In *Theoria* 16 (2001), 95-116, guest-edited by Josep Corbi. (The *Spanish* journal by that name.)

CAUSAL COMPATIBILISM AND THE EXCLUSION PROBLEM.

Terry Horgan

University of Memphis

In this paper I address the problem of causal exclusion, specifically as it arises for mental properties (although the scope of the discussion is more general, being applicable to other kinds of putatively causal properties that are not identical to narrowly physical causal properties, i.e., causal properties posited by physics). I summarize my own current position on the matter, and I offer a defense of this position. I draw upon and synthesize relevant discussions in various other papers of mine (some collaborative) that bear on this topic.[1]

In section 1 I describe the problem as I construe it, and some principal theoretical options for dealing with it. In section 2 I briefly summarize some observations by David Lewis about ways that many of our concepts, and the terms expressing them, are governed by implicit, contextually variable, discourse parameters; this is background for the discussion to follow. In section 3 I summarize my own approach to problem of causal exclusion, which incorporates the claim that the notions of causation and causal explanation are context-sensitive in a way that involves implicit parameters. In section 4 I defend my approach, arguing that it fares well in terms of overall theoretical costs and benefits.

1. The Problem of Causal Exclusion.

The problem can be put this way: Each of the following five statements is prima facie credible, and yet they are jointly inconsistent.

1. Physics is causally closed.
2. Mental properties are real, and are instantiated by humans.
3. Mental properties are causal properties.
4. Mental properties are not identical to physical causal properties.

5. If physics is causally closed, then all causal properties are physical causal properties.

Statement 1, the thesis of the causal closure of physics, is the claim that every physical event or state is completely causally determined—to the extent that it is causally determined at all—on the basis of physical laws plus prior physical events and states, and that the laws of physics are never violated. Statements 3 and 4 are to be understood as making conditional claims—claims about what mental properties are like, if there are any such properties and they are instantiated by humans. Statement 2, then, asserts the implicit antecedent of statements 3 and 4. By 'physical property' I mean, essentially, the kind of property posited in fundamental physical theory—i.e., a physics-level property.

Each of statements 1-5 has considerable prima facie plausibility. Let us consider them in turn. Statement 1 has enormously strong support on the basis of current scientific knowledge. In the case of humans, for instance, the bodily motion that constitutes action is all basically muscular contraction and relaxation, caused by physical activity in the central nervous system. Statements 2 and 3 are deeply embedded in our common-sense conceptual scheme and in our explanatory practices; they are claims we normally consider to be amply well warranted both by introspection and by the utility of common-sense mentalistic explanation.

Statement 4 is well warranted, inter alia, by virtue of considerations of multiple realization. It appears to be conceptually and nomologically possibile for mental properties to be realized by a multiplicity of different physical causal properties, either across different creature-kinds or even within a given creature-kind. So realization cannot be identity, because a single mental property cannot be identical to several distinct physical-realization properties.

Statement 5 can be defended by the following, initially very plausible-looking, reasoning. Physical causal properties evidently "do all the causal" work with respect to the generation of physical states and events, thereby apparently "excluding" non-physical properties from having any genuine causal role vis-à-vis physical states and events. Furthermore, since mental properties plausibly are supervenient on the physical and are realized by physical properties, ultimately what causes the instantiation of mental properties too is also physical: since the instantiation of a mental property always involves the

instantiation of some physical realizing-property, the physical cause(s) that generate the instantiation of the realizing-property thereby generate the instantiation of the mental property.

Another, closely related, line of reasoning in favor of claim 5 goes as follows. If indeed mental properties are distinct from physical properties and are causal properties, then either (a) certain states or events diachronically depend in part on the prior instantiation of some mental property (or properties) as a causally necessary condition of their occurrence, or else (b) certain states or events are causally overdetermined by both physical and mental sufficient conditions. Consider possibility (a). The states or events that supposedly depend partially on prior mental-property instantations as causally necessary conditions cannot be physical, because this would violate the causal closure of physics; nor can the effect-states and effect-events be mental, because the causal closure of physics guarantees that the instantiation of a mental property M (on a given occasion) depends causally only upon the physical cause(s) that generate the instantiation of whatever physical property realizes M (on that occasion). So possibility (a) is precluded. As for possibility (b), surely it is the case (given the causal closure of physics) that mental properties, if they are causal properties at all, can only be causally efficacious via the physical properties that realize them; but that is not real causal overdetermination after all, since there is no "independent causal route" leading from cause to effect. (Again we are back to the physical property "doing all the causal work.")[2]

Although each of statements 1-5 has substantial initial credibility, they are jointly inconsistent; so at least one of them must be false. Several potential philosophical positions can be indentified, each of which responds to this conundrum by rejecting one of the five statements and retaining the other four:

i. Causal emergentism. (Denies statement 1)

ii. Eliminativism. (Denies statement 2)

iii. Epiphenomenalism. (Denies statement 3)

iv. Identity materialism. (Denies statement 4)

v. Causal compatibilism. (Denies statement 5)

