

MENTAL CAUSATION

Suppose that, for every event, whether mental or physical, there is some physical event causally sufficient for it. Suppose, moreover, that physical reductionism in its various forms fails—that mental properties cannot be reduced to physical properties and mental events cannot be reduced to physical events. In this case, how could there be mental causation? More specifically, how could mental events cause other mental events, physical events, and intentional physical actions? The primary goal of this paper is to answer this question.¹

The explanation that emerges is based on three guiding ideas. First, a mental event (rather than a competing physical event) causes a subsequent *mental event* because of the special strength of certain fundamental psychological laws, namely, laws upon which acceptable nonreductive functional definitions may be based. Second, a mental event (rather than a competing physical event) causes a subsequent *physical event* because of (a) the strength of these psychological laws plus (b) the strength of relevant psychophysical correlations. When the details of this picture are worked out, mental-to-mental and mental-to-physical causation will turn out to be species of *trumping preemption*, wherein the power to trump competing physical causes derives from these two factors.² Third, when a mental event (rather than a competing physical event) causes a subsequent *intentional action*, we have an instance of what I call *essential-constituent causation*—where one essential constituent of the intentional action is physical and the other mental. After having given this three-stage account in a deterministic setting, I will show how to extend it to a probabilistic setting.

An subsidiary goal concerns the question of just how strong the psychophysical correlations are. Specifically, does the mental supervene on the physical as a matter of metaphysical necessity, or does it supervene in some weaker fashion? The goal is to construct an account that remains neutral with respect to this highly controversial question, for, other things being equal, it is best to steer clear of avoidable controversies. If a successful neutral account can be given, a significant dialectical point follows: such an account will undermine an interesting new theoretical argument in favor of metaphysical supervenience (hereafter, simply ‘supervenience’), namely, that

supervenience must be adopted as a premise in any successful account of mental causation.³ I should note, finally, this neutral account is also compatible with various forms of naturalism (and with their denials).

Another subsidiary goal of the paper is to explain, not just how a mental event can be a cause of a mental or physical effect, but also how it can be *the* cause—or, at least, why in a particular context it is correctly deemed to be *the* cause. Some philosophers think that this sort of question is entirely a matter of pragmatics (interest, salience, etc.). But even if pragmatic considerations are involved, it does not follow that there are not objective criteria that, relative to a context, make it correct to identify one event as *the* cause of another (rather than, for example, one of two overdetermining causes, one of two joint causes, or some other alternative). A full account should make clear what these criteria are. If correct, our account will do this.

A final subsidiary goal is to provide the resources to answer a question receiving much attention of late, namely, how to distinguish between genuine (justification preserving) inferences and merely incidentally caused sequences of thought.⁴ The answer is that genuine inferring is a species of mental causation of the sort explained by our account (i.e., a species of mental causation underwritten by the sort of laws of rational psychology upon which nonreductive functional definitions may be based). I will not, however, have space to develop this account of inference here.

Before beginning, I should elaborate upon my starting points. First, I will assume that mental properties are not reducible to either first-order or second-order physical properties. That is, I will assume that the identity thesis and reductive functionalism are mistaken.⁵ (My reason for rejecting the ordinary identity thesis is based on multiple realizability intuitions together with a rebuttal of the scientific-essentialist (i.e., necessary a posteriori) response.⁶ My reason for rejecting reductive functionalism (both “American” and “Australian”) is that reductive functional definitions require the wrong sorts of things to be the contents of our self-consciousness: the contents would have to be propositions involving physical “realizations” rather than mental properties themselves.⁷) Some readers will of course be unwilling to abandon these reductive theses. This paper, however, should still be of interest to them, insofar as a standard objection to the

non-reductionist alternative is that it is unable to account for mental causation. For, if correct, the present account will answer this objection.

Second, even though reductive functionalism is mistaken, nonreductive functionalism is correct.⁸ (Or, more cautiously, I will assume that there is a family of distinctively strong mental-to-mental ties of the sort that would be assured by reductive functionalism if it were correct.) By ‘nonreductive functionalism’ I mean that form of functionalism that identifies the standard mental properties and relations (being in pain, thinking, etc.) with the *unique* sequence of properties and relations (R_1, R_2, \dots) that satisfies an appropriately general psychological theory (hereafter, A). Accordingly, we have:

x is in pain iff_{def} there is a unique sequence R_1, R_2, \dots satisfying A and x has R_1 .

x thinks p iff_{def} there is a unique sequence R_1, R_2, \dots satisfying A and x is related by R_2 to p .

And so forth.

Third, I will assume that an adequate account of mental causation must explain the role *mental properties* play in mental causation, and so must go beyond a coarse-grained, token-identity account (perhaps in the spirit of Donald Davidson’s) which does not explicitly provide such an explanation. If the role of mental properties can somehow be explained in a fine-grained framework, then, plausibly, it can be reworked into an account of mental causation constructed in a setting of coarse-grained events. In view of this, since the fine-grained framework is so easy to work with—and since I find it to be more plausible in any case—I will assume that this framework is correct. Finally, as indicated above, I will assume weak causal closure: every actual event has some physical event that is causally sufficient for it. Of course, weak causal closure does not entail strong causal closure, namely, that every actual event has physical and only physical causes. Causing an event is intuitively a very different matter from merely being causally sufficient for an event. It is in the logical gap between weak and strong causal closure that mental causation lives. Failure to appreciate this opening in logical space has led many philosophers to the premature conclusion that mental causation is untenable unless mental properties are somehow reducible to physical properties.⁹

1. An Empirical Test for Causes

So how is mental causation possible in a world like ours? Let us begin by considering an idealized case that does not involve mental causation. Suppose that we have correctly narrowed down the candidate causes of a given effect e to two preceding events, c and d , which are simultaneous with each other. That is, suppose that we have applied various other tests for causes of e , and c and d are the only events that have passed all of them. Suppose, moreover, that one of the following holds: (1) c is the cause of e , (2) d is the cause of e , (3) each one individually causes e (i.e., c and d overdetermine e), (4) neither one individually causes e (rather, c and d jointly cause e). In addition, let us for now make two further simplifying assumptions: first, there are no relevant intermediary effects falling temporally between these two events and the effect (this will be important later); second, the laws are deterministic and, in particular, that the flow of events leading to e falls under them. (In section §6 I will suspend the assumption of determinism and show how the account works in a probabilistic setting.) It will be very important to bear in mind that these suppositions already rule out various candidate counterexamples to the test I am about to propose (e.g., this is achieved by the supposition that we have correctly narrowed down the candidate causes of a given effect e to two preceding events, c and d , which are simultaneous with each other).

In such a setting, how would one go about settling empirically which of the alternatives (1)-(4) holds? When practicable, we would use the following two-part screening-off method : (i) we would hold the background conditions b fixed as much as possible and see what happens in situations where events of one type, say, type c , are present but where, rather than having an event of the competing type d , we have instead an event of some relevant alternative type d' ; and conversely, (ii) we would hold the background conditions b fixed as much as possible and see what happens in situations where events of the other type d are present but where, rather than having an event of the competing type c , we have instead an event of some relevant alternative type c' .¹⁰

We will be interested in four outcomes, corresponding to the above four cases (1)-(4). And each of these outcomes has two parts corresponding to the test's two parts (i) and (ii). These outcomes are as follows.

Outcome (1). In part (i) of the method, the background conditions *b* are held constant as much as possible and we consider what happens when a *c*-type event occurs but, rather than being accompanied by a *d*-type event, some relevant alternative *d'*-type event occurs. Suppose that, in these type-(i) situations, *e*-type effects typically (perhaps always) *do* occur. In part (ii) of the method, *b* is again held constant as much as possible and a *d*-type event occurs but, rather than being accompanied by a *c*-type event, some relevant alternative *c'*-type event occurs. Suppose that, in these type-(ii) situations, *e*-type effects typically (perhaps always) *fail* to occur. If this pair of results were to obtain, we would be led to conclude (almost always correctly) that *c* is *the* cause of *e*.

Outcome (2). Suppose, instead, that things are the other way round. In situations of type (i), *e*-type effects typically (perhaps always) fail to occur, while in situations of type (ii), *e*-type effects typically (perhaps always) do occur. From this pair of results, we would be led to conclude (almost always correctly) that *d* is *the* cause of *e*.

The remaining outcomes (3) and (4) are specified analogously.

By way of illustration, consider a case of trumping preemption, for example, Bas van Fraassen's major/sergeant case.¹¹ The major and sergeant are both shouting various commands to the troops, who in cases of conflict obey the superior officer, and in cases in which only one officer gives a command, obey that officer. Suppose the major and the sergeant both shout "Advance!" simultaneously. Which event, if either, is the cause of the troops advancing? The screening-off test directs us to consider (holding the background conditions constant) what happens (i) in situations in which the major shouts "Advance!" and the sergeant shouts some relevant alternative command, say, "Retreat!" and (ii) in situations in which the sergeant shouts "Advance!" and the major shouts some relevant alternative command, say, "Retreat!". In the former type of situation, the troops typically—indeed, always—advance, and in the latter, they typically fail to advance. Thus, we have a case of Outcome (1), supporting the conclusion that the major's shouting "Advance!" is the cause of the troops advancing. Intuitively, this is the right result.

Notice that in part (i) of the major/sergeant example, not only do the troops typically advance, they *always* advance. By contrast, in part (ii), although the troops typically fail to advance,

this is not always the outcome; they sometimes advance, namely, when the major shouts nothing at all (i.e., he shouts the null command, as it were). This pattern—(i) always passing the first part of the test and (ii) typically failing the second part—turns out to be the pattern exhibited in cases of mental causation. In the following discussion I will be concerned with this pattern.

In the test we are supposed to consider what happens when a c-type event occurs, not in the presence of a d-type event, but rather in the presence of some relevant alternative d'-type event. This gives rise to the question of which types of alternative events are relevant alternatives. For example, in the major/sergeant case what alternative things is the sergeant permitted to do? Shoot the major? Bribe the troops? Incite mutiny among the troops? No. These are completely out of the question. And so are subvocalizing “Advance!”, voting for “Advance!”, listening for “Advance!”, and so forth. The relevant alternatives to shouting “Advance!” are shouting “Retreat!”, “Stand Ready!”, and other types of sanctioned battlefield commands. That is, they arise from “toggling” the original command, replacing it with other types of sanctioned battlefield commands.

These points suggest the following useful notation. I will write $c(\alpha)$ to highlight a salient constituent α of event c . And I will write $c(\alpha')$ for the event that arises from $c(\alpha)$ by replacing α with α' . For example, in the context of the major/sergeant case, $c(\text{“Advance!”})$ is the major’s shouting “Advance!”, $c(\text{“Retreat!”})$ is the major’s shouting “Retreat!”, and so on for the other relevant alternatives. Likewise, $d(\text{“Advance!”})$ is the sergeant’s shouting “Advance!”, and so on. This notation allows us to state our test much more briefly. Suppose that $d(\alpha')$ -type events are relevant alternatives to $d(\alpha)$ -type events. Then, in the first part of the test we are to consider (holding the background conditions constant) what happens when a $c(\alpha)$ -type event occurs and a $d(\alpha')$ -type event occurs instead of a $d(\alpha)$ -type event. Analogously for the second part of the test.

Notice that in the major/sergeant example the relevant alternatives to the major's shouting “Advance!” and to the sergeant’s shouting “Advance!” are commensurable—in each case, they arise from toggling an identical parameter, namely, the type of sanctioned battlefield command. In the general case, symbolized with $c(\chi)$ and $d(\delta)$, the relevant alternatives need not arise from toggling parameters that are identical in this way. That is, the parameters χ and δ may differ. As we

will see, mental causation exhibits this possibility. To handle such cases, we will generalize our formulation of the test in the obvious way.

