MIND-BODY DUALISM AND MENTAL CAUSATION

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Benjamin G. White
May 2016

Examining Committee Members:

Gerald Vision, Advisory Chair, Philosophy Department
Miriam Solomon, Philosophy Department
David Wolfsdorf, Philosophy Department
Kenneth Aizawa, External Member, Rutgers Newark

ABSTRACT

The Exclusion Argument for physicalism maintains that since every physical effect has a sufficient physical cause, and cases of causal overdetermination (wherein a single effect has more than one sufficient cause) are rare, it follows that if minds cause physical effects as frequently as they seem to, then minds must themselves be physical in nature. I contend that the Exclusion Argument fails to justify the rejection of interactionist dualism (the view that the mind is non-physical but causes physical effects). In support of this contention, I argue that the multiple realizability of mental properties and the phenomenal and intentional features of mental events give us reason to believe that mental properties and their instances are non-physical. I also maintain (a) that depending on how overdetermination is defined, the thesis that causal overdetermination is rare is either dubious or else consistent with interactionist dualism and the claim that every physical effect has a sufficient physical cause, and (b) that the claim that every physical effect has a sufficient physical cause is not clearly supported by current science. The premises of the Exclusion Argument are therefore too weak to justify the view that minds must be physical in order to cause physical effects as frequently as they seem to.

TABLE OF CONTENTS

Page

ABSTRACT.....ii

LIST OF TABLES.....iv

LIST OF FIGURES.....v

CHAPTER

1. THE EXCLUSION PROBLEM.....1

2. HOW NOT TO SOLVE THE EXCLUSION PROBLEM.....25

3. THE CASE FOR DUALISM AND THE NATURE OF REALIZATION.....57

4. THE MULTIPLE REALIZABILITY OF MENTAL PROPERTIES….....100

5. THE NOMOLOGICAL IRREDUCIBILITY OF MENTAL PROPERTIES.....149

6. THE IMMATERIALITY OF MENTAL EVENTS.....198

7. MENTAL CAUSATION WITHOUT OVERDETERMINATION.....261

8. THE CAUSAL SELF-SUFFICIENCY OF THE PHYSICAL.....292

9. THE CAUSAL EFFICACY OF THE MIND.....344

BIBLIOGRAPHY.....391

LIST OF TABLES

Table Page

1. Types of realization.....95

LIST OF FIGURES

Figure Page

1. Spectral sensitivities of the three types of cone receptors in the human eye.....106

2. Wavelength distributions of two metameric lights.....107

3. Example of potential neural circuitry for opponent processing.....112

4. Chromatic response curves for standard observer.....113

5. Example of simultaneous chromatic contrast.....115

6. Experimental set-up for the double slit experiment.....329

7. Gradual accumulation of particles in the double-slit experiment.....330

8. Diagrammatic representation of the causation of lung cancer.....367

9. Diagrammatic representation of the causation of intentional action.....372

CHAPTER 1

THE EXCLUSION PROBLEM

It is a belief widely and firmly held that our minds have a causal impact on the world by affecting the motion of our bodies and thence the states of other physical objects in our surrounding environment. When a person gets up from the couch to rummage about in the fridge, it is, we say, their *desire* for food and *belief* that there is desirable food in the fridge (which are states of their mind) that cause them to stand up, walk to the kitchen, open the fridge, and have look around. Similarly, when one cuts one's finger, it is, we say, the resulting *sensation* of pain (another mental state) that causes one to wince and say "Ow!" One of the most long-standing objections to mind-body dualism is that the assumption that the mind is non-physical makes it very difficult, if not impossible to see how the mind could play any role in the causation of such effects. In contemporary philosophy of mind, this objection has taken on the form of what is now arguably the main argument against any non-epiphenomenalist variety of dualism: the Exclusion Argument for physicalism. Roughly stated, the Exclusion Argument holds that since every physical effect has a sufficient physical cause, and the physical effects of mental events are not typically overdetermined (meaning that they do not normally have more than one sufficient cause), the fact that mental events cause physical effects entails that such events must themselves be physical in nature. Though often treated as a knock-down refutation of all forms of dualism that involve a commitment to the causal efficacy of mental events, I think that the Exclusion Argument fails to provide sufficient evidence

to justify belief in the falsity of such a position. The aim of this dissertation will be to substantiate this claim.

The present chapter clears the way for the execution of this task by clarifying the central positions to be discussed in what follows and providing a more precise formulation of the Exclusion Argument along with a brief account of some of the historical trends that led to its emergence as one of the primary contemporary arguments against mind-body dualism. The chapter then closes with a short summary of the remaining chapters.