Causal emergentism construes mental properties as fundamental force-generating properties; they generate new forces over and above those generated by the causal properties of physics, so that the net force affecting the distribution of matter is different from the net physical force. This position saves mental causation at the price of denying the causal closure of physics. Eliminativism denies that there are genuine mental properties instantiated by humans; this position repudiates our ordinary notion of mentality altogether, and mental causation along with it. Epiphenomenalism denies that mental properties are really causal properties; this position retains mentality, but it abandons mental causation and mentalistic causal explanation as illusory. Identity materialism claims that mental causal properties are really just identical to certain physical causal properties; this position saves mental causation at the price of denying that mental properties are multiply-realizable properties.[3] Causal compatibilism claims that even though physics is causally closed, and even though mental properties are multiply realizable and hence are not identical to physical causal properties, mental properties are causal properties nonetheless. This position asserts that there is genuine causation and genuine causal explanation at multiple descriptive/ontological levels, and that despite the causal closure of physics, physics-level causal and causal-explanatory claims are not really incompatible with mentalistic causal and causal-explanatory claims.[4]

I am a causal compatibilist: I advocate repudiating statement 5 and retaining statements 1-4. I will offer an articulation of causal compatibilism that puts enough flesh on the bones of the abstractly described position to constitute a coherent, conceptually stable, version of the view. I will also offer an account of why and how the prima facie plausible reasoning in support of statement 5 is mistaken—an account that explains why this reasoning is so intuitively powerful (despite being in error). I will turn to these tasks in section 3, after first laying some groundwork in the next section.

2. Scorekeeping in a Language Game.

My version of causal compatibilism builds upon a general point about certain terms and concepts that is articulated, illustrated, and argued for by David Lewis (1973/1983): viz., that these terms and

concepts often are partially governed by certain implicit, context relative, parameters. These parameters are elements of what Lewis calls the "score in the language game." They include, for instance, presuppositions (e.g., that France presently has exactly one king); factors determining the referent, in context, of a given definite description; factors determining the standards for contextually correct applicability of vague terms like 'bald', 'flat' or 'hexagonal'; and contextually variable factors operative in modal and counterfactual discourse (e.g., factors that get formalized in possible-world semantics as the accessibility relation over possible worlds, and the similarity ordering over possible worlds).

Lewis makes three especially pertinent points about such implicit parameters. First, as competent thinkers and speakers we deal with them so naturally that we often do not even notice them. Take definite descriptions, for instance. Frequently, Lewis points out, more than one object within a contextually determined domain of discourse will be a potentially eligible referent of 'the F'. When this happens, the proper referent will be the most salient F in the domain, according to some contextually determined salience ranking. We take this implicit context relativity so much in stride that we often are not even aware of it. Lewis gives this example:

> Imagine yourself with me as I write these words. In the room is a cat, Bruce, who has been making himself very salient by dashing madly about. He is the only cat in the room, or in sight, or in earshot. I start to speak to you:
>
> > The cat is in the carton. The cat will never meet our other cat, because our other cat lives in New Zealand. Our New Zealand cat lives with the Cresswells. And there he'll stay, because Miriam would be sad if the cat went away.
>
> At first, "the cat" denotes Bruce, he being the most salient cat for reasons having nothing to do with the conversation. If I want to talk about Albert, our New Zealand cat, I have to say "our other cat" or "our New Zealand cat." But as I talk more and more about Albert, and not any more about Bruce, I raise Albert's salience by conversational means. Finally, in the last sentence of my monologue, I am in a position to say "the cat" and thereby denote not Bruce but rather the newly-more-salient Albert (1983, p. 241).

Second, implicit context-relative parameters frequently get altered through a process Lewis calls accommodation: something is said that requires some parameter to have a new value, in order for what is said to be true (or otherwise acceptable): so that parameter thereby takes on that new value. Concerning salience and definite descriptions, he says:

> One rule, among others, that governs the kinematics of salience is a rule of accommodation. Suppose my monologue has left Albert more salient than Bruce; but the next thing I say is "The cat is going to pounce on you!" . . . . What I have said requires for its acceptability that "the cat" denote Bruce, and hence that Bruce be once again more salient than Albert. If what I say requires that, then straightaway it is so (1983, p. 242).

Third, often if a context-relative parameter is one we would naturally think of as involving standards that can be either raised or lowered, then accommodating upward will seem more natural than accommodating downward. Concerning context-relative standards of precision for terms like 'hexagonal' and 'flat', for example, Lewis remarks:

> I take it that the rule of accommodation can go both ways. But for some reason raising the standards goes more smoothly than lowering. If the standards have been high, and something is said that is . . . [acceptable] . . . only under lowered standards, then indeed the standards are shifted down. But what is said . . . may seem only imperfectly acceptable. Raising our standards, on the other hand, manages to seem commendable even when we know that it interferes with our conversational purposes (1983, p. 245).

In my view, various philosophically interesting concepts are among those that are governed by implicit contextual parameters, and this fact is importantly involved in philosophical puzzles that arise in connection with those concepts. The concept of causation is one of these, as is the closely related concept of causal explanation. These are, as I will put it, contextually parameterized notions. This idea will figure centrally in what follows, as will the three points about contextual parameters lately emphasized.

3. Causal Compatibilism Articulated.

The version of causal compatibilism I favor rests upon certain claims about the notions of causation and causal explanation, including claims about contextual paramaterization. The claims certainly do not constitute a complete account of causation and causal explanation; but that is not needed, for present purposes. Various ways of developing more complete accounts would be compatible with what I say here.

3.1 Scorekeeping in the causal explanation game.

I advocate a construal of causation and causal explanation, with specific application to the case of mental causation and mentalistic causal explanation, that includes three central ideas. The first is a conception of causal-explanatory relevance for properties, involving systematic patterns of counterfactual dependence.

