rightarrow$ the pigeon pecks" is false, since the pigeon does not peck at any red but non-crimson objects. Another illustration is given by a variant of Woodward's example in which the monkey performs some highly specific action A_1 when and only when it is in a distinctive neural state N_{11}, which is one among many realizers of an intention I_1. Then the difference-making cause of the monkey's action is the subvenient neural state N_{11} rather than the supervenient intention I_1; the latter can be realized by other neural states that do not lead to the action.[39]

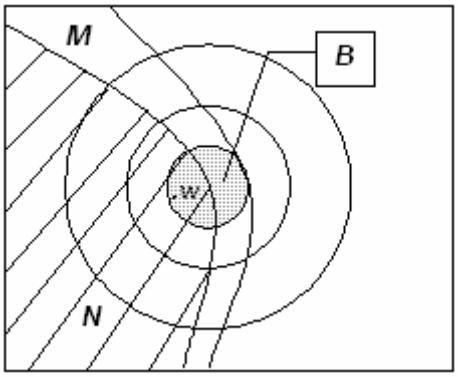
Although the conditions for upwards exclusion deserve more discussion, we devote the rest of our discussion to the conditions for downwards exclusion. We state necessary and sufficient conditions below. But a further point emerges from our analysis. It is often assumed that when a

higher-level causal relation holds, it does so in virtue of a lower-level causal relation. For example, according to Kim's *causal inheritance principle*, if property *F** supervenes on property *F* then *F**'s causal powers are identical with, or at least determined by, *F*'s causal powers.[40] But the opposite is true: when the exclusion principle holds, a higher-level property *F** can be a difference-making cause for some effect *G only if* the lower-level property *F* that realizes *F** is *not* a difference-making cause of that effect. The necessary and sufficient conditions for downwards exclusion are the following:

> *Necessary and sufficient conditions for downwards exclusion*: An instance of downwards exclusion occurs if and only if *M* is a difference-making cause of *B* and *B* is present in some closest $\neg N$-worlds that are *M*-worlds.

Again, a proof is given in the appendix. Figure 4 represents an example of downwards exclusion. As before, the diagonally lined region represents the *N*-worlds, the larger convex region the *M*-worlds, and the shaded region the *B*-worlds; *M*, *N* and *B* are present in the actual world, *w*.

Figure 4



It is easy to check that conditions (1a) and (1b) for *M* to be a difference-making cause of *B* are met. Further, *B* is present in some – in fact, all – of the closest $\neg N$-worlds that are *M*-worlds, namely those in the shaded part of the innermost sphere, as required by the conditions for downwards exclusion. Indeed, *N* is not a difference-making cause of *B*, since the counterfactual "*N* is absent $\Box\!\!\rightarrow$ *B* is absent" is false. (Once more, we assume that the innermost sphere contains worlds besides the actual world.)

How can we interpret the conditions for downwards exclusion? Recall that a causal relation between *M* and *B* is realization-sensitive if small perturbations in the way in which *M* is realized result in the absence of *B*, or formally, if *B* is absent in all those *M*-worlds that are closest $\neg N$-worlds. Similarly, call the causal relation between *M* and *B* *realization-insensitive* if *B* continues

to be present even under some small perturbations in the realization of *M*, or formally, if *B* is present in some closest $\neg N$-worlds that are *M*-worlds. We can then see that, if there exists a higher-level causal relation, it excludes a lower-level one if and only if it is realization-insensitive. This is verified in Figure 4 where the closest *M*-worlds, whether or not *M* is realized by *N* in them, are ones in which *B* is present.