2. A Philosophical Test for Causes

The aim of the above test is only to provide an empirically sufficient (vs. necessary) condition for concluding that c is the cause of e (or that d is the cause of e , or that each is an overdetermining cause of e , or that they together jointly cause e). But often, especially in cases of philosophical significance (e.g., mental causation), it is not practically or even nomologically possible to perform this empirical test. Nevertheless, the empirical test suggests an analogous philosophical test that overcomes this shortcoming, namely, a test in which we consider *hypothetically* the same sorts of combinations considered in the empirical test. The intention is that, in contexts in which the competitors have been correctly narrowed down to c and d , the test provides a metaphysically *sufficient condition* for c 's being the cause of e (and so forth).

Suppose, as before, that the best competitors for being the cause of e have been narrowed down to $c(\chi)$ and $d(\delta)$.¹² (This supposition restricts the range of examples to which the test I am about to propose is applicable; see the close of the opening paragraph of §1.) Suppose $c(\chi')$ and $d(\delta')$ are, respectively, relevant alternatives to $c(\chi)$ and $d(\delta)$. Let b be the background conditions. Then, just as in the empirical test, we have four relevant outcomes, each consisting of two parts. In the case of outcome (1) they are: (i) For all relevant alternatives $d(\delta')$, in the nearest world(s) in which b and $c(\chi)$ occur and $d(\delta')$ occurs instead of $d(\delta)$, e still occurs.¹³ (ii) For most (typical) relevant alternatives $c(\chi')$, it is not the case that, in the nearest world(s) in which b and $d(\delta)$ occur, and $c(\chi')$ occurs instead of $c(\chi)$, e occurs. If (i) and (ii) are satisfied, (1) tells us that $c(\chi)$ is *the* cause of e .¹⁴ That is, with $c(\chi)$ and $d(\delta)$ as best competitors, (1.i) and (1.ii) taken together provide a sufficient condition for $c(\chi)$'s being *the* cause of e . Outcomes (2)-(4), which are specified in the obvious way, provide corresponding sufficient conditions for what causes e .¹⁵

To illustrate how this test for sufficient conditions works, consider the major/sergeant case once again. The results are as follows. (i) Consider the nearest world(s) in which: the original background conditions hold, the major shouts "Advance!", and the sergeant shouts one of the

battlefield commands other than “Advance!” (say, “Retreat!”). In all such worlds the troops still advance. Moreover, this holds for every relevant alternative to the sergeant’s shouting “Advance!”.

(ii) Consider the nearest world(s) in which: the original background conditions hold, the sergeant shouts “Advance!”, and the major shouts one of the battlefield commands other than “Advance!” (say, “Retreat!”). In all such worlds the troops fail to advance. Moreover, this holds for most (typical) relevant alternatives to the sergeant’s shouting “Advance!”. (The null command is an exception.) Thus, according to the test, the major’s shouting “Advance!” prevails as the cause of the troops advancing. The right result.¹⁶

A refinement of the screening-off test is required to deal with certain cases in which $c(\chi)$ and/or $d(\delta)$ bears a *holistic* relationship to the background. In such cases, it might not be possible for both the *entire* original background b to be held constant and $c(\chi)$ to occur without $d(\delta)$, or conversely. A convenient way to deal with this issue is by *factoring* the background conditions b into the background conditions b_c relevant to c and the background conditions b_d relevant to d .

An example involving mental causation will make it clearer how factoring the background conditions works. Let x be someone with body y .¹⁷ Let $m_1(A)$ be the mental event of x ’s thinking that A . Let $m_2(\neg\neg A)$ be the subsequent mental event of x ’s thinking that $\neg\neg A$. Suppose that x is in the sort of situation in which we would ordinarily consider $m_1(A)$ to be the cause of $m_2(\neg\neg A)$. For example, a situation in which, upon thinking that A , x comes to think $\neg\neg A$ by virtue of inferring it from the prior thought. We may suppose, for instance, that x is a logic student working on a proof one of whose premises is the proposition that A and whose eventual conclusion is supposed to involve the double negative. Let $p_1('A')$ be a brain event correlated with $m_1(A)$, and $p_2(' \neg\neg A')$ be a brain event correlated with $m_2(\neg\neg A)$. For example, $p_1('A')$ might be the event of *this* particular Mentalese token of ‘ A ’ being in y ’s Thinking Box, and $p_2(' \neg\neg A')$ might be the event of *that* particular Mentalese token of ‘ $\neg\neg A$ ’ being in y ’s Thinking Box.¹⁸

Now the holistic nature of a person’s total mental state is legendary (and is indeed encoded in nonreductive functional definitions). The relationship between $m_1(A)$ and its cognitive background b_{m_1} is a case in point. For example, there might be trouble in part (ii) of the test where

we are supposed to consider what happens in the nearest worlds in which: the entire background condition b still holds, p_1 occurs, and some relevant alternative to m_1 occurs instead of m_1 . But given that b is the *entire* background, it includes not just the physical background relevant to $p_1('A')$ but also the cognitive background b_{m_1} relevant to m_1 (i.e., to x 's thinking that A). Accordingly, the envisaged test world might not be possible, for b_{m_1} includes so much information about x 's concurrent auxiliary cognitive states that it might require the presence of m_1 (i.e., x 's thinking that A).¹⁹ So, rather than trying to hold constant the *entire* background b , we can factor b into b_{m_1} and b_{p_1} and then run the two parts of the test while holding constant whichever is relevant. In part (ii), for example, we are to consider whether $m_2(\neg\neg A)$ occurs in the nearest world(s) in which: the physical background b_{p_1} holds, $p_1('A')$ occurs, and $m_1(B)$ occurs instead of $m_1(A)$. Factoring thus allows the physical background b_{p_1} of $p_1('A')$ to be constant while relevant portions of the mental background b_{m_1} may vary as needed for the occurrence of $m_1(B)$.

The notion of a *wide event* allows us to further simplify our phrasing of the test (I am not strictly speaking committed to them). Suppose that $w_{m_1}(A)$ is the wide event of x 's thinking that A against the relevant cognitive background b_{m_1} . Likewise, let $w_{p_1}('A')$ be the corresponding wide physical event (namely, this particular tokening of ' A ' in y 's Thinking Box against the relevant physical background b_{p_1}).²⁰ Then in part (i) we are to consider whether $m_2(\neg\neg A)$ occurs in the nearest world(s) in which $w_{m_1}(A)$ occurs but in which relevant alternatives $p_1('B')$ occur instead of $p_1('A')$. In part (ii) we are to consider whether $m_2(\neg\neg A)$ occurs in the nearest world(s) in which $w_{p_1}('A')$ occurs but in which relevant alternatives $m_1(B)$ occur instead of $m_1(A)$.

(Before proceeding, a few general comments about the test should be helpful. To begin with, there is no commitment to the test's always having determinate outcomes. The idea is that, if there are determinate outcomes, the identity of the cause is settled accordingly. It should also be borne in mind that, though this test invokes the notion of nearest worlds, it is not committed to a nearest-worlds analysis of counterfactuals or a nearest-worlds or counterfactual analysis of causation. In fact, it is compatible with noncounterfactual, realist analyses such as those of Fred Dretske, Michael Tooley, and David Armstrong. Similarly, it is not committed to anything like Lewis's doctrine of

“Humean supervenience.” So, for example, in specifying the conditions under which the test may be applied, I make free use of causal and nomological notions, and in making judgments about the nearness of worlds, I will rely freely on nomological notions (see especially §§3-4). The claim is only that the test yields (or generally yields) the intuitively correct judgments concerning causation in the kinds of cases at issue. And even if it should turn out that the test falls short of this in certain cases, its application to relevant test cases would nonetheless provide an illuminating way to develop the three guiding ideas mentioned above. In this way, even if the test should fall short, it would still provide insight into the nature of mental causation.)

3. Mental-to-Mental Causation

My objective in this section is to outline an explanation of why laws governing mental-to-physical transitions have a special trumping power, and how this makes it correct, in the envisaged context, to say that m_1 , not p_1 , is the cause of m_2 . We will assume that the best competitors for cause of m_2 have already been narrowed down to m_1 and p_1 (as above, this assumption is important because it rules out various candidate counterexamples). I will also continue to assume that determinism holds until §6, where I will adapt the account to a more realistic probabilistic setting. Finally, I will assume that, among the psychological background conditions included in $w_{m_1}(A)$, is the fact that both x and x 's generally good cognitive conditions persist *throughout* the relevant period.²¹ Traditional epiphenomenalism was concerned with mental-to-mental causation. According to it, mental events are never caused by mental events because the psychological laws characterizing the relevant mental-to-mental transitions are only *derived* laws. These laws are underwritten by *basic* physical-to-physical laws together with nomologically or causally necessary psychophysical principles. Nonreductive functionalism, if correct, shows what is wrong with this idea and, in turn, with this traditional form of epiphenomenalism. Given nonreductive functionalism (which was one of our starting points), the standard mental properties are defined as the unique satisfiers of an appropriately general psychological theory A . A logical consequence of these definitions is that the indicated mental-to-mental laws are basic in an especially strong sense: they are *metaphysically necessary*.²² To see why, consider the nonreductive functional definition of

thinking. (To simplify the presentation, I will assume, without loss of generality, that ‘thinks’ is the only psychological constant in psychological theory A.)

x thinks p iff_{def} there is a unique relation R satisfying theory A and x is related by R to p .

Since definitions hold necessarily, this definition has the following as an immediate consequence: necessarily, if x thinks p , then the original psychological theory A is true. In symbols,

$\square (x \text{ thinks } p \rightarrow \text{theory } A \text{ is true}).$

Therefore, since $m_1(A)$ is the event of x 's thinking the proposition that A and since the occurrence of the wide event $wm_1(A)$ trivially entails the occurrence of the constituent event $m_1(A)$, it follows that: necessarily, if $wm_1(A)$ occurs, then the psychological theory A is true. In symbols:

$\square (wm_1(A) \rightarrow \text{theory } A \text{ is true}).$

Now suppose A contains, or entails, the principle that, if $wm_1(A)$ occurs, then $m_2(\neg\neg A)$ will occur.

Then, the last conclusion implies that: necessarily, if $wm_1(A)$ occurs, this principle is true. In symbols:

$\square (wm_1(A) \rightarrow (wm_1(A) \rightarrow m_2(\neg\neg A))).$

Therefore, by simplification, it follows that:

$\square (wm_1(A) \rightarrow m_2(\neg\neg A)).$

That is, necessarily, if $wm_1(A)$ occurs, then $m_2(\neg\neg A)$ will occur. This is the conclusion we sought.

Naturally, this generalizes to other mental-to-mental conditionals contained in, or entailed by, psychological theory A .

In our ensuing discussion it will be helpful to have the following terminology. Let *primary psychology* be the psychological theory, or theories, upon which correct nonreductive functional definitions can be based, and let *primary psychological laws* be the mental-to-mental conditionals belonging to, or entailed by, primary psychology.²³ In this terminology, the above conclusion is neatly stated thus: given nonreductive functional definitions, primary psychological laws hold necessarily. In informal terms, the moral is this. It is in the very nature of mental properties to interact with one another in accordance with the primary laws of psychology, and nonreductive

functional definitions record this fact. This is the vision to which functionalist philosophy, once separated from its unsuccessful reductionist ambitions, has been pointing all along.²⁴

Still, this does not yet show that there really is mental-to-mental causation, for example, that $m_1(A)$ is the cause of $m_2(\neg\neg A)$. Nor does it show why an event like $m_1(A)$, together with correlated physical events such as $p_1('A')$, is neither an overdetermining cause nor a joint cause. But, predictably, showing these things turns on the special modal status of principles of primary psychology. Let us apply our two-part test.