In causal explanation the effect phenomenon e, described as instantiating a phenonemon type E, is shown to depend in a certain way upon the cause phenomenon c, described as instantiating a phenomenon of type C. Often the dependence involves the fact that c and e are subsumable under a counterfactual-supporting generalization—either a generalization that directly links C to E, or else a more complicated generalization whose antecedent cites a combination of properties that includes C.[5] But in order for the cited properties C and E to be genuinely explanatorily relevant to the causal transaction between c and e, it is not enough that c caused e and c and e are subsumable under such a generalization. Rather, C and E must fit into a suitably rich pattern of counterfactual relations among properties.

It is important to understand how this feature is related to the structure of scientific laws. The generality of the fundamental laws of the natural sciences, for example, does not consist merely in their having the logical form, "All As are Bs." It consists, rather, in the fact that they are systematic in scope and structure, so that a wide range of phenomena are subsumable under relatively few laws. One major source of their systematicity is that (1) the laws cite determinable properties, namely magnitude-properties, where the determinants are quantitatively specific instances of these properties, and that (2) the

laws contain universal quantifiers ranging over these quantitative determinant-values (in addition to the universal quantifiers ranging over the non-numerical entities in the law's domain). Newtonian velocity, for example, is not a single determinate property but an infinite array of determinate properties, one for each real value of the determinable V. The resultant generality of a physical law consists largely in the existence of a whole (typically infinite) set of specific nomically true principles, each of which is a specific instantiation of the law with specific numerical values "plugged in" for the determinant-variables. Rich patterns of counterfactual dependence, of the sort that are a crucial feature of successful causal explanation in science, are reflected by the truth of such sets of specific law-instantiations.

Second: Often several distinct patterns of counterfactual dependence, all subsuming a single phenomenon, will involve different descriptive/ontological levels, for example microphysical, neurobiological, macrobiological, and psychological. Consider, for instance, instances of human behavior, vis-à-vis the level of common-sense intentional psychology, so-called folk psychology. There are robust patterns of counterfactual dependence among the state types (including act types) posited by folk psychology—patterns systematizable via generalizations containing universal quantifiers ranging over suitable determinant-values. These determinant-values are not quantitative, but instead are propositional (or intentional); i.e., they are the kinds typically specified by 'that'-clauses. Take, for instance, relations between actions and reasons. The intentional mental properties that constitute reasons (namely belief types, desire types, and other attitude types), in combination with act types, clearly figure in a rich and robust pattern of counterfactual dependence of actions upon reasons that rationalize them, a pattern conforming to the following generalization:

> For any subject S, desire-content D, and action A, if S wants D and S believes that doing A will bring about D, then ceteris paribus, S will do A.

There are also rich patterns of counterfactual dependence among folk psychological mental states themselves, again systematizable by suitable ceteris paribus generalizations involving quantification over propositional/intentional determinant-values. Wanting, believing, etc. figure in these generalizations as determinable properties, and the generalizations characterize vast (possibly infinite), highly structured,

counterfactual-dependence relations among the corresponding determinant properties—a different specific dependence relation for each specific instantiation of the propositional variables in the generalizations.

Third: The closely related concepts of causation and causal explanation are contextually parameterized notions, with an implicit contextual parameter keyed to a specific descriptive/ontological level; I will call this the level-parameter. The contextually relevant counterfactual-dependence patterns, for purposes of evaluating the truth or falsity of causal and causal-explanatory statements in specific contexts of usage, are those patterns that reside at the level determined by the contextually operative level-parameter.[6]

When we bring together the three key ideas just described, the following picture results. A single phenomenon can perfectly well be subject to a variety of different causal explanations, involving properties from a variety of different counterfactual-dependence patterns at different descriptive/ontological levels. Often various different causal and explanatory claims with respect to a given phenomenon, involving properties from various different descriptive/ontological levels, all will be objectively true, since each is grounded in some objective counterfactual-dependence pattern. But the different kinds of causal and causal-explanatory claims will be tethered to different contexts of causal inquiry—contexts in which the level-parameter has different settings, involving different kinds of objective counterfactual-dependence pattern.

Let me elaborate this picture. As I said, causal explanation typically involves fitting a given phenomenon into some pattern of counterfactual dependence—often a pattern systematizable by an associated generalization. Such patterns exist at each of the various levels corresponding to the various sciences. Which kinds of dependence patterns and generalizations are most germane typically will be a context-relative matter, governed largely by the interests of those doing the explaining and inquiring. Choice of descriptive vocabulary normally will have a very heavy influence on the default settings for the contextually relevant parameters, the operative "score" in the causal-explanation game. If we pose our questions and offer our answers in psychological vocabulary, for instance, then normally the relevant

patterns of counterfactual dependence will be ones involving psychological properties, with their associated generalizations—including the generalizations of folk psychology.

Sometimes, as I said, a single phenomenon is susceptible to causal explanations at several different levels of description. For instance, a piece of human behavior, described in terms of specific muscular contractions and relaxations, will be causally explainable (at least in principle) neurophysically; and that same phenomenon, described as an action, also might be explainable mentalistically.[7] These differing causal explanations are not in competition with one another; the neurophysical explanation does not "screen off" or "exclude" the psychological one. On the contrary: since there are robust patterns of counterfactual dependence at both levels of description, and since these patterns themselves are compatible with one another, different causal explanations can be given that fit the particular behavioral phenomenon into either pattern.