It is natural to assume that the similarity relation over possible worlds has a structure like that in Figure 4, giving rise to a realization-insensitive causal relation between *M* and *B*. Indeed, in discussing Woodward's and Yablo's examples above, we made this assumption without justifying it at the time. For example, in arguing that the target's being crimson is not a difference-making cause of the pigeon's pecking, we claimed that the counterfactual "The target is not crimson $\Box\!\!\rightarrow$ the pigeon does not peck" is false because if the target had not been crimson it would have been some other shade of red, in which case the pigeon would still have pecked. But this is analogous to assuming a similarity relation according to which some of the closest $\neg N$-worlds are *M*-worlds in which *B* is present. Likewise, in arguing that the monkey's being in neural state N_{11} is a not a difference-making cause of A_1, we claimed that the counterfactual "N_{11} is absent $\Box\!\!\rightarrow$ A_1 is absent" is false because if N_{11} had been absent, some other neural property would have realized the intention I_1, which would still have led to the performance of A_1. Once again, this reasoning made an assumption analogous to the one that some of the closest $\neg N$-worlds are *M*-worlds in which *B* is present.

Why is this assumption so easy to make? Why is it so natural to assume that when there is a higher-level causal relation, it is realization-insensitive? One reason is that we intuitively require difference-making causal relations to hold in various possible situations, not just in the actual one. We take this to be an empirical observation about ordinary speakers' practices of assigning truth conditions to causal claims. Woodward has made a similar observation, developing some insights of Lewis's discussion of "sensitive" and "insensitive" causation.[41] Like Lewis, Woodward defines counterfactual dependence between events in terms of non-backtracking counterfactuals, stating that *e*'s occurrence *counterfactually depends* on *c*'s occurrence if and only if:

(3a) *c* occurs $\Box\!\!\rightarrow$ *e* occurs.

(3b) *c* doesn't occur $\Box\!\!\rightarrow$ *e* doesn't occur.

Woodward agrees with Lewis that the counterfactual (3a) is true if its antecedent and consequent are true, but notes that ordinary speakers regard it as crucial to the truth of a causal claim that these counterfactuals should remain true under various changes to the actual circumstances: they should be *insensitive*. He further notes that the insensitivity of counterfactual (3a) has more weight in an ordinary speaker's causal judgment than that of counterfactual (3b).[42]

Woodward gives many illuminating examples. But let us consider two examples from Lewis's original discussion. One of Lewis's examples of insensitive causation is shooting someone at point blank range with a large calibre bullet: as we vary the background circumstances in ways that are not too unlikely or far-fetched, it continues to be true that if the shooter had fired the bullet, the victim would have died. An example of sensitive causation, by contrast, involves the following scenario: Lewis writes a strong letter of recommendation that causes *X* to get a job he would not have got otherwise, which in turn causes *Y*, who would have gotten the job in the absence of Lewis's letter, to take a job in a distant city, where she meets and marries a person; she has children with this person, and these children in turn have children, and so on. Call one of *Y*'s descendants *N*. As Lewis notes, the causal statement "Writing his letter of recommendation caused *N*'s death" is very sensitive. Counterfactual (3b) is true: if Lewis had not written his letter of recommendation, *Y* would not have met and married the person she did and her descendants, including *N*, would not have come into existence and later died. Woodward argues that the reason we are nonetheless reluctant to regard this as a true causal statement is that counterfactual (3a) is very sensitive here. There are ever so many changes to the actual circumstances that are not too unlikely or far-fetched which would undermine its truth. For example, if the university department in the distant city had chosen someone other than *Y*, or if *Y* had not lingered at the party quite as long as she did, she would not have met her future husband and so on. Woodward's crucial observation is that despite the existence of a counterfactual dependence between Lewis's writing the letter and *N*'s death, it is the extreme sensitivity of counterfactual (3a) that undermines the credibility of the associated causal claim.[43]

The main outstanding issue is how to specify which variations in background conditions are admissible in evaluating causal claims. On our approach, this issue comes down to the question of which similarity relation to employ in assessing the relevant counterfactuals. The worlds which differ admissibly in background conditions from the actual world belong to the innermost sphere around it. But which conditions are admissibly varied and which must be

invariant under change? Woodward says this is a context-sensitive matter, and disciplinary-specific rules may play an important role.[44] For example, he observes that in economics fundamental causal relations are sometimes required to remain invariant under changes to economic agents' states of information and under changes to relative costs. Further, he notes that genetic causal relations are usually required to remain invariant under changes to imposed environmental conditions, especially changes in human social conventions.