In part (i), for each relevant alternative to p_1 , we are to consider what happens in the nearest world(s) in which the wide mental event $wm_1(A)$ occurs but in which the relevant alternative to p_1 occurs instead of p_1 . Is it the case that, in the indicated class of nearest world(s), $m_2(\neg\neg A)$ still occurs? Yes. The reason is that there is a metaphysically necessary law of primary psychology connecting the occurrence of $wm_1(A)$ and the subsequent occurrence of $m_2(\neg\neg A)$, thereby creating the strongest sort of pressure for $m_2(\neg\neg A)$ to occur. Since there is no equally strong physical-to-physical law creating a contrary pressure, $m_2(\neg\neg A)$ does indeed occur. Hence, $m_1(A)$ passes one half of our test.

In part (ii) of the test we are to consider what effect results in the nearest world(s) in which wp_1 occurs and relevant alternatives to m_1 occur instead of m_1 .²⁵ Before answering this question, however, we must take a moment to look more closely at what these alternatives are. For guidance, consider the major/sergeant example once again. We saw that, in the context of that example, the relevant alternatives to shouting "Advance!" are not subvocalizing "Advance!", voting for "Advance!", listening for "Advance!", and so forth. The right constituent to toggle is not what the sergeant is doing with respect to "Advance!". Rather, the type of sanctioned battlefield command ("Advance!", "Retreat!", etc.) is the right constituent to toggle. Our logic-problem example is parallel. The propositional attitude is not the right constituent to toggle: the relevant alternatives to thinking that A are not desiring that A, doubting that A, remembering that A, and so forth. Rather the proposition which x is thinking is the right constituent to toggle.

But which propositions are the relevant alternatives to the original proposition A? In the

major/sergeant example, if the alternatives to “Advance!” were restricted to commands that were more or less equivalent to the original command (e.g., “Forward!”, “Advance, you idiots!”, “OK men, do what we did at San Juan Hill!”) or to commands that more or less entail the original command (e.g., “Advance with vigor!”, “Advance quickly!”). In such a case, the test would not correctly locate the cause, for each alternative to “Advance!” is itself sufficient for the troops advancing. The problem is avoided only if the range of relevant alternative battlefield commands is kept broad—going far beyond those that are (more-or-less) equivalent to “Advance!” or that (more or less) entail “Advance!”. Indeed, the range of alternatives must comprise the full range of sanctioned battlefield commands. Similarly, in our logic-problem case the range of alternative propositions B must be kept comparably broad—going far beyond propositions that are (more-or-less) equivalent to A or (more or less) A-entailing. In other words, parity requires that the relevant alternatives go far beyond propositions such as A & A and the like. As far as this example is concerned, just about any proposition B appropriate to a logic exam is an appropriate alternative to A. Of course, it may be that various pragmatic factors serve to narrow the range of alternatives a bit. There is nothing wrong with this, as long as a substantial range of alternatives survives.

We are now ready to consider the second half of our test. Our focus is on the wide physical event $wp_1('A')$ which involves the narrow physiological event $p_1('A')$ plus relevant features of the physical background. The latter features include the physical correlates of x’s cognitive background—that is, physical correlates of x’s auxiliary cognitive contents (including a great many auxiliary A-ish contents) and physical correlates of x’s generally good cognitive conditions (intelligence, attentiveness, memory, etc.). Let B be a typical relevant alternative to A. What happens in the nearest world(s) in which $wp_1('A')$ occurs and $m_1(B)$ occurs instead of $m_1(A)$?

The answer is that $m_2(\neg\neg A)$ fails to occur in at least *some* nearest “ $wp_1('A')$ & $m_1(B)$ ” world(s). The argument turns on the holistic character of mind (discussed at the close of §2)—specifically, the interplay of $m_1(B)$ with x’s auxiliary cognitive contents and x’s cognitive conditions. In the actual world x’s cognitive conditions are generally good. In a given “ $wp_1('A')$ & $m_1(B)$ ” world, either x’s cognitive conditions would be generally good, or they would be

degraded. Suppose the former. Then, since x is thinking that B (instead of A) in that world, x 's generally good cognitive conditions (intelligence, attentiveness, memory, etc.) would require his auxiliary cognitive contents to harmonize with B rather than A ; specifically, x 's auxiliary cognitive contents would have to be B -ish in character (rather than A -ish). For example, since x is now thinking that B (instead of A), x 's generally good cognitive conditions would require that x no longer be in his original auxiliary state of *being aware* that he is thinking A ; instead, he would have to be aware that he is thinking B . And so forth.²⁶ On the other hand, suppose that x 's cognitive conditions are degraded in the given " $wp_1('A') \ \& \ m_1(B)$ " world. Then, since in the actual world x has a stupendous number of dispositional mental properties associated with his generally good cognitive conditions, x would in the given world have to lose these properties and acquire a very different set of new dispositional properties. This stupendous departure from the actual would be at least as great as that associated with the shift (just contemplated) from A -ish to B -ish auxiliary cognitive contents. Consequently, among the nearest " $wp_1('A') \ \& \ m_1(B)$ " worlds, there are at least *some* in which x retains his generally good cognitive conditions and instead undergoes relevant shifts from A -ish to B -ish auxiliary cognitive contents.

Now consider the wide mental event $wm_1(B)$ that accompanies the wide physical event $wp_1('A')$ in one such nearest " $wp_1('A') \ \& \ m_1(B)$ " world. This wide mental event is constituted of x 's generally good cognitive conditions, x 's B -ish auxiliary cognitive contents, and the narrow event of x 's thinking that B . We saw earlier that nonreductive functional definitions imply that, necessarily, if $wm_1(A)$ occurs, then $m_2(\neg\neg A)$ occurs. The same would hold *mutatis mutandis* for most relevant alternatives B to A : necessarily, if $wm_1(B)$ occurs, then $m_2(\neg\neg B)$ occurs. Accordingly, for most relevant alternatives B to A , there will be at least some " $wp_1('A') \ \& \ m_1(B)$ " worlds in which the necessary laws of primary psychology would send $wm_1(B)$ to the succeeding event $m_2(\neg\neg B)$ rather than $m_2(\neg\neg A)$. The desired result.

Our conclusion, therefore, is that m_1 wins both halves of the screening-off test. Therefore, given that either m_1 or p_1 is the cause of m_2 (or each separately causes m_2 , or they jointly but not separately cause m_2), it follows that m_1 is the cause of m_2 . Because of the modal strength of

primary psychology, m_1 trumps p_1 as the cause of m_2 .

4. Transition to Mental-to-Physical Causation

My next main goal is to show how to extend the foregoing ideas to obtain an account of mental-to-physical causation. The purpose of this transitional section is to build a bridge to the eventual full account by developing an account that applies to a certain circumscribed family of cases. I begin with two preliminaries aimed at a general physical characterization of relevant physical correlates of mental events.

The first is a sketch of an account (which I will use for heuristic purposes) of what it is for a being to have a body. Consider a being U who has a rich mental life. And consider a body V that has a functional architecture of the sort contemplated by language-of-thought functionalism. V has a Raw Experience Box, Belief Box, Desire Box, Decision Box, and so forth. Various words and sentences of Mentalese are tokened in these boxes. (For simplicity, I will pretend for a moment that Mentalese is English, and I will often use single quotes where, strictly, corner quotes are needed.) The following input-output conditionals hold for V : if there is damage to V 's exterior, 'pain' is tokened in V 's Raw Experience Box; if F -ing is a certain kind of macroscopic bodily motion that V can exhibit and there is no external force impeding V from exhibiting F and some token or other of 'Do F ' is in V 's Decision Box, then V will exhibit F ; and so forth. These conditionals (with normal-conditions clauses included) hold with nomological necessity. Finally, suppose U and V are biconditionally related as follows: U experiences E iff some token or other of 'E' is in V 's Raw Experience Box; U believes that S iff some token or other of 'S' is in V 's Belief Box; U desires that S iff some token or other of 'S' is in V 's Desire Box; U decides to do F iff some token or other of 'Do F ' is in V 's Decision Box; and so forth. And suppose these biconditionals hold with nomological necessity.

If these and kindred conditions were fulfilled, I would be inclined to say that U *has a body* and, in particular, that V is U 's body. I find it plausible, moreover, that U would have a body only if some such conditions were fulfilled. These considerations suggest that the notion of having a body can be explicated along the following lines (for arbitrary agents u and bodies v): u has body v iff u

has a suitable array of mental properties and v has such and such organization and u 's mental contents and the Mentalese tokenings in v 's modules match up in so and so nomologically necessary fashion.²⁷ Let us suppose that something along these general lines succeeds. Of course, someone might hold that some of the indicated psychophysical biconditionals (or at least the right-to-left or left-to-right halves of such biconditionals) hold with a necessity stronger than nomological necessity. To accommodate this idea, we need only relax the account, requiring instead that the psychophysical biconditionals hold with a necessity that is *no weaker* than nomological necessity (and maybe stronger). Relaxing the account on this point guarantees that it will be consistent with the neutral stance we are trying to take on the question of supervenience.²⁸

Now for the second preliminary point. Two paragraphs above we were pretending for simplicity that Mentalese is English. When we stop doing this, the analysis would have something like the following form (for arbitrary agents u and bodies v): u has body v iff_{def} there exists a content function c from physical types (which play the role of Mentalese expressions) to propositions such that it is at least nomologically necessary that, for all p , u believes p iff, for some physical type s for which $c(s) = p$, a token of s is in v 's Belief Box; u desires p iff, for some s for which $c(s) = p$, some token or other of s is in v 's Desire Box; and so forth.²⁹ This formulation serves to isolate classes of physical types relevant to the problem of mental causation. For example, in our logic-problem case we know that the student x has body y , so the analysis tells us that there exists a content function c of the indicated sort.³⁰ Let α be the physical types whose tokens in y have content A .³¹ Then gp_1 is defined to be the event of an α 's being tokened in y 's Thinking Box. The event gp_2 is defined analogously except that the proposition that $\neg\neg A$ takes the place of the proposition that A . Hereafter, let us call gp_1 , gp_2 , and other such events *general brain events*. Associated with this notion is an important family of (at least) nomologically necessary psychophysical biconditionals: m_1 iff gp_1 ; m_2 iff gp_2 ; and so forth.³² (Hereafter, called "*the psychophysical biconditionals*.")

Now for the first step in our account of mental-to-physical causation. The guiding idea is that the psychophysical biconditionals have a certain special trumping power. At the outset, I stated

the highly plausible principle that the extent of (nonmagical, e.g., nontelekinetic) mental-to-physical causation in a given world is a function of two quantities: (a) the strength of laws governing relevant mental-to-mental transitions in that world and (b) the strength of relevant psychophysical correlations in that world. In the previous section we saw that the former are as strong as can be: they are governed by *necessary* primary psychological laws thus ensuring extensive mental-to-mental causation. Given this, the extent of mental-to-physical causation is a function of just the strength of relevant psychophysical correlations. For example, if the psychophysical biconditionals were also to hold necessarily, extensive mental-to-physical causation would likewise be ensured.³³ But very few people believe that the psychophysical biconditionals hold necessarily. The point is that the psychophysical biconditionals are so strong that they are a hallmark of the actual world and a large sphere of worlds surrounding it, a hallmark sufficiently rudimentary to underwrite mental-to-physical causation. Nearly every party to the contemporary debate over mental causation is already committed to this idea, as we will see.

Of course, virtually no one, and surely no contemporary advocate of supervenience, believes that a mere neurological event in isolation from the larger brain to which it belongs is metaphysically sufficient for *x* to think that *A*. The neurological event must be situated in a brain, and not just any brain but one that is operating in accordance with relevant physical laws. Absent such laws, the brain would be “dead.” Therefore, they need a weaker supervenience principle. One such principle is Kim’s “strong supervenience” (which is the principle Yablo invokes in his account of mental causation, if only for sake of illustration). But many materialists find this principle too strong, and in any case a weaker (but still rather strong) Horgan-Lewis-Jackson style principle will do.³⁴ Let me explain.