The compatibility of these different levels of explanation largely stems from inter-level supervenience relations. Since the higher-order, psychological, patterns and generalizations are supervenient upon underlying physical facts and laws, the mental properties that are causal properties at the psychological level have their causal efficacy via the causal efficacy of physical causal properties that realize them. The higher-order causal properties of psychology do not generate physical or mechanical forces over and above the physical forces produced by fundamental physical properties, and they do not intrude upon the causal-explanatory closure of physics vis-a-vis physical phenomena as physically described. Rather, mental properties causally explain certain phenomena in a way that is complementary to physical causal explanation, by fitting those phenomena into certain robust higher-order patterns of counterfactual dependence that conform to systematic, non-accidental, higher-order generalizations.

When one considers causal explanation in a detached philosophical way, it is appropriate simultaneously to describe natural properties at various different theoretical levels of description as all being "causal properties"; and I have been doing so in recent paragraphs. On the other hand, just as contextual parameters often determine which kind of causal explanation is appropriate in context, such parameters also typically govern the notion "causal property" itself. In context, the properties that count

as causal properties normally will be the ones that figure in the relevant kind of causal explanation. That is, from within an engaged perspective of causal-explanatory inquiry, the properties that qualify as causal will all fall within a contextually eligible range of candidates, as delimited by the current score in the causal-explanation game.

The fact that certain phenomena are susceptible to multiple causal explanations, involving natural properties from different levels in the hierarchy of the sciences, is not properly considered a matter of "causal overdetermination." Overdetermination is instead an intra-level notion, and is governed by the same contextually variable parameters that typically govern the notion of a causal property. The idea is essentially this: given the specific level of description that is contextually appropriate for causal explanation, several properties are co-instantiated, all at the relevant level of description, each of which is such that its instantiation is independently causally sufficient (in the circumstances) for the effect. I.e., even after we contextually fix the operative score in the causal-explanation game in a way that restricts the relevant patterns of counterfactual dependence to those involving properties at a specific level in the hierarchy of the sciences, we still find several properties instantiated that each figure in the contextually relevant dependence patterns in a way that makes each property-instantiation an independently causally sufficient condition for the effect. So mental properties and the neurophysical properties that realize them do not causally overdetermine their effects, because they figure in distinct counterfactual dependence patterns at distinct theoretical levels.

Since the notions of causal property and causal overdetermination are both governed by contextually variable parameters, the properties we may properly cite, when we are tallying an inventory of properties or factors that were causally operative with respect to a given phenomenon, fall within the range determined by the current score in the causal-explanation game. Mental properties fall within the contextually eligible range when the score is set for psychological explanation, whereas the neurophysical properties that realize them fall within the contextually eligible range when the score is set for neurophysical explanation. In a normal context of psychological explanation it is not appropriate to count the neurophysical realizers as causal properties in addition to the mental properties, whereas in a normal

context of neurophysical causal explanation it is not appropriate to count the mental properties as causal properties in addition to the neurophysical ones. In either context, such double-counting goes contrary to the contextually operative score in the causal-explanation game.

3.2. How causal exclusion reasoning goes wrong.

On this account, causal exclusion reasoning goes wrong because it wrongly treats the notions of causation and causal explanation a though they are not governed by an implicit level-parameter, when in fact they are. If one ignores the level-parameter (which is easy to do, since it is not explicit), then it will appear that properties that are either causal or non-causal, simpliciter—and thus that the causal closure of physics just leaves no room for other properties to be causal. In order to be causal, they either would have to be additional fundamental force-generating properties (going contrary to the causal closure of the physical), or would have to be overdetermining causal properties (going contrary to the lack of an independent causal route to the effect). But if indeed the notions of causation and causal explanation have an implicit level-parameter, then this is just the wrong picture of the matter. The basically mistaken idea is that properties are causal, or not causal, punkt. This is something like asking what time it is on earth, rather than asking what it is in a given time zone. Which properties count as causal depends upon the parameters governing engaged causal inquiry.

This is not causal irrealism, or explanatory irrealism. For, the relevant counterfactual dependence patterns are all objectively real. Given a contextually fixed value of the level-parameter, it is a perfectly objective matter that certain properties are causal, and that certain phenomena are causally explainable by appeal to the instantiation of those properties. Causation and causal explanation are perspectival and interest-relative, to be sure. But they are also objective, because they involve the way particular phenomena fit it into real, objective, patterns of counterfactual dependence.

Now admittedly, there is also a more detached perspective one can take (and often does take, in philosophical inquiry in which several pertinent levels of description are considered simultaneously). One can talk about causal properties at multiple levels of description, and about causal claims and causal

explanations that do not compete with one another. (I have been talking that way in this paper.) Such talk is legitimate too, on the view I am proposing, but it needs to be properly understood. Since the notions of causation and causal explanation have an implicit level-parameter, when one talks in this level-spanning way one's statement in effect ranges over different allowable settings of the level-parameter, and one's talk of properties at two different levels as both being causal in effect implicitly links their being causal to the respectively relevant parameter-settings. So suppose, for example, that I claim that a certain mental property M, and the physical property P that realizes M (on a particular occasion of instantiation), are both causal properties. My claim can be approximately paraphrased this way:

> There are settings S1 and S2 of the level-parameter such that physical property P is a causal property under S1 and mental property M is a causal property under S2.[8]

Since there is no such thing as a causal property simpliciter, on the view I am proposing, the claim in question does not assert that M and P are both causal properties simpliciter.

3.3. Causal exclusion as a cognitive illusion.

There is a strong and persistent intuition, I realize, to the effect that physical causal properties "screen off" or "exclude" higher-order properties from any genuine causal role. Insofar as one gives credence to this intuition, my scorekeeping story about the notions of causation and causal explanation is apt to seem like sophistry.