In accordance with these remarks, we suggest that in the special sciences higher-level causal relations are typically required to be invariant under changes to the way in which higher-level properties are physically realized. The realization-insensitivity we naturally assume for causal relations involving mental properties is an instance of this more general phenomenon. If it is correct that realization-insensitivity is a general requirement in higher-level causal claims, then it follows that the conditions for downwards exclusion are generally satisfied.[45] But this in turn entails that higher-level causal relations such as that between *M* and *B* obtain even though there is no underlying causal relation between the neural realizer *N* and *B*. In such cases, we have good reason to believe in the causal autonomy of higher-level properties.

VI. CONCLUSION

We started by considering Kim's formulation of the exclusion principle, as employed in his argument against non-reductive physicalism, and rejected it on the basis of a conception of causation as difference-making. Under this conception, a mental property can be a proportional difference-making cause of a physical property even when it supervenes on a neural property that is causally sufficient for the physical property.

Despite refuting Kim's formulation of the exclusion principle, this conception of causation allowed us to formulate a revised exclusion principle not vulnerable to the same counterexamples. The revised principle permits different logically equivalent formulations – an upwards and a downwards one – which highlight different ways in which the principle can apply: the commonly recognized possibility that a lower-level cause excludes a higher-level one, and the less commonly recognized reverse possibility that a higher-level cause excludes a lower-level one. However, we showed that the new principle is not an *a priori* truth, since it can be false of some causal systems. Nonetheless, the systems that falsify it are very special in that they involve higher-level causal relations that are realization-sensitive: small perturbations that

change the realizer of the higher-level property from its actual realizer also change the truth-value of the relevant causal claims. Of course, one possibility consistent with the existence of realization-sensitive higher-level causal relations is the reductionist claim that mental properties are identical with their neural realizers. We do not deny that some mental properties may be reductively identified with their neural realizers. After all, the causal profile of a given system is an empirical matter. However, if there exist no realization-sensitive causal relations at the higher-level, the exclusion principle is true. Moreover, when there are some higher-level causal relations that are realization-insensitive, the conditions for downwards exclusion are met and the higher-level properties involved in those relations are causally autonomous. A comparison with Woodward's discussion of sensitive and insensitive causation suggested that the realization-insensitivity requirement is a special case of the more general requirement that genuine causal relations should be insensitive to variation in background conditions.

The lesson of our discussion is that whether or not non-reductive physicalism is tenable depends on the empirical characteristics of each causal system in question. For systems exhibiting some instances of downwards exclusion, non-reductive physicalism is vindicated at least minimally, i.e., the three theses of non-reductive physicalism are true with respect to some higher-level properties. For systems exhibiting only instances of upwards exclusion and no instances of higher-level causes compatible with lower-level ones, non-reductive physicalism is false, since its third thesis is false with respect to *every* higher-level property. For other systems, neither a reductive stance nor a non-reductive one is ruled out.

APPENDIX: PROOFS

Definitions. Let Ω be a non-empty set of possible worlds. Any property is represented by a subset $P \subseteq \Omega$. We write $\neg P$ to denote the negation of *P*, i.e., $\neg P = \Omega \setminus P$. We assume that, for each world $w \in \Omega$, there exists a *system of spheres of worlds* around *w* (a set of subsets of Ω), denoted $\mathbf{S}_w$, with the following properties:

Nestedness: for any $S, T \in \mathbf{S}_w$, $S \subseteq T$ or $T \subseteq S$.

Weak centring: for every $S \in \mathbf{S}_w$, $w \in S$.

Exhaustiveness: $\Omega \in \mathbf{S}_w$.

Limit assumption: for every $P \subseteq \Omega$ with $P \neq \varnothing$, $\bigcap_{S \in \mathbf{S}_w : S \cap P \neq \varnothing} S \in \mathbf{S}_w$.