Call a world *nonalien* iff every natural property instantiated in it is instantiated in some nomologically possible world.³⁵ Call two worlds *complete physical duplicates* iff they are alike in all physical respects, both qualitatively and numerically (that is, they are alike in the concrete physical particulars existing in them, in the distribution of physical qualities and relations instantiated in them, and in the physical laws holding in them). And call two worlds *complete*

duplicates simpliciter iff they are alike in *all* aspects, both qualitatively and numerically (that is, they are alike in the concrete particulars existing in them, in the distribution of qualities and relations instantiated in them, and in the laws holding in them). Now consider the following rather strong supervenience principle (which virtually all contemporary materialists accept):

Among nonalien worlds, any two that are complete physical duplicates are complete duplicates simpliciter.³⁶

This principle tells us that in each nonalien world the physical facts (the concrete physical facts together with the physical laws of that world) fix all the facts.³⁷

Although this supervenience principle ensures extensive nonmagical mental-to-physical causation (see (i) and (ii) below), it entails a still weaker supervenience principle strong enough to ensure the same thing, as I will now explain.³⁸ Notice that every physically possible world is a complete physical duplicate of some nomologically possible world.³⁹ Since every nomologically possible world is by definition nonalien, it follows that every nonalien physically possible world is a complete physical duplicate of some nonalien nomologically possible world. Therefore, the above supervenience principle implies that every physically possible nonalien world is a complete duplicate of a nomologically possible nonalien world. Thus, every physically possible nonalien world is nomologically possible. In other words,

In all nonalien worlds, if the physical laws hold, so do all other laws, including the psychophysical biconditionals.

Let us call this principle *nonalien metaphysical supervenience*. As before, virtually all contemporary materialists would accept this principle.⁴⁰

As I indicated, this principle supports an account of mental-to-physical causation. I will go through the usual two steps.

Part (i). For all relevant alternatives 'B' to 'A', is it the case that in all nearest worlds in which $wm_1(A)$ occurs and $gp_1('B')$ occurs instead of $gp_1('A')$, the physical event $gp_2(' \neg \neg A')$ still occurs? Yes, if (as is extremely plausible) these nearest test worlds are nonalien. Choose an arbitrary such world. Since $wm_1(A)$ occurs and $gp_1('B')$ occurs instead of $gp_1('A')$ in this world, there would have to be a large number of violations of the psychophysical biconditionals at

t_1 — specifically, $m_1(A)$ iff $gp_1('A')$ and $m_1(B)$ iff $gp_1('B')$ and also a host of psychophysical biconditionals dealing with x 's cognitive conditions and auxiliary cognitive contents. But the principle of nonalien metaphysical supervenience tells us that these biconditionals would have to hold at t_1 if the physical laws hold at t_1 . Therefore, by contraposition, it follows that the physical laws do not hold at t_1 . In other words, there would have to be a “miraculous” break (as David Lewis would call it), not just in the psychophysical biconditionals, but also in the physical laws.⁴¹ This break in the physical laws would be only momentary, however; they would snap right back into effect at t_2 . Otherwise, the test world would have gratuitous miraculous breaks in its determinative structure, namely, the physical laws, contradicting the hypothesis that this test world is a *nearest* test world. Given that the physical laws snap right back into place at t_2 : nonalien metaphysical supervenience implies that psychophysical biconditionals also snap right back into place at t_2 . In particular, $m_2(\neg \rightarrow A)$ iff $gp_2(' \neg \rightarrow A')$ would hold at t_2 . But, by the reasoning of §3, the occurrence of $wm_1(A)$ at t_1 necessitates the occurrence of $m_2(\neg \rightarrow A)$ at t_2 . It follows, therefore, that $gp_2(' \neg \rightarrow A')$ also occurs at t_2 . The desired result.

Part (ii). For most relevant alternatives B to A , is it the case that, in all nearest worlds in which $wgp_1('A')$ occurs and $m_1(B)$ occurs instead of $m_1(A)$, the physical event $gp_2(' \neg \rightarrow A')$ still occurs? No, if (as is extremely plausible) these nearest test worlds are nonalien. Amongst the nearest test worlds, there are at least some worlds like those described in part (ii) of §3. That is, amongst the nearest test worlds, there are at least some worlds in which x 's cognitive conditions are still good and x 's auxiliary cognitive contents are B -ish in character (rather than of A -ish in character). For such test worlds, the reasoning of §3 shows that the occurrence of $wm_1(B)$ at t_1 necessitates the occurrence at t_2 of $m_2(\neg \rightarrow B)$ in place of $m_2(\neg \rightarrow A)$. Now, just as in part (i), a large number of psychophysical biconditionals are violated at t_1 in the indicated test worlds. But since such worlds are nonalien, the reasoning of part (i) shows that there are miraculous momentary breaks in the physical laws at t_1 but they snap right back into effect at t_2 . Consequently, by nonalien metaphysical supervenience, the psychophysical biconditionals would also snap right back into place at t_2 . In particular, $m_2(\neg \rightarrow B)$ iff $gp_2(' \neg \rightarrow B')$ would hold at t_2 . But, as we saw, $m_2(\neg \rightarrow B)$

occurs in place of $m_2(\neg \neg A)$ at t_2 . Therefore, $gp_2(' \neg \neg B')$ occurs in place of $gp_2(' \neg \neg A')$ at t_2 . Once again, the desired result. Thus, parts (i) and (ii) show that $m_1(A)$ is the cause of $gp_2(' \neg \neg A')$.

The above account of mental causation would also go through using a *much* weaker supervenience principle. Consider the supervenience conditional: if the physical laws hold, then all the laws hold (including the psychophysical biconditionals). Call a world *nominally supervenient* iff this conditional holds in it. Nonalien metaphysical supervenience tells us that *every* nonalien world is nominally supervenient whereas all that is needed for the above account is the following very weak principle:

The closest nonalien test worlds are nominally supervenient.

This weaker supervenience principle tells us that, even if there are nonalien worlds (and perhaps even nonalien test worlds) that fail to be nominally supervenient, the *closest* nonalien test worlds are all nominally supervenient. This principle—call it *nonalien nomic supervenience*—is something that nearly *all* parties to the contemporary debate would be willing to accept, including a great many who would identify themselves as anti-materialists. When this principle is used in place of nonalien metaphysical supervenience, the resulting account goes through just as before, thus establishing that the added strength of metaphysical supervenience is superfluous.

Of course, since nonalien nomic supervenience is so much weaker than nonalien metaphysical supervenience, it is not surprising that there should be equally effective supervenience principles of intermediate strength. For example:

Amongst nonalien worlds that duplicate, or very closely resemble, the actual world prior to a time t but depart from the actual world in one or more events that occur at t , those worlds that are nearest to the actual world are nominally supervenient.⁴²

Since our account goes through using a variety of plausible supervenience principles that are far weaker than nonalien metaphysical supervenience, our account is (as we hoped) neutral on the question of metaphysical supervenience. (Indeed, in the probabilistic setting of Appendix I, there is no need for any supervenience principles, for no laws—psychophysical or physical—are broken in the nearest test worlds.)

Summing up, in the competition between $m_1(A)$ and $gp_1('A')$, $m_1(A)$ prevails as the cause

of $gp_2('A')$. In addition, analogous reasoning shows that $gp_1('A')$ prevails over $p_1('A')$ as the cause of $gp_2('A')$.⁴³ Therefore, by transitivity, $m_1(A)$ prevails over $p_1('A')$ as the cause of $gp_2('A')$. From the outset, however, we agreed that $p_1('A')$, $gp_1('A')$, and $m_1(A)$ are (from their temporal distance t_1) the only reasonable competitors from their temporal distance (i.e., at t_1 for being the cause of $gp_2('A')$). Therefore, since $m_1(A)$ prevails over both $p_1('A')$ and $gp_1('A')$, it follows that $m_1(A)$ is *the* cause of the physical event $gp_2('A')$. Thus, we have an account of one form of mental-to-physical causation. My next task is to show how this form of mental-to-physical causation provides the basis for an account of more familiar forms of mental-to-physical causation (e.g., ringing doorbells, talking, writing, etc.).

Before proceeding, let us note how the combination of the laws of primary psychology and the psychophysical biconditionals add up to the special trumping power alluded to earlier. Consider once again how a human being differs from an epiphenomenal system. The difference is the product of two factors: first, necessary laws of primary psychology create an autonomous pressure for mental-to-mental transitions to occur; second, nonalien nomic supervenience (or nonalien metaphysical supervenience) ensures that in all closest nonalien test worlds the psychophysical biconditionals—and so, in particular, their mental-to-physical halves—snap right back into place immediately after momentary degradations. When we “compose” these two “arrows”—the mental-to-mental arrow ($wm_1\text{-to-}m_2$) and the mental-to-physical arrow ($wm_2\text{-to-}gp_2$)—the product is a mental-to-physical “arrow” ($wm_1\text{-to-}gp_2$) that is “stronger” than the competing physical-to-physical arrow ($wgp_1\text{-to-}gp_2$); that is, we get a mental-to-physical arrow that can trump the physical-to-physical arrow. It can trump the physical-to-physical arrow because of the absolute strength of the mental-to-mental arrow and because nonalien nomic supervenience (or nonalien metaphysical supervenience) ensures that the psychophysical biconditionals—and, in particular, the ensuing mental and physical events—will be in sync. By contrast, in a superficially similar epiphenomenal system (e.g., a machine with a biconditionally connected monitoring device), there simply are no basic laws (corresponding to laws of primary psychology) that connect the successive states of the monitoring device. As a result, the “arrow” from earlier states of the

monitoring device to ensuing states of the machine lacks such trumping power; on the contrary, the “arrow” from the machine’s earlier states to its ensuing states has the trumping power. Similarly, in a system consisting of two duplicate computers wired together and operating in parallel (i.e., biconditionally), the basic physical laws governing successive states of one computer are (we may assume) the same as those governing successive states of the other. Consequently, there is no possibility of trumping one way or the other.⁴⁴

5. Mental-to-Physical Causation: Physical Behavior and Intentional Action

In our dialectical context, we are supposing a tripartite distinction between a mental event, the associated general brain event, and the specific brain event of which the general brain event is a determinable. For example, we have been supposing that there are distinctions between the mental event of x’s thinking that A and, correlated with this mental event, the specific neurological event and associated general brain event. Analogous distinctions of course hold in the case of decision. Take, for example, a case in which x decides to press the doorbell button. In this case we have the events of x’s deciding to press the doorbell button, the corresponding *general* brain event correlated with x’s decision, and the *specific* neurological event of which the general brain event is a determinable. Adapting our earlier notation, let us hereafter refer to these three events as m_2 , gp_2 , and p_2 , respectively. Of course, these three events are preceded by a similar triplet of events associated with x’s antecedent (derived) desire to press the doorbell. Let these three events be m_1 , gp_1 , and p_1 .

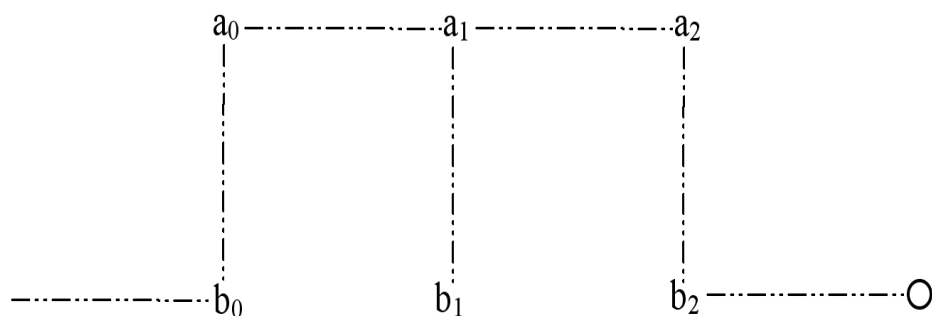
Now consider the event of my pressing the button (i.e., my intentionally pressing the button). This event is of course distinct from the event of the finger’s displacing the button by moving along exactly *this* path. Let this latter kinetic event be k . Of course, my pressing the button does not require k ; the finger need only have some appropriate pure motion or other. Let gk be this general kinetic event—the finger’s exhibiting one or another pure motion of the relevant sort. Now my intentionally pressing the button differs from gk (and k) in so far as the former event involves an intentional factor, namely, my concurrent intending, or my concurrent trying, to press it.⁴⁵ Call this concurrent intending (concurrent trying) m_3 . Since my pressing the button involves both

factors (the general kinetic event and the concurrent intending) as essential constituents, it is what we may call a *hybrid event*. This hybrid event is an *intentional action*.⁴⁶ Call it a_1 .