Although I think the intuition is mistaken, I acknowledge the burden of explaining its strength and persistence. In my view, the intuition is best explained as a subtle "cognitive illusion," analogous in some ways to perceptual illusions that persist even when knows they are illusory. The illusion arises because our cognitive mechanisms for accommodating implicit discourse-parameters get tripped up by the somewhat abnormal, level-spanning, philosophical mode of discourse. I will set forth the proposed explanation, and then return to the charge of sophistry.

The very posing of the causal-exclusion question creates an <u>atypical</u> discourse-context, one in which the implicit level-parameter tends to shift in a way that generates conceptual puzzlement. One

tends to undergo a series of cognitive steps something like the following. (Note that some of these are cognitive "acts of omission.")

1. Focusing on the causal explanation of behavior at some theoretical level more fundamental than the psychological level--e.g., neurochemical explanation of specific muscle movements.
2. Accommodating, automatically and subliminally, to the level-parameter appropriate to this kind of causal explanation.
3. Failing to notice that such accommodation has occurred, or that context-sensitive level-parameters are operative for the notions of cause and causal explanation.
4. Noticing that for the kind of natural-science causal explanation under consideration, intentional properties (if any) of the causally operative states and structures are quite irrelevant.
5. Shifting focus to the role of mental properties in the causal explanation of action.
6. Failing to accommodate to the level-parameter appropriate for mentalistic causal explanation.
7. Finding it intuitively plausible that if mental properties are not just identical to neurochemical properties, then they never have genuine causal/explanatory relevance at all.

The crucial component in such a process is step 6, which paves the way for the state of philosophical puzzlement that arises at step 7. One key factor contributing to step 6 is the overarching failure to notice, at the level of reflective consciousness, that a context-sensitive level-parameter is operative at all for the notions of causation and causal explanation, or that accommodation is going on. Another is a subliminal cognitive resistance to the kind of accommodation that actually would be required at step 5. To accommodate properly would be to acquiesce in standards of causal-explanatory relevance that are lower, on a scale we might call "comparative degree of causal-explanatory fundamentality," than the standards

already operative after step 2; and accommodation involving lowering of standards often does not go smoothly (as Lewis points out).

In short: Notions like cause and causal explanation are normally governed by an implicit, contextually variable, level-parameter that determines which level of description, and which kind of counterfactual dependency among properties, are appropriate to focus on in giving a contextually appropriate causal explanation of a given phenomenon. Competent language/concept-users normally keep track of such implicit contextual parameters automatically and subliminally, typically without even noticing them. But in philosophical contexts where one is asking simultaneously about the causal efficacy of mental properties and of physical properties, these implicit parameters have no fixed, stable settings—i.e., there is no determinate score in the language game. In such contexts, one's subliminal cognitive mechanisms for handling implicit discourse-parameters tend to gravitate toward a setting of the level-parameter that is appropriate for talking about physical causation qua physical. This produces the conscious intuition that mental properties "do no real causal work" and hence are epiphenomenal.

This intuition is mistaken, given the causal-compatibilist approach to causation and causal explanation. Nevertheless, the psychological sources of the mistake are very subtle indeed. The psychologically normal operation of subliminal language-processing and concept-wielding cognitive mechanisms interacts with a somewhat abnormal, level-spanning, philosophical mode of discourse to produce a non-veridical intuition. Moreover, because the intuition is caused in this way, it will tend to persist even in those who come to believe that it is mistaken. The intuitive plausibility of causal-exclusionary reasoning is thus a kind of "cognitive illusion," analogous to perceptual illusions like the Muller-Lyer illusion:

The cognitive mechanisms that make the top horizontal line look longer than the bottom one tend to persist in generating this appearance, even in those who become convinced that the two lines are really

the same length. Likewise, the cognitive mechanisms that make the idea of causal exclusion seem so plausible will tend to sustain this mistaken intuition even in those who come to believe that it is mistaken and to understand why it arises.

"Yet more sophistry!" one might be inclined to think, insofar as one continues to give credence to the causal-exclusion intuition. But note well the dialectical state of play: what is on offer is a purported explanation of why that intuition is so strong and persistent while also being mistaken. So it would be too hasty to reject the explanation, and the underlying scorekeeping treatment of causation, simply because of uncritical reliance on the intuition itself.

4. Causal Compatibilism Defended.

The problem of causal exclusion, like many philosophical problems, is simultaneously about matters metaphysical and about the conceptual/semantic workings of certain philosophically important concepts and terms. Metaphysically, it concerns the nature of causation, and the criteria for legitimate causal explanation. But it also concerns the concepts of causation and causal explanation. One investigates the workings of the concepts, in order to come to grips with the metaphysical issue.

I use the term 'ideology' for the investigation into the workings of philosophically important concepts (and the terms that express them). In my view, ideology is a broadly empirical intellectual enterprise, even though (1) often it can be effectively pursued from the comfort of one's armchair, and (2) often the ideological hypotheses advanced in philosophy are, if true at all, conceptually grounded necessary truths. Philosophers propound broadly ideological hypotheses, and they argue for them on the basis of broadly empirical considerations not unlike those that operate in empirical science—even though the data cited in support of these hypotheses are largely available from the armchair. The types of data that can figure in philosophical ideological reflection include the following:

1. Intuitive judgments about what is correct to say concerning various concrete scenarios, actual or hypothetical.
2. Facts about standardly employed warrant-criteria for the use of various concepts.