For each $P \subseteq \Omega$ with $P \neq \varnothing$ and each $w \in \Omega$, define the *smallest P-permitting sphere* around *w* as

$$\min_P(w) = \bigcap_{S \in \mathbf{S}_w : S \cap P \neq \varnothing} S .$$

By the limit assumption, $\min_P(w) \in \mathbf{S}_w$ provided $P \neq \varnothing$. Define the *closest P-worlds* to *w* as

$$f_P(w) = \min_P(w) \cap P.$$

For any $P, Q \subseteq \Omega$, the conditional $P \mathrel{\square\!\!\rightarrow} Q$ is true in world *w* if all the closest *P*-worlds to *w* are *Q*-worlds, formally $f_P(w) \subseteq Q$.

Question. Let *M* be some (supervenient) mental property, *N* some (subvenient) neural property, and *B* some behavioural property. We are interested in whether the following two claims are compatible:

(1) *M* is a difference-making cause of *B*. Equivalently:

(1a) $M \mathrel{\square\!\!\rightarrow} B$.

(1b) $\neg M \mathrel{\square\!\!\rightarrow} \neg B$.

(2) *N* is a difference-making cause of *B*. Equivalently:

(2a) $N \mathrel{\square\!\!\rightarrow} B$.

(2b) $\neg N \mathrel{\square\!\!\rightarrow} \neg B$.

Assumption. We assume that *N* realizes *M* in the actual world *w*, i.e., $w \in N \subseteq M$.

Result 1.

(i) (1a) is true in $w \Rightarrow$ (2a) is true in *w*.

(ii) (1b) is true in $w \Rightarrow$ [(2b) is true in $w \Leftrightarrow f_{\neg N}(w) \cap M \subseteq \neg B$]

Proof of (i): Suppose (1a) is true in *w*, i.e., $\min_M(w) \cap M \subseteq B$. Since $w \in N \subseteq M$, we have $\min_M(w) = \min_N(w)$. Therefore $\min_N(w) \cap N \subseteq \min_M(w) \cap M$, whence $\min_N(w) \cap N \subseteq B$. So (2a) is true in *w*. **Q.E.D.**

Proof of (ii): Suppose (1b) is true in *w*, i.e., $\min_{\neg M}(w) \cap \neg M \subseteq \neg B$. What are the truth conditions for (2b) in *w*, i.e., for $\min_{\neg N}(w) \cap \neg N \subseteq \neg B$? Since $N \subseteq M$, we have $\neg M \subseteq \neg N$. Hence, for any $S \in \mathbf{S}_w$, if $S \cap \neg M \neq \varnothing$ then $S \cap \neg N \neq \varnothing$, and thus $\min_{\neg N}(w) \subseteq \min_{\neg M}(w)$. Since $\min_{\neg M}(w) \cap \neg M \subseteq \neg B$, it follows that $\min_{\neg N}(w) \cap \neg M \subseteq \neg B$. So

$\min_{\neg N}(w) \cap \neg N \subseteq \neg B$ if and only if $\min_{\neg N}(w) \cap (\neg N \setminus \neg M) \subseteq \neg B$

if and only if $\min_{\neg N}(w) \cap \neg N \cap M \subseteq \neg B$

if and only if $f_{\neg N}(w) \cap M \subseteq \neg B$. **Q.E.D.**

Result 2.

(i) (2a) is true in $w \Rightarrow [$(1a) is true in $w \Leftrightarrow f_M(w) \cap \neg N \subseteq B]$.

(ii) (2b) is true in $w \Rightarrow [$(1b) is true in $w \Leftrightarrow f_{\neg M}(w) \setminus \min_{\neg N}(w) \subseteq \neg B]$.