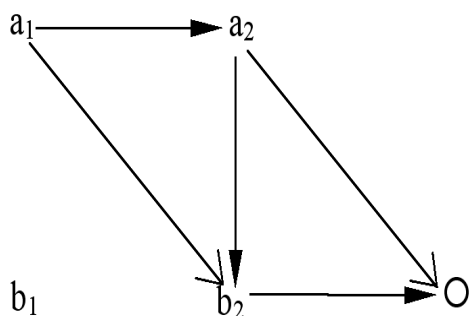
Thus, we have four effects to consider: the specific kinetic event (k), the general kinetic event (gk), the concurrent intending (m_3), and the hybrid (intentional-*cum*-kinetic) event (a_1). Our goal is to explain how in a world like ours the decision to press the doorbell (m_2) can be the cause of the general kinetic event (gk) and intentional action (a_1). The explanation will be guided by an analogy involving a system in which the pattern of causes mirrors that which we find in genuine mental causation.

At the close of the previous section, we considered a system consisting of two duplicate computers wired together and operating in parallel (i.e., biconditionally). Our tests show that the pattern of causation in such a system is one of joint causation. In cases of genuine mental causation, however, a mental event is not a mere joint cause but rather *the* cause of various ensuing events—including events involving physical behavior and intentional action. (At least, this is what we ordinarily say in relevant contexts of evaluation.) There is, however, a different kind of system of computers in which the pattern of causes mirrors that in genuine mental-to-physical causation.

Consider, as before, two computers (let them be A and B) that are wired together so that their internal events (or states) are biconditionally correlated with causal or nomological necessity in the usual way; for example, a_1 and b_1 are so related, as are a_2 and b_2 . Unlike the previous system of computers, however, most of the circuits in B's internal processor have been blown. In spite of this, B is still attached to an external monitor, and it still displays outputs on it, for example, a circle \bigcirc

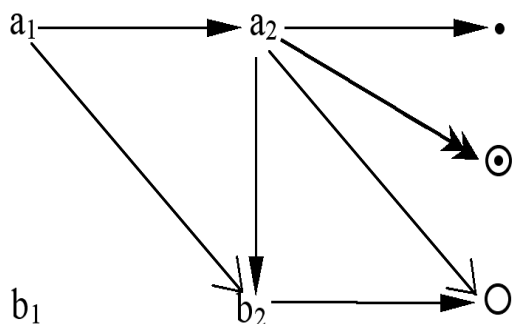


at its center. This is why I said only that “most” of B’s circuits have been blown: although the circuits relating to B’s internal-to-internal transitions have been blown, those relating to its internal-to-external transitions are still intact. We then have the following picture. Arrows represent relations of cause as we intuitively take them to be (solid-headed arrows for proximal causes and open-headed arrows for distal causes).



a_1 is the cause of a_2 and of b_2 . Of course, a_1 causes b_2 *via* a_2 . That is, a_1 is the distal cause of b_2 , and this is so because a_1 is the proximal cause of a_2 and a_2 is the proximal cause of b_2 . Finally, b_2 is the proximal cause of the appearance of \bigcirc , and since a_2 is the proximal cause of b_2 , it is the distal cause of \bigcirc .

Next we complicate the example slightly by supposing that we disconnect B from the monitor and connect A to it instead. And we suppose that in this case a dot \bullet appears on the monitor instead of \bigcirc . The resulting causal situation is unchanged except that a_2 is now the proximal cause of \bullet . (Of course, nothing causes \bigcirc because \bigcirc no longer occurs.) With this in mind, consider a situation in which both A and B are connected to the monitor and in which a bullseye \odot appears at the center of the screen instead of \bullet or \bigcirc alone. The intention is that the appearance of \odot is a hybrid event consisting of two essential constituents: an appearance of \bigcirc and an appearance of \bullet . The intuitive causal picture is then as follows (with the double-headed arrow for essential-constituent causation).



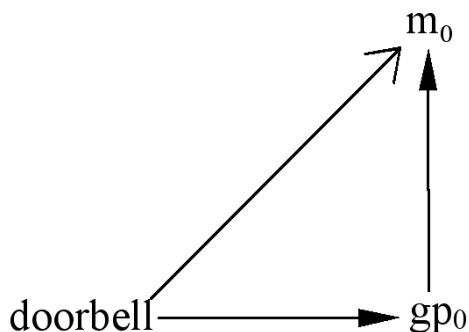
As before, b_2 is the proximal cause of O , and a_2 is its distal cause. And, as before, a_2 is the proximal cause of \bullet . Of course, b_2 does not cause \bullet , either proximally or distally. (b_2 is not a proximal cause of \bullet since a_2 is *the* proximal cause of \bullet . Nor does \bullet have b_2 as a distal cause: since a_2 is the proximal cause of b_2 and not the other way round, b_2 cannot cause \bullet *via* a_2 and thereby qualify as a distal cause of \bullet .) With b_2 ruled out, there is no other relevant cause of \bullet (from temporal distance t_2) besides a_2 . And we have just seen that a_2 is a cause of both \bullet and O (the proximal cause of \bullet and the distal cause of O). Thus, a_2 is the only event that is a cause of both \bullet and O (from temporal distance t_2). At the same time, a_2 is neither a joint nor an overdetermining cause of either \bullet or O . That is, a_2 does not (together with some other event) either jointly cause or causally overdetermine either \bullet or O . (On the one hand, a_2 is neither a joint nor overdetermining cause of \bullet because it is the proximal cause of \bullet , and neither b_2 nor any other relevant event is a cause of \bullet , proximally or distally. On the other hand, a_2 is neither a joint nor overdetermining cause of O : a_2 and b_2 , which are the only relevant candidates, are not in competition because a_2 is the distal cause and b_2 the proximal cause of O ; but only if they are in competition can they jointly cause or overdetermine O .)

Thus, we have the following: (i) \bullet and O are the essential constituents of \odot ; (ii) a_2 is the only event that is a cause of both \bullet and O (from temporal distance t_2); (iii) a_2 is neither a joint nor an overdetermining cause of either \bullet or O . Given this, we may infer that a_2 is *the* cause of \odot (from temporal distance t_2). Not only is this inference intuitive in its own right (surely the person on the

street would say that a_2 is the cause), it is validated by an intuitively compelling general principle: if (i) events e_1 and e_2 are the essential constituents of e_3 and (ii) c is the only event that is a cause of both e_1 and e_2 (from a given temporal distance), and (iii) c is neither a joint nor an overdetermining cause of either e_1 or e_2 , then c is *the* cause of e_3 .⁴⁷ Let us call causation of this sort *essential-constituent causation* and this principle, *the principle of essential-constituent causation*.

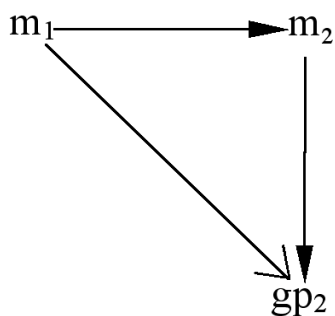
The idea is that the above pattern of causes mirrors the pattern of causes in the cases of mental-to-physical causation with which we are presently concerned. In particular, the relation between our decisions and our subsequent intentional actions is essential-constituent causation. I will be guided by this idea in what follows.

The first step concerns the proximal/distal distinction as it arises in the setting of physical-to-mental causation. Consider the analogy to sense-perception. Suppose that, prior to forming the derived desire to press the doorbell (m_1), it appeared to x that there was a doorbell in plain view and, given the psychophysical biconditionals, that ‘There is a doorbell in plain view’ was tokened in the Appearance Box in x ’s body. Both of these events—the appearance and the tokening—were caused by there being a doorbell in plain view.



Of course, the latter event caused the appearance *via* the tokening. In other words, there being a doorbell in plain view was the distal cause of the appearance, and this is so because it was the proximal cause of the tokening and the tokening was the proximal cause of the appearance.⁴⁸ Now what causes x to decide to press the doorbell (m_2) and the associated general brain event (gp_2)? Is it

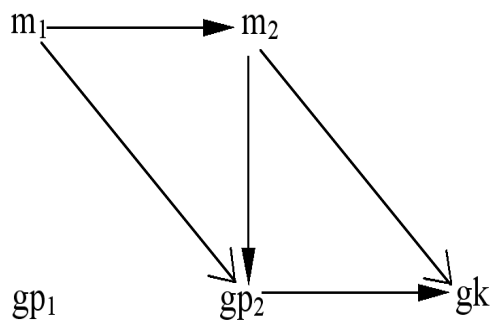
x's desire to press the doorbell (m_1), or is it one of the associated physical events (p_1 or gp_1)? The argument from the previous section shows *mutatis mutandis* that, in a competition with p_1 and gp_1 , m_1 prevails as the cause of both m_2 and gp_2 .⁴⁹ (I am again supposing that primary psychology includes a law to the effect that, if wm_1 occurs, so does m_2 .) Since p_1 , gp_1 , and m_1 are the only reasonable competitors for cause of m_2 and gp_2 (from their temporal distance, t_1), we may conclude that m_1 is indeed the cause of m_2 and gp_2 . That is, x's desire to press the doorbell is the cause of both x's decision to press it and the associated tokening of 'Press it'. But note the parallelism between this case and the sense-perception case: just as there being a doorbell in plain view caused the appearance of the doorbell *via* the tokening, so m_1 causes gp_2 *via* m_2 . That is, m_1 is the distal



cause of gp_2 , and this is so because it is the proximal cause of m_1 and m_1 is the proximal cause of gp_1 .⁵⁰ This is the point I wanted to make. (Of course, we saw the analogous thing in the case of the wired-together computers a moment ago: a_1 caused b_2 *via* a_2 . On this score, then, the analogy between mental-to-physical causation and these wired-together computers is intact.)

Now we come to the mental causation of physical behavior. Suppose for a moment that m_2 , gp_2 , and p_2 are in ordinary competition for being the cause of the general doorbell pressing motion (gk). Then, if we apply our standard test, we reach the conclusion that gp_2 is the cause. (Consider m_2 versus gp_2 : wm_2 in absence of gp_2 does not lead to gk , whereas wgp_2 in absence of m_2 does, so gp_2 prevails. Next consider gp_2 versus p_2 : wgp_2 in absence of p_2 leads to gk , but, since wp_2 in

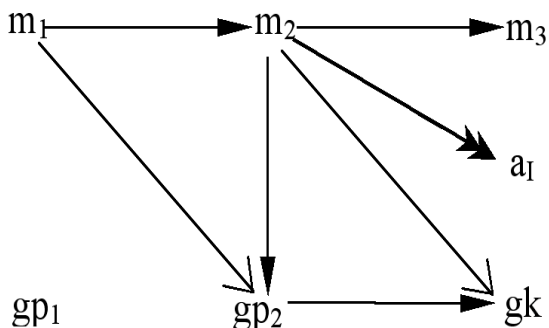
absence of gp_2 is not possible— p_2 is a determination of gp_2 —the other half of the test is not applicable. Thus, gp_2 prevails once again.) But are m_2 and gp_2 genuine competitors? gp_2 is intuitively the proximal cause of gk (just as, in the sense perception analogy, there being a doorbell in plain view was the proximal cause of the tokening of ‘There is a doorbell in plain view’). And we just saw that m_2 is the proximal cause of gp_2 . Therefore, m_2 is the distal cause of gk (just as, in the sense perception analogy, there being a doorbell in plain view was the distal cause of its appearing to x that there was a doorbell in plain view). So m_2 and gp_2 are not genuine competitors (unlike, e.g., cigarettes or the asbestos as the cause of the illness; the bullet or the poison as the cause of the death; the major’s shouting or the sergeant’s shouting as the cause of the advancing, etc.).