3. General background knowledge, including untendentious scientific knowledge.
4. Facts about the key sociolinguistic purposes served by various concepts.
5. Facts about conflicting judgments or judgment-tendencies concerning the correct use of certain concepts vis-à-vis various actual or hypothetical scenarios.

Facts of all these kinds can go into the hopper of wide reflective equilibrium whereby ideological claims are defended in philosophy. One makes a case for a certain ideological hypothesis—for instance, the contention that the meaning of natural-kind terms depends on the language users' environment—by arguing that it does a good job, all things considered, of accommodating and explaining the relevant data—for instance, our intuitive judgments about how to describe "Twin Earth" scenarios—in a plausible and theoretically unified way. Needless to say, the epistemic strength of such argumentation depends in part upon the overall theoretical benefits and costs of the given ideological hypotheses, as compared with the overall benefits and costs of competing hypotheses. Typically the argument is to the effect that the ideological hypothesis in question explains and accommodates the salient, largely armchair-accessible, data better than the available alternatives, and that it fares better than the alternatives in terms of overall simplicity, coherence with other well-warranted beliefs both theoretical and common-sensical, conformity with ordinary epistemic standards governing the relevant concepts, and so forth.

In the case of the causal exclusion problem, as with many problems in philosophy, data of type 5 plays a prominent role. We are up against a fairly deep internal tension among our beliefs, belief-tendencies, and intuitions. Each of statements 1-5 in section 1 is initially plausible and is prima facie well warranted, and yet something has to give (because they are jointly inconsistent). So no philosophical treatment of the problem can simply accommodate all five statements as correct. What is needed, rather, is an approach that does the following. First, it should accommodate <u>most</u> of them—preferably, four of the five. Second, it should provide a theoretically <u>respectful</u> treatment of whatever statement it rejects as false—a treatment that construes the reasoning and intuitions behind the rejected statement as the product of some fairly subtle error, rather than as any kind of egregious blunder or gross logical/conceptual

howler. For, as Aristotle said, "whenever a reasonable explanation is given of why a false view appears true, this makes us more confident of the true view" (Nichomachean Ethics VII.14, 1154a24-25).

In section 3 I set forth a number of ideological hypotheses about the notions of causation and causal explanation—hypotheses that together constitute a version of causal compatibilism vis-à-vis the causal exclusion problem. Let me now briefly defend this causal compatibilist position, by appeal to how it fares with respect to data of types 1-5.

First, type 1. One has strong intuitive judgments, with respect to specific scenarios of various kinds, about matters of mental causation and mentalistic causal explainability. For many such scenarios, one intuitively judges that the agent's behavior is the causal result of the agent's mental processes qua mental. A plausible empirical assumption about these judgments is that normally they are the straightforward product of one's own conceptual/semantic competence with the notions of causation and causal explanation, and with common-sense mentalistic notions—and hence are normally true. Compatibilism accommodates such judgments straightforwardly, by allowing them to be true under paradigmatic attribution-conditions. Incompatibilist positions of various kinds, however—causal emergentism, eliminativism, epiphenomenalism, identity materialism—all hold mentalistic causal and causal-explanatory claims hostage to a very demanding additional condition, over and above the conditions that are clearly satisfied under paradigmatic attribution-conditions—viz., the requirement that mental properties are fundamental force-generating properties. Ceteris paribus, one ideological hypothesis is better than another if the former accommodates the attributional practices of competent users of the relevant concept(s) better than the other. So in this respect, compatibilism does better than incompatibilism.

Data of type 2 reinforces these considerations. The evidential standards that are normally employed, in one's practice of attributing mental causes and proferring mentalistic causal explanations (and also in the corresponding default assumptions one routinely and implicitly makes about mental causation as operative in oneself and in others) are standards under which such attributions (and such default assumptions) are regarded as epistemically warranted independently of the availability or lack of

availability of any strong evidence for or against the contention that mental properties are fundamental force-generating properties. A plausible empirical assumption is that the epistemic standards one employs, when one makes confident intuitive judgments that various behaviors and mental events are caused by prior mental events qua mental—and when one confidently adopts and maintains the default presumption that most ordinary human behavior, and many mental events, are caused by prior mental events qua mental—are appropriate epistemic standards, given the ideological workings of common-sense mentalistic concepts and the concepts of mental causation and mentalistic causal explanation. For, the use of grossly inappropriate epistemic standards, in the confident intuitive deployment of a concept, typically reflects a definciency in one's conceptual/semantic competence with concept. Compatibilism accommodates the epistemic standards governing normal judgments and default assumptions about mental causation, since it treats these standards as appropriate. Incompatibilism, on the other hand, entails that such standards are much too lax; under suitable epistemic standards, mental-cause attributions and mentalistic causal explanations are only warranted when one has good evidence that humans instantiate mental properties that are fundamental force-generating properties. Ceteris paribus, an ideological hypothesis that accommodates the epistemic standards normally accompanying a concept's deployment is better than ideological hypothesis entailing that those standards are seriously deficient. So in this respect too, compatibilism does better than incompatibilism.

Data of type 3 favors compatibilism. Attributions and default assumptions about mental causation are absolutely fundamental to our conception of ourselves as agents—creatures whose behaviors constitute not merely bodily motions, but actions performed for reasons. In order to count as an action, after all, an item of behaviour must be caused by mental states qua mental—for instance, by belief/desire combinations that constitute reasons. Human concepts emerge pragmatically, in ways that serve the purposes for which those concepts are employed. In general, therefore, concepts do not have satisfaction conditions built into them that are so demanding that they thwart the very purposes the concepts serve. Attributions and presuppositions of human agency play a ubiquitous and fundamental role in virtually every aspect of human life; this is so whether or not mental properties are fundamental force-generating

properties. So it would be purpose-thwarting for the concept of a causal property to require all causal properties to be fundamental force-generating properties (since this requirement could very well fail to be satisfied). Compatibilism, therefore, accords better than incompatibilism with certain central purposes that the concept of causation serves.