Proof of (i): Suppose (2a) is true in *w*, i.e., $\min_N(w) \cap N \subseteq B$. What are the truth conditions for (1a) in *w*, i.e., for $\min_M(w) \cap M \subseteq B$? Since $\min_M(w) = \min_N(w)$ (as noted in part (i) of the earlier proof), it follows that $\min_M(w) \cap N \subseteq B$. So

$\min_M(w) \cap M \subseteq B$ if and only if $\min_M(w) \cap (M \setminus N) \subseteq B$

if and only if $\min_M(w) \cap M \cap \neg N \subseteq B$

if and only if $f_M(w) \cap \neg N \subseteq B$. **Q.E.D.**

Proof of (ii): Suppose (2b) is true in *w*, i.e., $\min_{\neg N}(w) \cap \neg N \subseteq \neg B$. What are the truth conditions for (1b) in *w*, i.e., for $\min_{\neg M}(w) \cap \neg M \subseteq \neg B$? Since $\neg M \subseteq \neg N$, it follows that $\min_{\neg N}(w) \cap \neg M \subseteq \neg B$. Moreover, we know that $\min_{\neg N}(w) \subseteq \min_{\neg M}(w)$ (as noted in part (ii) of the earlier proof). So

$\min_{\neg M}(w) \cap \neg M \subseteq \neg B$ if and only if $(\min_{\neg M}(w) \setminus \min_{\neg N}(w)) \cap \neg M \subseteq \neg B$

if and only if $\min_{\neg M}(w) \cap \neg M \setminus \min_{\neg N}(w) \subseteq \neg B$

if and only if $f_{\neg M}(w) \setminus \min_{\neg N}(w) \subseteq \neg B$. **Q.E.D.**

All the results stated in the main text of the paper follow immediately from results 1 and 2.

ENDNOTES

[*] We are grateful to Richard Bradley, Nancy Cartwright, David Chalmers, Franz Dietrich, Daniel Hausman, Christopher Hitchcock, Graham Macdonald, Huw Price, Kai Spiekermann, Daniel Stoljar, Laura Valentini, Stephen Yablo, seminar participants at the ANU, Brown University, CUNY, LSE and MIT, and conference participants at the 2008 Sydney-Tilburg Conference on Reduction and the Special Sciences for helpful comments and discussion.

[1] Jaegwon Kim, *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Cambridge, MA: MIT Press, 1998) and *Physicalism or Something Near Enough* (Princeton, NJ: Princeton University Press, 2005).

[2] Kim acknowledges that, literally, properties are not causes of other properties: it is instances of properties that cause instances of other properties. We follow Kim in using this convenient shorthand way of speaking of properties as causes and effects. Sometimes we speak of states as the instances of properties that are the causes and effects.

[3] Kim's own argument is more subtle and less vulnerable to objection than this very simplified version. But we set aside the details here and focus on the exclusion principle itself, which is indispensable in all versions of the argument.

[4] See, for example, Kim, *Physicalism or Something Near Enough* (op. cit.).

[5] One might regard realization-sensitivity as a plausible criterion for identifying higher-level properties with their physical realizers.

[6] Kim's principle with such a clause might read: *except in cases of genuine overdetermination*, if a property *F* is causally sufficient for a property *G*, then no distinct property *F** that supervenes on *F* causes *G*. One might object that the examples below are not genuine counterexamples because they involve cases of overdetermination. We do not explore this objection in detail, as there are significant disanalogies between examples of overdetermination and our examples. In overdetermination cases, the rival causes are logically and metaphysically independent of each other, so that one can ask meaningful counterfactual questions about what would happen if one of the rival causes occurred without the other occurring. Would the victim have died if one member of the firing squad had not fired? Would the fire have occurred if the short circuit had occurred without the lightning strike? In the examples below there are significant logical or metaphysical dependencies between the rival causes, so these counterfactual questions cannot be asked nontrivially. This is not just an artificial feature of the examples, or a superficial disanalogy with overdetermination cases. It is essential for overdetermination that (i) the effect would have occurred if one of the causes was present without the other, and (ii) the effect would not have occurred if neither had been present. When the causes are logically and metaphysically independent, condition (i) can be non-trivially satisfied, but when they are not, it cannot. See also Larry Shapiro and Elliott Sober, "Epiphenomenalism: The Do's and the Don'ts", in Gereon Wolters and Peter Machamer, eds., *Studies in Causality: Historical and Contemporary* (Pittsburgh: University of Pittsburgh Press, 2007) and James Woodward, "Mental Causation and Neural Mechanisms", in Jakob Hohwy and Jesper Kallestrup, eds., *Being Reduced: New Essays on Reduction, Explanation, and Causation* (Oxford: Oxford University Press, 2008).