The initial temptation to take them to be genuine competitors was mistaken: the fact is that *each one* causes gk —one distally and the other proximally. We have thus arrived at our first goal: since m_2 is the distal cause of gk , mental causation of this general type of physical behavior is vindicated.

We come, finally, to the mental causation of intentional action. It is here that the recent wired-together computers analogy and the notion of essential-constituent causation come to bear. To begin with, the reasoning from the previous section adapts *mutatis mutandis* to show that m_2 prevails over gp_2 (and p_2) as the proximal cause of m_3 (i.e., the decision, not the general brain event or the specific brain event, is the cause of the ensuing concurrent intention to be pressing the doorbell). Furthermore, gp_2 is neither a proximal nor a distal cause of m_3 . (gp_2 is not a proximal cause of m_3 since m_2 is *the* proximal cause of m_3 . Nor is gp_2 a distal cause of m_3 : since m_2 is the

proximal cause of gp_2 and not the other way round, gp_2 cannot cause m_3 *via* m_2 and thereby qualify as a distal cause of m_2 .) Put another way, since gp_2 is neither a proximal nor a distal cause of m_3 , it in no way counts as a cause of m_3 . With gp_2 ruled out, there is no other relevant cause of m_3 (from temporal distance t_2) besides m_2 . In fact, not only is m_2 a cause of m_3 , it is also a cause of gk , we saw in the previous paragraph. Thus, m_2 is the only event that is a cause of both m_3 and gk (from temporal distance t_2). But m_3 and gk are the essential constituents of a_I . That is, the concurrent intention and the general kinetic event are the essential constituents of the intentional action. At the same time, our reasoning in the wired-together computer case adapts *mutatis mutandis* to show that (since m_2 is the proximal cause of m_3 and gp_2 is in no way a cause of m_3 and since m_2 is the distal cause of gk and gp_2 is the proximal cause of gk) m_2 is neither a joint nor an overdetermining cause of either m_3 or gk .



Hence, we arrive at the following conclusions: (i) m_3 and gk are the essential constituents of a_I ; (ii) m_2 is the only event that is a cause of both m_3 and gk (from temporal distance t_2); (iii) m_2 is neither a joint nor an overdetermining cause of either m_3 or gk . It follows by the principle of essential-constituent causation (and it is intuitive in its own right) that m_2 is the cause of a_I . That is, the decision is the cause of the intentional action.

6. Mental Causation in a Probabilistic Setting

Heretofore we have been operating under the simplifying assumption that the physical laws and psychological principles relevant to mental causation are deterministic. Specifically, we have

supposed that the physical laws backing physical-to-physical transitions of the w_{p_0} -to- p_1 type and the w_{p_1} -to- k type (and w_{gp_0} -to- p_1 type and w_{gp_1} -to- k type) are deterministic. Likewise, for the general principles of primary psychology backing w_{m_i} -to- m_{i+1} type transitions. In the present section, we suspend this assumption. In the case of psychology, most people believe this is realistic, and they think the principles of psychology are typically probabilistic (except perhaps those covering mental-to-mental transitions in cases where cognitive conditions are effectively ideal). In this connection, we will assume that the principles of primary psychology have the form of conditional probability statements: the probability of m_{i+1} given $w_{m_i} = s$. That is, $\Pr[m_{i+1} | w_{m_i}] = s$. We allow that s might be a vague probability, for example, “high,” “very high,” “extremely high,” “effectively 1,” “low,” “very low,” “effectively 0.” (In subsection (b) below we will consider what happens when ‘=’ is replaced with ‘ \geq ’, ‘ $>$ ’, ‘ \leq ’, or ‘ $<$.’.)

In this setting, the thesis that the principles of primary psychology are necessary is understood as saying that, for *all* auxiliary conditions γ (where w_{m_i} & γ is logically independent of m_{i+1}), it is necessary that, if the probability of w_{m_i} & γ is nonzero, the probability of m_{i+1} given w_{m_i} & γ is the same as the probability of m_{i+1} given simply w_{m_i} . So, for example, suppose that the primary psychological law tells us that the probability of m_{i+1} given $w_{m_i} = n$. And suppose that w_{m_i} & γ is logically independent of m_{i+1} . Then, in every world in which $\Pr[w_{m_i} \& \gamma] \neq 0$: $\Pr[m_{i+1} | w_{m_i} \& \gamma] = n$.⁵¹ What follows turns on this invariance.

(a) *Suspending the assumption of determinism.* The probabilistic counterparts of our two-part test turn out much the same as it did in the deterministic setting. Let us suppose that it is a principle of primary psychology that the probability of m_2 given $w_{m_1}(A) = n$. That is, $\Pr[m_2 | w_{m_1}(A)] = n$. The test then goes as follows.

Part (i). Consider the wide event $w_{m_1}(A)$ associated with $m_1(A)$ —namely, x ’s thinking that A in generally good cognitive conditions (intelligence, attentiveness, memory, etc.) and with relevant auxiliary cognitive contents (being aware of thinking that A , trying to prove $\neg\neg A$, seeming to remember having just inferred A from some prior thought, contemplating an inference from A to $\neg\neg A$, recognizing the validity of such an inference, etc.). Let ‘ B ’ be some relevant alternative to

‘A’. Choose a nearest world in which $wm_1(A)$ occurs and $gp_1('B')$ occurs instead of $gp_1('A')$. Our question concerns, not whether $m_2(\neg\neg A)$ will occur at t_2 , but rather the singular probability at t_1 that $m_2(\neg\neg A)$ will occur at t_2 . Specifically, the question is whether in our test world that singular probability suffers a big drop relative to the corresponding singular probability in the actual world (i.e., relative to n)?⁵² The answer is that it suffers no drop at all, for we are given that it is a principle of primary psychology that the probability of m_2 given $wm_1(A) = n$. Therefore, if $wm_1(A)$ & γ is logically independent of $m_2(\neg\neg A)$, then in every world in which $\Pr[wm_1(A) \& \gamma] \neq 0$:

$$\Pr[m_2(\neg\neg A) \mid wm_1(A) \& \gamma] = n.$$

Let β consist of all the concrete *physical* facts about our test world at t_1 (including, of course, the fact that $gp_1('B')$ occurs at t_1). Then, in every world in which $\Pr[wm_1(A) \& \gamma] \neq 0$:

$$\Pr[m_2(\neg\neg A) \mid wm_1(A) \& \beta] = n.$$

Relative to our chosen test world, the condition $wm_1(A) \& \beta$ invoked in this last conditional probability statement consists of *all* the concrete facts about the test world that have any bearing on the singular probability at t_1 of $m_2(\neg\neg A)$. So in effect this conditional probability statement tells us the singular probability at t_1 of $m_2(\neg\neg A)$ in the test world. That is, in this world, $\Pr_{t_1}[m_2(\neg\neg A)] = n$. Precisely the same probability as in the actual world. And the same thing holds for every other nearest “ wm_1 & $wgp_1('B')$ ” world if there is more than one. Finally, this result generalizes to most (or typical) relevant alternatives ‘B’ to ‘A.’ Thus, $m_1(A)$ passes the first half of our test.

Part (ii). For the various relevant alternatives B to A, we are to consider test worlds in which $wgp_1('A')$ occurs and in which $m_1(B)$ occurs instead of $m_1(A)$. For most (typical) B, is it the case that in at least some of the nearest “ $wgp_1('A')$ & $m_1(B)$ ” worlds the singular probability at t_1 of $m_2(\neg\neg A)$ suffers a big drop relative to the corresponding probability in the actual world? Yes. Choose a typical relevant alternative B. In at least some of the nearest “ $wgp_1('A')$ & $m_1(B)$ ” worlds, the new wide mental event $wm_1(B)$ that occurs instead of the actual-world wide event $wm_1(A)$ involves generally good cognitive conditions on x 's part and auxiliary cognitive contents that are B-ish rather than A-ish. (Call such worlds “B-ish test worlds.”) This assessment is reached by pretty much the same reasoning as in §3. And (again by the reasoning in §3) there will

be a principle of primary psychology applicable to this wide event $wm_1(B)$. In §3, however, we were making the simplifying assumption that the principles of primary psychology are deterministic. Given this assumption, the indicated deterministic principle would have told us that, in these B-ish test worlds, the probability at t_1 of $m_2(\neg \rightarrow B)$ is identical to *one*. For typical B, this result implies that in these B-ish test worlds the probability at t_1 of $m_2(\neg \rightarrow A)$ is *zero*. That is, this probability suffered the *maximum* possible drop it could undergo. Why is this? Because in generally good cognitive conditions and with B-ish auxiliary cognitive contents, it would be simply absurd for x to infer $\neg \rightarrow A$ from the altogether unrelated proposition B. Now, in our present probabilistic setting, we can combine this §3 reasoning with the kind of probabilistic reasoning just given in part (i) to get the result we are seeking. Namely, relative to the indicated B-ish test worlds, the probability at t_1 of $m_2(\neg \rightarrow A)$ is *extremely* low (indeed, as x 's cognitive conditions go up, its value approaches 0 as a limit). The underlying intuitive idea, once again, is that in generally good cognitive conditions and with B-ish auxiliary cognitive contents, it would be extremely unlikely for x to infer $\neg \rightarrow A$ from the altogether unrelated proposition B. The fact that the relevant (probabilistic) psychological principle holds necessarily ensures that this is so in each B-ish test world. For instance, in the logic problem example, this law tell us that, for all allowable γ :

$$\Pr[m_2(\neg \rightarrow A) \mid wm_1(B) \ \& \ \gamma] = k.$$

where k is some extremely low probability. Now let α consist of all the concrete physical facts about the test world at t_1 (including the fact that $wgp_1('A')$ occurs at t_1). Then:

$$\Pr[m_2(\neg \rightarrow A) \mid wm_1(B) \ \& \ \alpha] = k.$$

(Recall that k may be a vague probability—“extremely low,” “virtually nil,” etc.) Since $wm_1(B)$ consists of every concrete mental fact relevant to the logic problem, the condition $wm_1(B) \ \& \ \alpha$ invoked in this last conditional probability statement consists of *all* the concrete facts about the test world that have any bearing on the singular probability at t_1 of $m_2(\neg \rightarrow A)$ in the test world. So in effect the conditional probability statement just tells us the singular probability at t_1 of $m_2(\neg \rightarrow A)$. That is, in the test world, $\Pr_{t_1}[m_2(\neg \rightarrow A)] = k$. Thus, the singular probability in the test world suffer an extremely big drop from what it was in the actual world.

When we combine this outcome with the outcome of part (i), we see that $m_1(A)$ prevails over $gp_1('A')$ as the cause of $m_2(\neg\neg A)$. So, given that $m_1(A)$ and $gp_1('A')$ are the two best competitors for being the cause of $m_2(\neg\neg A)$, $m_1(A)$ is correctly identified as the cause. Moreover, the same reasoning extends *mutatis mutandis* in the case of mental-to-physical causation in the probabilistic setting.