Data of type 4 reinforces these latest considerations. From the epistemic vantage point of the educated layperson concerning large-scale developments in science, and specifically concerning the neurophysical etiology of human behavior and the neurophysical realization of mentality, the hypothesis that physics is causally closed is extremely well warranted. Moreover, the psychophysical type-type identity theory is very likely false, because it appears both conceptually and nomically possible (given current scientific knowledge) that mental properties are multiply realizable in different actual or possible species of creatures with mentality—or in a single species of creatures, or even in a single creature at a specific moment in its life. (Indeed, it is a very live epistemic possibility that mental properties are multiply realizable in humans, even individual humans at a specific moment.) Thus, there are strong grounds to believe that mental properties simply are not fundamental force-generating properties; they do not meet the stringent requirements that incompatibilist views impose upon them. All the more reason, then, to think that it would be purpose-thwarting for the concept of a causal property to operate in such a way that only fundamental force-generating properties count as genuine causal properties.

When these considerations are fed together into the hopper of wide reflective equilibrium, they reinforce one another epistemically in such a way that their combined epistemic weight is very powerful. The compatibilist hypothesis simply accords better with the relevant data of types 1-4 than does the incompatibilist hypothesis. In particular, compatibilism explains why our ordinary intuitive attributions and presumptions of agency and mental causation, and the accompanying justification standards we rely upon in making such attributions and presumptions, operate in a way that is essentially orthogonal to the question of whether mental properties are fundamental force-generating properties. They do so because their truth simply does not require mental properties to have that feature.

There remains, of course, the task of dealing with considerations of type 5. An adequate ideological treatment of the concepts of causation and causal explanation, vis-à-vis the problem of causal exclusion, owes a credible account of why people's intuitions can be easily and naturally pulled in the direction of thinking that if physics is causally closed and mental properties are not identical to physical causal properties, then mental properties are excluded from having any genuine causal role. Incompatibilist positions can explain this tendency fairly straightforwardly, as the putative product of our conceptual/semantic competence with the concepts of causation and causal explanation. Compatibilists, however, have the burden of explaining credibly why the tendency arises so strongly and naturally even though is allegedly mistaken—and the closely related burden of explaining credibly what is wrong with exclusionary arguments of the kind set forth in section 1.

My own proposed explanation, drawing upon the ideological hypothesis that the concepts of causation and causal explanation are governed by an implicit, context-sensitive, level-parameter, was set forth in sections 3.2 and 3.3 above. The explanation is respectful: it treats exclusionary intuitions, and the kinds of reasoning associated with them, as a very natural byproduct of the workings the cognitive mechanisms whereby people normally take implicit countextual parameters in stride often without even noticing them. So the burden of explaining data of type 5 can be satisfactorily discharged, given the version of compatibilism that includes the ideological hypothesis that causation and causal explanation are contextually parameterized notions governed by an implicit level-parameter.

There are theoretical costs associated with my position, admittedly. Ceteris paribus, a theory incorporating the ideological hypothesis just mentioned is more complex than a theory eschewing it. But when this hypothesis is considered in a wider dialectical context—with an eye on the strong evidence for compatibilism provided by data of types 1-4, I submit that the theoretical advantages of the position outweigh this cost. On balance, the overall weight of the evidence points toward compatibilism, and specifically toward the version of compatibilism I have described here.

Let me make three final points. First, the form of compatibilism I propose, with its emphasis on an implicit level-parameter, is fairly generic. There are various potential ways of specifying in more detail

the kinds of relations within families of properties in virtue of which the members of a given family count as causal properties (under a suitable setting of the level-parameter). Various more specific treatments of causation and causal explanation—in philosophy of mind, in metaphysics, and in philosophy of science—could potentially be harnessed to my generic position, by taking on board the idea of an implicit level-parameter.

Second, although the challenge posed by the problem of causal exclusion demands a philosophical response, many treatments of causation in the philosophical literature have failed to address it clearly and explicitly. When one faces this challenge squarely, I think one should come to appreciate the attractions of grafting the the level-parameter hypothesis onto whatever general philosophical treatment of causation one might be inclined to embrace.

Third, data of types 1-4 strongly supports compatibilism over against incompatibilism, whether or not the best version of compatibilism ultimately turns out to be one that incorporates the hypothesis of an implicit level-parameter. In principle, there could turn out to a version of compatibilism with a different and better way of handling data of type 5—a different and better way of explaining why causal-exclusion intuitions are mistaken and how causal-exlusionary reasoning goes wrong. But if there is such an alternative kind of compatibilism, then I would certainly like to know what it is.

REFERENCES

Graham, G. and Horgan, T. (1991). In Defense of Southern Fundamentalism, Philosophical Studies 62, 107-34.

Graham, G. and Horgan, T. (1994). Southern Fundamentalism and the End of Philosophy, Philosophical Issues 5, 219-47.

Henderson, D. and Horgan, T. (2000). What Is A Priori, and What Is It Good For?, Southern Journal of Philosophy 39, Spindel Conference Supplement, The Role of the Empirical and the A Priori in Philosophy.