[7] It is straightforward to extend our account of causation as a difference-making relation to probabilistic systems, though generalizing our results about the limitations of the revised exclusion principles would be more technical.

[8] Thus we set aside causal systems involving pre-emption and overdetermination as these complications are not germane to our discussion. Of course, if our examples could be understood as involving overdetermination, this simplifying assumption would be unwarranted. But we have already suggested that the examples are dissimilar to cases of overdetermination.

[9] Contrary to our formulation, the exclusion principle is sometimes formulated in terms of events, but then its implications reappear as implications about the properties associated with events. Our formulation also differs slightly from other property formulations, which do not restrict the competing properties *F* and *F** to ones related by supervenience but typically restrict them to simultaneously instantiated ones. Clearly, some restriction is needed, as many properties, instantiated at different times in a causal chain, can be causes of some effect. We impose the restriction above since it is the most relevant one for our argument. The exclusion principle by itself does not imply that a physical property *F* will always be available. An additional assumption that the physical world is causally closed is needed to ensure that there is a physical property *F* to compete with the higher-level property *F**.

[10] In James Woodward, "Mental Causation and Neural Mechanisms" (op. cit.).

[11] S. Musallam, B. D. Corneil, B. Greger, H. Scherberger and R. A. Andersen, "Cognitive Control Signals for Neural Prosthetics", *Science*, 305 (July 2004): 258-262.

[12] Specific intentions about how limbs are to be moved to reach the target are encoded elsewhere in the motor cortex.

[13] See, for example, Carl Craver, *Explaining the Brain* (New York: Oxford University Press, 2007), ch. 6; Peter Menzies, "The Exclusion Problem, the Determination Relation, and Contrastive Causation", in Jakob Hohwy and Jesper Kallestrup, eds., *Being Reduced: New Essays on Reduction, Explanation, and Causation* (Oxford: Oxford University Press, 2008); James Woodward, "Mental Causation and Neural Mechanisms" (op. cit.); Stephen Yablo, "Mental Causation", *Philosophical Review*, 101, 2 (April 1992): 245-280.

[14] Yablo, "Mental Causation" (op. cit.).

[15] Ibid. The constraint says, roughly, that a property *F* is proportional to an effect *G* just in case *F* screens off all its determinates from *G* and *F* is not screened off by any of its determinables from *G*. The present terminology of "screening off" was introduced in Matthew McGrath, "Proportionality and Mental Causation: A Fit?", *Philosophical Perspectives* 12 (December 1998): 167-176. This terminology is to be understood as follows: property *F screens off* a property *H* from another property *G* if and only if, for any object *x*, if *x* were *F* but not *H*, then *x* would still be *G*.

[16] It is tempting to interpret the supervenience relation as a straightforward determination relation because both are relations of asymmetric necessitation: determinates necessitate their determinables but not vice versa; and similarly, subvening neural properties necessitate their supervening mental properties, but not vice versa.

[17] See, for example, Eric Funkhouser, "The Determinable-Determinate Relation", Nous, 40, 3 (September 2006): 548-569, and Menzies, "The Exclusion Problem, the Determination Relation, and Contrastive Causation" (op. cit.). The kind of necessitation involved in the claim that determinates necessitate their determinables differs from that involved in the claim that subvenient neural properties necessitate their supervening mental properties. The first kind

is logical or metaphysical. It can be spelled out in terms of inclusion of sets of possibilia, defined over an unrestricted universal set of all possible worlds. Take all the possibilia in those worlds; then the set of crimson possibilia is included in the set of red possibilia. The second kind of necessitation is contingent and non-logical; physicalists do not wish to rule out dualism as logically impossible. One way of capturing this contingency is to restrict the set of worlds over which the supervenience relation is defined to *minimal physical duplicates* of the actual world, which contain the same physical objects, physical properties and relations, and physical laws as the actual world, and nothing else. A physical, neural property *N* then subvenes (and necessitates) a mental property *M* just in case the set of possibilia that instantiate *N* is included in the set of possibilia that instantiate *M*, where the possibilia are restricted to the set of minimal physical duplicates of the actual world. The standard features of the supervenience relation follow from this: any possibilia from this restricted set that differ in the property *M* must differ in the property *N*; and any possibilia in this set that agree in respect of *N* must agree in respect of *M*. See David Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996), Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press, 1998), and David Lewis, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61, 4 (December 1983): 343-377.