(b) Physically boosted psychological probabilities. In the preceding subsection, we assumed that the conditional probabilities involved in the principles of primary psychology take a specific (though perhaps vague) value rather than merely setting lower or upper bounds for such values. In this subsection we consider what happens when this assumption is suspended and lower bounds or upper bounds are employed instead—that is, when the conditional probabilities are not equal to some unique value, but rather are: greater than n , greater than or equal to n , less than n , or less than or equal to n (for appropriate n). The resulting principles take the following form:

$$\Pr(m_{i+1} \mid wm_i) \geq n.$$

$$\Pr(m_{i+1} \mid wm_i) \leq m.$$

When the probability values of the psychological laws are relaxed in this manner, a significant new possibility is opened up. Namely, the associated conditional probabilities in the actual world may have some value greater than n or less than m . For example, perhaps the following specific values hold in the actual world:

$$\Pr(m_{i+1} \mid wm_i) = n^+.$$

$$\Pr(m_{i+1} \mid wm_i) = m^-.$$

for some specific $n^+ > n$ and $m^- < m$. (Like n and m , n^+ and m^- may be vague.) With respect to the first of these two conditional probabilities, let us say that the actual-world value n^+ is *boosted* from the minimum value n required by primary psychology or, more simply, just that the value is *boosted*.⁵³ One way this could happen is that the boosted value might be determined by a contingent basic law of psychology. In this case, the above account of mental causation would go through substantially unchanged. But another way this could happen is that the boosted value might be *derived* from the presence of contingent basic laws of physics (together with the relevant

psychological biconditionals). In this event, the foregoing account needs to be elaborated in a certain way. This may be done in two steps corresponding to the two parts of the test.

Part (i). Let 'B' be a relevant alternative to 'A'. Consider the nearest " $w_{m_1(A)} \& gp_1('B')$ " world(s). By the reasoning in part (i) of subsection (a), we know that, relative to the indicated world(s), the singular probability at t_1 of $m_2(\neg \neg A)$ might drop from n^+ to as low as n , but no lower. Does this constitute a big drop? In order to decide this question, let us first consider the second part of the test.

Part (ii). Let B be a relevant alternative to A. Consider the nearest " $w_{gp_1('A')} \& m_1(B)$ " world(s). By the reasoning in part (ii) of subsection (a), we know that, relative to the indicated world(s), the singular probability at t_1 of $m_2(\neg \neg A)$ must drop from m to some *extremely* low value (indeed, as x 's cognitive conditions go up, its value approaches 0). Once again, the intuitive idea is that in the envisaged circumstance it would be utterly absurd of x to infer $\neg \neg A$ from a wholly unrelated proposition B. And the fact that the relevant (probabilistic) psychological principle holds necessarily ensures that this is so in the world(s) under consideration. Clearly in this case the drop in probability approaches the maximum drop possible. Thus, our answer in part (ii) is that the result is an extremely big drop in probability.

With this answer in hand, we may return to the question in part (i): Does the drop from n^+ to n constitute a big drop? No. For, relative to the standard set by the extremely big probability drop incurred by its sole competitor in part (ii), the drop from n^+ to n is comparatively small. Moreover, since the same pair of conclusions is reached for most (or typical) alternatives 'B' to 'A' and B to A, $m_1(A)$ prevails over $gp_1('A')$ as the cause of $m_2(\neg \neg A)$. The conclusion is that, relaxing the probability values of the psychological laws so that they merely specify lower or upper bounds, in no way threatens mental-to-mental causation. Moreover, analogous reasoning will also allow us to vindicate mental-to-physical causation in this setting as well. (By extending this reasoning, it can also be shown that the account also holds if, in the worlds invoked in our test, the psychophysical biconditionals are probabilistic, not deterministic; see Appendix.)

(c) *Physical boosts from a lower bound of zero.* The preceding subsection suggests an important limiting case. Should we accept every casual comment ordinary people make about mental causation? Certainly not. There had better be room for the scientific discovery that some ordinary attributions of mental causes are mistaken. For example, suppose ordinary people commonly say that sensing blue causes one to have a feeling of relaxation. And suppose empirical psychologists confirm that, as a matter of fact, whenever human beings sense blue for a certain length of time, a feeling of relaxation is very likely to ensue (and suppose that no other colors have this effect). Suppose, moreover, that empirical psychologists, in collaboration with physical theorists, show that this fact is derivable ultimately from physical law together with the psychophysical biconditionals. And let us accept the obvious fact that the sensing-blue/relaxation correlation is no part of primary psychology—that is, primary psychology assigns a lower bound of *zero* to the probability of having the feeling of relaxation, given that one senses blue. Would ordinary people have been right in saying that sensing blue causes relaxation? Not according to our account. Instead, the cause would be a corresponding brain event. My intuition is that this is exactly right, and this is so regardless of what ordinary people happen to say. Otherwise there would never be any room for the scientific discovery that ordinary attributions of mental causes are sometimes mistaken. In such a dialectical situation, I can see no rational grounds for denying this conclusion. What we have is just one more collision of prescientific judgment and empirical science. Absent an argument to the contrary, science wins. After all, there was never a reason to think that epiphenomenalist claims were mistaken in *every* case.

NOTES

¹This formulation of the question is pretty much Stephen Yablo's ("Mental Causation," *Philosophical Review* 101, 2, April 1992: 245-80) and Jaegwon Kim's (e.g., "The Mind-Body Problem: Taking Stock after Forty Years," *Philosophical Perspectives* 11, 1997: 185-207). Yablo's splendid paper played an important role in the development of my own account.

²The notion of trumping preemption is Jonathan Schaffer's; see his "Trumping Preemption" (*Journal of Philosophy* 97, 2000: 165-81).

³Stephen Yablo (op.cit.) does not himself give such an argument, but someone might try to use his account for this purpose. Originally, I thought that, by relativizing determination relations to background physical laws or by incorporating them into "megawide" events, one could adapt Yablo's account to satisfy this secondary goal, but I found certain difficulties with this project, which I now believe apply to Yablo's account itself (see "Yablo on Mental Causation" in my *Integrity of Mind*, forthcoming). If this is right, the present account should be of particular interest to advocates of supervenience precisely because it provides them with a way to avoid epiphenomenalism that is compatible with their view (just as Yablo's account promised to do).

⁴Paul Boghossian, "Blind Reasoning," *The Aristotelian Society* supplementary volume 77, 2003: 225-248. Timothy Williamson, "Understanding and Inference," *The Aristotelian Society* supplementary volume 77, 2003: 249-93.

⁵For "Australian functionalism," see, e.g., David Lewis, "A Defense of the Identity Theory" (*Journal of Philosophy* 63, 1966: 17-25). For "American functionalism," see, e.g., Sydney Shoemaker, "Some Varieties of Functionalism" (*Philosophical Topics* 12, 1981: 83-118).

⁶See my "Mental Properties," *Journal of Philosophy* 91, 1994: 185-208.

⁷For this argument, see my "Self-Consciousness," *Philosophical Review* 106, 1997: 69-117. The only way for a functional definition to avoid this problem of unwanted content is for the psychological theory upon which the definition is based to be an implicit definition of the standard mental properties—that is, this psychological theory must be sufficiently strong that it is *uniquely* satisfied by the standard mental properties.

⁸Sydney Shoemaker, for example, has abandoned reductive functionalism in favor of nonreductive functionalism. The Self-consciousness Argument (ibid.) provides one of his reasons. His other reason has to do with mental causation: if reductive functionalism were correct, physical realizer events would supplant mental events as causes of our thoughts and actions. For these two reasons, Shoemaker has now abandoned reductive functionalism in favor of nonreductive functionalism. See Shoemaker, "Realization and Mental Causation," *Physicalism and Its Discontents*, Barry Loewer and Carl Gillette (eds.), Cambridge: Cambridge University Press, 2001.

⁹For example, Jaegwon Kim, "The Mind-Body Problem: Taking Stock after Forty Years."

¹⁰By background conditions I mean such things as standard temperature and pressure. This is the ordinary notion, according to which laws themselves (laws of physics, etc.) are not genuine background conditions.

¹¹Cited by Schaffer, *ibid.* For some people, the most convincing cases are those that involve laws of nature rather than conventional laws (e.g., military laws).

¹²That is, suppose, as before, that we have applied various other reliable tests for causes of e, and c and d are the only events that have passed all of them.

¹³The phrase ‘world(s)’ is to be understood in the obvious way. Each $d(\delta')$ determines a class of nearest worlds having the indicated features (i.e., the class of the nearest worlds in which b and $c(\chi)$ occur and $d(\delta')$ occurs instead of $d(\delta)$). Condition (i) requires that, for each $d(\delta')$, if the associated class of worlds contains exactly one world, e occurs in that world, and if this class contains more than one world, e occurs in each of them.

¹⁴As before, each $c(\chi')$ determines a class of nearest worlds having the indicated features (i.e., the class of the nearest worlds in which b and $d(\delta)$ occur and $c(\chi')$ occurs instead of $c(\chi)$). Condition (ii) requires that, for most $c(\chi')$, if the associated class of worlds contains exactly one world, e does not occur in it, and if this class contains more than one world, e does not occur in all of them.

¹⁵This test bears some resemblance to Lewis’s revised analysis of causation in “Causation as Influence” (*Journal of Philosophy* 97, 2000: 182-97) but was arrived at independently while grappling with certain difficulties I find in Stephen Yablo’s account (see note 3 above).

¹⁶In this version major/sergeant case it is understood that the sergeant and major are not coordinating their orders in any way. So understood, this case is a counterexample to Lewis’s original counterfactual account of causation: neither the major’s shouting nor the sergeant’s passes Lewis’s test; accordingly, the two shouts are wrongly judged to be joint causes when in fact the major’s shout is the sole cause of the advance. There are some variants of the example in which their shouting *is* coordinated and which are also counterexamples to Lewis’s test. (1) The major is strongly disposed to shout some order *iff* the sergeant does (and the sergeant is strongly disposed to shout some order *iff* the major does). (2) The major is strongly disposed to shout some order *only if* the sergeant does; but the sergeant is not strongly disposed to shout an order *only if* the major does. In variant (1), each shout passes Lewis’s test. Thus, the test yields the result that the two shouts overdetermine the advance when in fact it is the major’s shout alone that causes it. In variant (2), the major’s shout fails Lewis’s test whereas the sergeant’s passes it. Thus, the test yields the result that the sergeant’s shout is the cause of the advance when in fact it is once again the major’s shout that is the cause. As I have said, mental causation is a species of trumping preemption. Since the mental event and its correlated physical event are strongly disposed to occur together, mental causation will turn out to belong to the same species of trumping preemption as variant (1). (Lewis’s test of course yields the incorrect result that the mental and physical event overdetermine the effect, when in fact the mental event is the cause, or so we will show.)

¹⁷I put it this way because, other things being equal, it is desirable to have an account which is neutral on whether people are actually identical to their bodies or whether they merely *have* them. See Shoemaker “The Mind-Body Problem” and my “The Mind-Body Problem.”

¹⁸I invoke the framework of language-of-thought functionalism (Mentalese, Belief Boxes, etc.) for heuristic purposes only. In the eventual analysis it may be eliminated in favor of a more neutral formulation, and by relaxing certain details, we can arrive at formulations that mesh with various connectionist architectures as well.

¹⁹For example, b_{m_1} includes the fact that x is aware that he is thinking that A , and this—at least together with information about x ’s auxiliary cognitive conditions (intelligence, attentiveness, etc.)—plausibly requires x to be thinking that A .

²⁰As before, laws (vs. such things as standard temperature and pressure) are not genuine background conditions (see note 10) and so are not constituents of wide events in our sense. (For the same reason, ad hoc dispositional properties that merely code up laws of nature—e.g., the property of being a body such that $f = ma$ —are not constituents of wide events.) Such “megawide events” would trivialize the test by making the target effect an outright *logical consequence* of

megawide “causes.”

²¹It is commonplace that the antecedent of a diachronic law include restrictions on diachronic background conditions. For example, given the metaphysical possibility that the world suddenly cease to exist, a successful diachronic law typically includes (explicitly or implicitly) the condition that the world does not cease to exist at any time within the interval with which the law is concerned. Laws of psychology are a case in point. The assumption in the text is just that there is a wide event $wm_1(A)$ associated with the antecedent of such a psychological law. In the ensuing account, such wide events could be dropped in favor of a pair of entities: the narrow event $m_1(A)$ and a diachronic state of affairs s in which b prevails throughout the relevant period. (Incidentally, I believe that in the probabilistic setting of §6 the assumption about diachronic background conditions may be dropped.)