Henderson, D. and Horgan, T. (in press). What Does It Take to Be a True Believer? Against the Opulent Ideology of Eliminative Materialism. In C. Erneling and D. Johnson (eds.), Mind as a Scientific Object: Between Brain and Culture. Oxford.

Horgan, T. (1989). Mental Quausation, Philosophical Perspectives 3, 47-76.

Horgan, T. (1993a). The Austere Ideology of Folk Psychology. Mind and Language 8, 282-97.

Horgan, T. (1993b). Nonreductive Materialism and the Explanatory Autonomy of Psychology. In Wagner & Warner, eds., Naturalism: A Critical Appraisal. Notre Dame, 295-320.

Horgan, T. (1994). Nonreductive Materialism. In R. Warner and T. Szubka, eds., The Mind-Body Problem. Blackwell, 236-41.

Horgan, T. (1996). Kim on the Mind-Body Problem, British Journal for the Philosophy of Science 47, 579-607.

Horgan, T. (1998). Kim on Mental Causation and Causal Exclusion, Philosophical Perspectives 11, 165-84.

Horgan, T. (in press). Multiple Reference, Multiple Realization, and the Reduction of Mind. In F. Siebelt and B. Preyer, eds., Reality and Humean Supervenience: Essays on the Philosophy of David Lewis. Rowman & Littlefield.

Horgan, T. and Tienson, J. (1990). Soft Laws, Midwest Studies in Philosophy 15 (1990), 256-79.

Kim, J. (1993). Supervenience and Mind: Selected Philosophical Essays, Cambridge.

Kim, J. (1997). The Mind-Body Problem: Taking Stock after Forty Years, Philosophical Perspectives 11, 185-207.

Kim, J. (1998). Mind in a Physical World. MIT.

Lewis, D. (1966). An Argument for the Identity Theory, Journal of Philosophy 63, 17-25.

Lewis, D. (1973). Scorekeeping in a Language Game, Journal of Philosophical Logic 8, 339-59. Reprinted in Lewis (1983).

Lewis, D. (1983). Philosophical Papers, Volume 1. Oxford.

[1] See Graham and Horgan (1991, 1994), Henderson and Horgan (2000, in press), Horgan (1989, 1993a, 1993b, 1994, 1996, 1998, in press), Horgan and Tienson (1990).

[2] Reasoning of the kind set forth in this paragraph and the preceding one figures prominently in the writings of Jaegwon Kim, who has been much exercised by the problem of causal exclusion. See, for instance, Kim (1993, 1997, 1998). I discuss Kim's treatment of causal exclusion, and other aspects of his work in metaphysics and philosophy of mind, in Horgan (1996, 1998).

[3] Someone could espouse the compatibilist thesis and still affirm the type-identity theory on other grounds. But here we are considering positions that deny a single thesis and retain the others. (Fans of the identity theory do tend to invoke causal-exclusionary reasoning in favor it; see, for instance, the closing paragraphs of Lewis (1966).)

[4] How should Kim be situated, relative to this taxonomy of positions? In my view, his thinking has persistently manifested deep internal tensions, because of the extent to which he has remained in the grip of all of claims 1-5, and has tried to find a philosophical position that somehow accommodates the spririt of each of them. Many of his writings seem to me to exhibit what I call "angst-ridden causal compatibilism": he tries to give an account of how mental properties can be causal properties, given claims 1-4, an account which somehow accommodates the spirit of causal-exclusion reasoning like that I described above. In my view, this is an inherently unstable position. In more recent writings, he has moved toward a version of eliminativism about mental properties, while still wanting to be a "causal realist" and an "explanatory realist" about the mental; this too looks inherently unstable to me. I myself think his intellectual trajectory is pointing him toward a version of the psychophysical identity theory of the Armstrong-Lewis type, one which treats mental-property names as population-relative nonrigid designators of physical causal properties. Cf. Horgan (1998).

[5] Typically, I believe, such generalizations will be what are called "soft laws" in Horgan and Tienson (1990). In the case of psychology, for instance, this means that the generalizations will have ineliminable ceteris paribus clauses adverting not merely to potential lower-level exceptions resulting from physical

breakdown (e.g., having a stroke) or from external physical interference (e.g., being hit by a bus), but adverting to potential psychology-level exceptions as well.

[6] I think that other implicit parameters operate too, typically in an intra-level way. These are related to the contextually appropriate way of distinguishing between what counts as cause, and what counts instead as "background conditions."

[7] On some metaphysical schemes, an action is identical to a bodily motion (i.e., a combination of muscular contractions and relaxations); on other schemes, an action is constituted by, but is not identical to, a bodily motion. But the point about multiple causal explanations holds either way. Even if the action is not identical to the motion, one way to causally explain the action is to causally explain the motion that constitutes it.

[8] I call this an "approximate paraphrase" because I do not want to commit myself to saying that that this kind of quantification over parameter settings is part of the actual content of a claim like "M and P are noncompeting causal properties." I am inclined to think that normally, particular parameter-settings are not part of a statement's content, but rather are partial determinants of its specific content. If so, then it is a complex matter how best to describe the workings of implicit level-parameters in level-spanning discourse. Although the approximate paraphrase certainly illuminates what is going with the level-spanning claim that M and P are both causal properties, the paraphrase may not have quite the same content as the claim it illuminates. (A more adequate semantical treatment of such level-spanning claims, I suspect, would construe them as involving dynamic variation in the operative setting of the implicit level-parameter: the setting "jumps levels" as one moves, in the course of a single assertion, from saying that M is a causal property to saying that P is a causal property.)