[18] See, for example, Christopher Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy*, 98, 6 (June 2001): 273-299; Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2000); and James Woodward, *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press, 2003). In his forthcoming paper, "Causation, Exclusion, and the Special Sciences", *Erkenntnis*, Panu Raatikainen offers an independently developed analysis of mental causation in terms of an interventionist theory of causation.

[19] David Lewis, *Counterfactuals* (Oxford: Blackwell, 1973).

[20] Together with the limit assumption below, this entails that there exists a smallest sphere containing *w* and possibly other worlds. More on this later.

[21] Without this assumption, there could be an endless sequence of smaller and smaller *P*-permitting spheres around *w* but no smallest such sphere; Lewis also considers this case.

[22] Under this notion of similarity, the system has the property that if there is a set of worlds such that every member of the set is more similar to *w* than any non-member of the set, then the set is one of the spheres around *w*.

[23] David Lewis, "Counterfactual Dependence and Time's Arrow", *Nous*, 13, 4 (November 1979): 455-476.

[24] Indeed, nothing in the weightings of similarity and difference with respect to avoidance of miracles and maximization of spatiotemporal region of perfect match of particular fact which Lewis (ibid.) recommends for the causally relevant counterfactuals suggests that no world can be as similar to the actual world as it is to itself.

[25] Besides these formal constraints, our only other constraint on counterfactuals is that they are non-backtracking. If properties E_1 and E_2 are both effects of property *C*, it is tempting to reason that if E_1 had not been instantiated then *C* would not have been instantiated, in which case E_2 would not have been instantiated. But such reasoning involves backtracking, which must be banned if the counterfactual rendering of difference-making causation is to work. This can be achieved through several different ways of specifying the similarity relation. We do not commit ourselves to

one of them, but note that they carry different commitments regarding the status of our account. In "Counterfactual Dependence and Time's Arrow" (op. cit.), Lewis proposes a similarity relation for non-backtracking counterfactuals that gives special weight to the avoidance of miracles and the maximisation of the spatiotemporal region of match of particular facts. Lewis's specification of the similarity relation avoids any use of causal notions as he represents his counterfactual theory of causation as a reductive analysis. In *Making Things Happen: A Theory of Causal Explanation* (op. cit.), Woodward proposes a similarity relation for non-backtracking counterfactuals in terms of the causal notion of an intervention, which plays the same role as a miracle in Lewis's account. Woodward concedes that, due to his causally loaded semantics, his theory of causation cannot offer a reductive analysis, but argues that his account is nonetheless informative about causation. In eschewing any specification of the similarity relation for non-backtracking counterfactuals, we remain neutral on whether our counterfactual account is fully reductive or not.

[26] See, for example, Fred Dretske, "Referring to Events", *Midwest Studies in Philosophy*, 2, 1 (September 1977): 90-99; Christopher Hitchcock, "The Role of Contrast in Causal and Explanatory Claims", *Synthese*, 107, 3 (June 1996): 395-419; James Woodward, "A theory of singular causal explanation", *Erkenntnis*, 21, 3 (November 1984): 231-262; and Woodward, *Making Things Happen: A Theory of Causal Explanation* (op. cit.).

[27] If the causal claim that "*F*'s being present made the difference to *G*'s being present" implies that changing the situation from *F* being present to being absent also changes the situation from *G* being present to being absent, and vice versa, as expressed by the counterfactuals "*F* is present $\Box\!\!\rightarrow$ *G* is present" and "*F* is absent $\Box\!\!\rightarrow$ *G* is absent", then it is appropriate to say that "*F*'s being present rather than absent caused *G*'s being present rather than absent".