²²I am for now making the additional simplifying assumption that the indicated types of mental-to-mental transition are instances of psychological laws that belong to, or are entailed by, this general psychological theory A . In §6.c this assumption is dropped.

²³There might be more than one psychological theory that provides the basis for counterexample-free nonreductive definitions. If so, a candidate definition will count as a genuine definition iff it incorporates some *minimal* set of psychological laws sufficient making the definition counterexample-free. I hypothesize that every psychological law needed for my account of mental causation is included in at least one such minimal set and, in turn, is necessary.

²⁴Strictly speaking, nonreductive functional definitions need not be successful; it will be enough that the laws governing the envisaged sort of psychology are intuitively necessary (at least when all the relevant qualifiers are in place—‘*ceteris paribus*’, ‘psychologically normal’, ‘normal cognitive conditions’, ‘ideal cognitive conditions’, and the like). This special intuitive status of various fundamental psychological principles has animated philosophical psychology from Plato and Aristotle, through Descartes, and on to contemporary analytical functionalists.

Note that one can always construct artificial Cambridge properties, and associated Cambridge events, that necessitate one another (thereby creating the false appearance of their having genuinely lawful necessary relations to one another); in our context we are supposing that clear-cut Cambridge events have already been dispensed with. Primary psychological properties, by contrast, are genuine natural, non-Cambridge properties (likewise, for the associated events and laws). Not only is this intuitively compelling; it is a near corollary of our assumption at the outset that psychological properties are in no way reducible to physical properties. In any case, for dialectical purposes, we need only the weak premise that mental properties are not *clear-cut* Cambridge properties; this is undeniable.

²⁵Of course, if $m_1(A)$ supervenes on $wp_1('A')$, the second half of our test is inapplicable: it would be metaphysically impossible for $wp_1('A')$ to occur in the absence of $m_1(A)$. In this case, the first half of our test suffices to show that $m_1(A)$ is the cause of $m_2(\neg\neg A)$. By the way, in §4 I will explain why even materialists should deny the present case of supervenience, so they too would need to consider part (ii) of the test.

²⁶Similarly, $wm_1(A)$'s auxiliary cognitive contents would also include such things as: trying to prove $\neg\neg A$, seeming to remember having just inferred A from some prior thoughts, contemplating an inference from A to $\neg\neg A$, recognizing the validity of such an inference, etc. These would all have to become B-ish in character as well.

²⁷I have the intuition that it is possible for a being u to have body v even if there were isolated psychophysical divergences—rogue tokenings or rogue thoughts. But it is still the case that *nearly*

all these psychophysical biconditionals must hold. (This is important, for the worlds contemplated in the two parts of our test involve rogue tokenings or rogue thoughts at t_1 .) Note, also, that other refinements might be needed to accommodate various *recherché* questions (e.g., whether *u* can have two bodies or whether *u* and *u'* can share a body). See my “The Mind-Body Problem.” Another refinement concerns the substitution of ‘causally necessary’ for ‘nomologically necessary’ throughout. Ultimately, I prefer this formulation (e.g., so that it would not be a contradiction to allow disembodiable beings whose thoughts have physical effects but only when these beings are embodied). When this and the other refinements are adopted, our account would need to be adjusted accordingly but would remain substantially the same.

²⁸Such an analysis would also provide a kind of explanation of why there should be correlations between our mental and physical properties: the existence of such correlations is an immediate consequence of the fact that we have bodies. One would be free, in turn, to explain this latter fact by positing a law that every being with mental properties has a body (and always the same body). Similarly, one would be free to posit a law that every body suitable for being the body of a being with mental properties is the body of a unique being with mental properties. The resulting picture is one that materialists without metaphysical axes to grind should be happy to embrace.

²⁹Analogously, in the case of Mentalese expressions for properties and relations (vs. propositions). Note that I am continuing to use the token-in-a-box idiom for heuristic purposes. Ultimately, it can be bypassed, and the physical types *s* can simply be physical properties (narrow or wide) of body *v*.

³⁰If there is more than one such function, let *c* be the union of them; that is, *c* is the maximal such function.

³¹As I understand them, various contemporary materialists think that events like *gp1* promise to play a role in higher-level neuroscience. When Noam Chomsky and John Searle tell us that in psychology and philosophy of mind, our focus should be on biological events at a higher level of abstraction, I think they have in mind something like general brain events.

³²For brevity I will henceforth omit ‘at least’ from the phrase ‘at least nomological necessity’.

³³Suppose that the psychophysical laws were indeed necessary. In that case, they would have to be basic as well. (For if they were derived, they would have to be derived from other laws that are both basic and necessary—either basic necessary psychological laws or basic necessary physical laws, or both. But basic physical laws are contingent, not necessary and so cannot play a role in the envisaged derivation. And the hypothesized necessary psychophysical laws cannot be derived from basic necessary psychological laws alone.) Thus, since the laws of primary psychology would also be both necessary and basic, the psychophysical laws and these psychological laws would have the power to trump the merely contingent basic physical laws. Of course, given the necessity primary psychology and the hypothesized necessity of the psychophysical biconditionals, there would be various *derived* physical-to-physical laws that are necessary. But in their role of explaining the occurrence of physical effects relevant to mental causation, necessary laws that are basic trump necessary laws that are only derived. So in this role, the envisaged derived physical laws would be trumped by the indicated basic laws of psychology and basic psychophysical laws.

I believe that there are very good arguments against the strong necessitarian view of laws (i.e., that all laws—including physical laws—are metaphysically necessary). On the supposition that this view is correct, however, the account of mental causation I am proposing would not be correct. In spite of this, the account of the laws of primary psychology would remain; and, given this, we could show that mental causation would (at least) be a case of causal overdetermination. The resulting picture would resemble that of Paul Pietroski, *Causing Actions*, New York: Oxford University Press, 2000.

³⁴Terry Horgan, “Supervenience and Microphysics” (*Pacific Philosophical Quarterly* 63, 1982: 29-43). David Lewis, “New Work for a Theory of Universals” (*Australasian Journal of Philosophy* 61, 1983: 343-77). Frank Jackson, *From Metaphysics to Ethics*, Oxford: Oxford University Press, 1998.

³⁵Every nomologically possible world is by definition nonalien, but not conversely. There are nonalien worlds in which one or more the laws of nature fail. In other words, nonalien worlds are limited only in what natural properties are instantiated in them, not in what laws hold in them. The point of the notion is to rule out natural properties having to do with things such as ghosts, telekinetic forces, and other such nomological impossibilities (as we may suppose them to be).

³⁶See Jackson, p. 12 f.

³⁷By the laws of a world w , I will simply mean those statements that are nomologically necessary in w (i.e., that hold in all worlds nomologically accessible from w). By the physical laws of w , I will simply mean those physical statements that are nomologically necessary in w . By physically possible relative to w , I mean those worlds in which all of w 's physical laws hold.

³⁸To simplify the discussion, I will hereafter focus on the actual world, making the assumption that it is nonmagical (nontelekinetic, etc.). So when I speak of physically possible worlds and nomologically possible worlds, I will mean worlds that are physical and nomologically possible relative to the actual world. The account will generalize to other worlds in which nonmagical causation would be explicable in the same way.

³⁹Proof: Suppose for reductio that this is not so for some physically possible world w . Then, w 's complete physical description D is true in no nomologically possible world. Hence, not- D is true in every nomologically possible world and so would be a physical law (cf. note 37). But, since w is physically possible, all of the physical laws, now including not- D , would have to be true in w , contradicting the fact that D is true in w . So the reductio hypothesis fails.

⁴⁰Given this supervenience principle, the worlds in which the physical laws hold are a proper subset of the worlds in which the psychophysical laws hold. On every other supervenience principle we will consider, the same proper inclusion holds at least for all of the nearest worlds relevant to mental causation. This is what I meant when I said that the psychophysical biconditionals are such a rudimental feature of the world that they underwrite mental causation.

⁴¹What exactly is it for a law to “break”? The answer depends on what a law is. I need not take a stand on this, for as far as I can see the story in the text goes through on any credible answer.

⁴²This principle suggests a natural generalization: amongst the nonalien worlds, even if there are some that fail to be nomically supervenient, those closest to the actual world are all nomically supervenient. Although this supervenience principle is stronger than that just given in the text, it is still much weaker than nonalien metaphysical supervenience.

⁴³Note that part (ii) of the test is inapplicable because p_1 entails gp_1 and in that sense is a determination of gp_1 . Yablo's theory works very nicely here.

⁴⁴This framework also provides the tools for explicating explicate whether and why mental-to-physical causation is possible on various traditional views of the mind-body relation such as animism and parallelism.

⁴⁵In the primary cases, there must be a concurrent intention (or concurrent trying). Our ensuing remarks need to be adjusted somewhat to handle less primary cases, for example, time-lag cases

(e.g., murdering by means of a time-bomb). Note, too, that I will be suppressing subtleties concerning the difference between intending and trying.

⁴⁶If, as some people hold, actions are not events, this thesis is mistaken and what follows in the text should instead be taken to be about the events associated with actions rather than the actions themselves (e.g., the event of my pressing the doorbell instead of my action of pressing the doorbell). Though oversimplified, doing this should suffice for our purposes, for an action causes (or at least explains) another action if the associated events are causally related in a parallel fashion. For example, if the event of deciding to press the doorbell causes the event of pressing the doorbell, then the mental act of deciding to press it causes (or explains) the act of pressing it.

⁴⁷I think this principle is “objective” in the following sense. It holds in all contexts in which: (a) we are interested in isolating the cause of e_3 if there is one; (b) we have narrowed down the class of competitors to events occurring at a given temporal distance from e_3 , and (c) we have made it clear that the class of competitors is not to be narrowed down by any further pragmatic factors such as salience or our interests.

⁴⁸Of course, there being a doorbell in plain view is not the proximal cause of the appearance, for if it had been, we would have a kind of magical causation in our sense: it would require a special power to produce effects in consciousness without going by way of the body.

⁴⁹Does this create any special problems for free will? No. In the present context we are assuming determinism. So we have two possibilities: either compatibilism holds and there is free will possible, or incompatibilism holds and there is no free will. Either way, no additional problem is created by the conclusion that x 's desire causes x 's decision. Suppose, on the other hand, that determinism does not hold and (as will be the case in §6) laws are probabilistic. Then the situation is more complicated. On the account in §6, x 's desire still causes x 's decision. This, however, does not imply that x 's decision was not free. After all, given that the laws are probabilistic, x could have (compatibly with the laws and all the conditions through t_1) decided either way. Some people, however, might be inclined to hold that a decision is free only if it has no cause and, therefore, that on the §6 account x 's decision is not free. Still others might hold that x 's decision was not free simply because there is a law entailing that, given x 's desire (and relevant background conditions) it is probable that x make the decision. Even though I find both of these positions implausible, further discussion lies beyond the scope of this paper.

⁵⁰As we saw in note 48, there being a doorbell in plain view was the distal cause of the appearance that there was a doorbell in plain view, for if it had been the proximal cause, that would have been a kind of magical causation. Analogously, m_1 is the distal, not proximal, cause of gp_2 , for if it were instead the proximal cause, that too would be kind of magical causation, this time akin to telekinesis.

⁵¹Consider the analogy between the foregoing and the fact that

$$[wm_i \rightarrow m_{i+1}]$$

implies that, for all γ ,

$$[(wm_i \ \& \ \gamma) \rightarrow m_{i+1}].$$

⁵²I have adapted this idea from Lewis, p. 177 (Postscript B, “Causation,” *Philosophical Papers II*).

⁵³For present purposes we need not discuss the second of these two conditional probabilities.