[28] Alternatively, it is also borne out by the truth of the contrastive statement: The target's being red rather than not red made it the case that the pigeon pecked rather than didn't peck.

[29] Similarly, the following contrastive statement is false: The target's being crimson rather than not crimson made it the case that the pigeon pecked rather than did not peck.

[30] In contrastive terms, the monkey's having intention I_1 rather than not having this intention made it the case that it performed A_1 rather than did not perform it.

[31] Again in contrastive terms, it is false that the monkey's having neural property N_{11} rather than not having this property made it the case that it performed A_1 rather than did not perform this act.

[32] In *Physicalism or Something Near Enough* (op. cit.), p. 18.

[33] Elizabeth Anscombe, "Causality and Determination", in *Inaugural Lecture* (Cambridge: Cambridge University Press, 1971).

[34] In James Woodward, "Sensitive and Insensitive Causation", *Philosophical Review*, 115, 1 (January 2006): 1-50.

[35] See Craver, *Explaining the Brain* (op. cit.).

[36] See Wesley Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton NJ: Princeton University Press, 1984) and Woodward, *Making Things Happen: A Theory of Causal Explanation* (op. cit.).

[37] Again, see Woodward, *Making Things Happen: A Theory of Causal Explanation* (op. cit.).

[38] For a discussion of this point, see Woodward, "Mental Causation and Neural Mechanisms" (op. cit.).

[39] One might speculate that whether a higher-level or lower-level property should be cited as a difference-making cause is determined by how the relevant effect property is described. Sometimes the effect property is a coarse-

grained, higher-level one, in which case it is most likely to be explained in terms of a coarse-grained, higher-level cause. But other times, as illustrated by the examples above, the effect can be explained only by a fine-grained, lower-level property. When this is so, we have an instance of upwards exclusion.

[40] Kim, *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (op. cit.).

[41] See Woodward, "Sensitive and Insensitive Causation" (op. cit.) and David Lewis, "Postscripts to 'Causation'", in *Philosophical Papers: Volume 2* (Oxford: Oxford University Press, 1986), respectively.

[42] Woodward says that the counterfactual (3a) is insensitive to the degree that there is a broad range of background conditions β, that are not too improbable or far-fetched, such that the following counterfactual is true:

(3c) *c* occurs in circumstances β different from the actual ones $\Box\!\!\rightarrow$ *e* occurs.

The insensitivity of counterfactual (3b) is defined similarly; and the insensitivity of the corresponding causal statement "*c* caused *e*" is defined as some weighted average of the insensitivity of counterfactuals (3a) and (3b), with most weight going to counterfactual (3a).

[43] We believe that Lewis and Woodward are describing in different terms the same phenomenon we have described. They say that ordinary speakers are typically inclined to regard causal statements like "*c* caused *e*" as true not just when a pair of counterfactuals like (3a) and (3b) are true, but when in addition counterfactual (3a) is insensitive. We say that ordinary speakers typically regard this causal statement as true just when counterfactuals (3a) and (3b) are true, where the counterfactual (3a) already has built into its truth conditions the requirement that *e* would occur if *c* were to occur in background circumstances different from the actual ones. The difference in our views is due to a difference about the truth conditions of counterfactuals like (3a). If one accepts Lewis's view that a counterfactual is true if it has a true antecedent and consequent, one has to build the additional requirement of the sensitivity of the counterfactual (3a) into the truth conditions of causal statements. In contrast, one does not have to add this requirement if one accepts the view that the truth of such a counterfactual already requires a connection between *c*'s occurrence and *e*'s occurrence under an admissible range of changes to the background conditions.

[44] In "Sensitive and Insensitive Causation" (op. cit.).

[45] To summarize the point, we are reluctant to acknowledge the existence of a higher-level causal relation unless it exhibits a certain degree of realization-insensitivity. But in this case the conditions for downwards exclusion are met, and the higher-level causal relation excludes any subvenient, lower-level one.