1. A taxonomy of dualisms and definition of physicalism

While mind-body dualism comes in many different forms, the Exclusion Argument really only applies to those varieties of dualism that reject epiphenomenalism. This, however, leaves a number of different views open to the Argument that seem to have little in common other than their shared commitment to the idea that the mind is non-physical, the most prominent among these being: Substance Dualism, Neutral Monism, Dual-Aspect Theory, Materialistic Panpsychism, and Emergentism. All five of these positions hold (or are at least consistent with the view) that the mind is causally efficacious. Of these five, only the first (Substance Dualism) holds that mental and physical properties are properties of distinct kinds of substances. The remaining four positions are all substance monistic forms of property dualism. The first two (Neutral Monism and Dual-Aspect Theory) hold that the only substances are neither physical nor mental, while the last two (Materialistic Panpsychism and Emergentism) hold that all

substances are physical.[1] Partly to simplify matters, and partly because I find it to be the most plausible of the various forms of dualism threatened by the Exclusion Argument, I'll largely limit my attention in what follows to Emergentism, understood as the conjunction of the following five theses:

> (a) Mental properties (i.e., properties that things exemplify insofar as they are endowed with intentionality and/or consciousness) and their instances are entirely distinct from, irreducible to, and incapable of being fully explained in terms of physical properties[2] and their instances.
>
> (b) Instances of mental properties have a causal impact on events in the physical world.
>
> (c) There is only one kind of substance in which both mental and physical of properties inhere.
>
> (d) All substances are physical substances.
>
> (e) Mental properties are exemplified only by aggregates of physical substances that satisfy certain conditions, such as exhibiting a certain complexity of structural and/or functional organization, or standing in certain causal relations to the surrounding environment.

---

[1] In its classical, Spinozistic form, Dual-Aspect Theory is the view that the mental and the physical are two distinct attributes or aspects of a single substance that is intrinsically neither physical nor mental. It is distinguished from Neutral Monism by the fact that the latter allows that a substance may exemplify physical properties without exhibiting any mental properties (or vice versa), whereas according to Dual-Aspect Theory, all finite modes (or substances, if one parts with Spinoza's belief that substance is one in both kind and number) have both a mental and physical aspect. Dual-Aspect Theory is hence a form of Panpsychism. It is distinguished from Materialistic Panpsychism by the fact that the latter holds that all substances are physical.

[2] I.e., those properties that, in Jackson's (1998, p.8) words, "are those that [(barring panpsychism)] we need to handle the non-sentient, they are broadly akin to those that appear in current physical science, [and] they are those we need to handle the relatively small."

(a)-(e) exemplify a progressive winnowing of the various strains of dualism mentioned above. (a) is a statement of mind-body dualism; as such it is an essential commitment of any dualist view, including epiphenomenalism. The addition of (b) rules out epiphenomenalism. Any position committed to (a) and (b) is hence vulnerable to the Exclusion Argument. (c) excludes Substance Dualism, leaving only substance monistic varieties of property dualism. (d) limits the field to Materialistic Panpsychism and Emergentism, the former of which is eliminated by (e). Thus, schematically, with the initials of each position listed next to the theses to which it is committed:

(a) E., M.P., D.A.T., N.M., S.D., Epi.

(b) E., M.P., D.A.T., N.M., S.D.

(c) E., M.P., D.A.T., N.M.

(d) E., M.P.

(e) E.

Although Epiphenomenalism is indeed a kind of dualism in the broad sense defined by (a), given that the Exclusion Argument threatens only those forms of dualism committed to both (a) and (b), it will at times be important that Epiphenomenalism be kept separate from those forms of dualism that view the mind as causally efficacious. For this reason, "dualism" and its grammatical variations should be read as excluding reference to Epiphenomenalism when the context makes it clear that it is only non-epiphenomenalist forms of dualism that are at issue. The phrase "interactionist dualism" will often be used to indicate that it is only forms of dualism that are committed to both (a) and (b) that are being referred to.

While dualism is certainly not extinct, its adherents are today something of an endangered species, due largely to the rise of physicalism as the predominant metaphysical outlook in Anglo-American philosophy. Given the variety of different forms that physicalism may take (e.g., type-physicalism, realizer physicalism, token- or "non-reductive" physicalism, supervenience physicalism, superdupervenience physicalism, *et al.*), the central commitments of the position have been notoriously difficult to formulate. That said, I'll be adopting D. Gene Witmer's (2001, p.69) definition of physicalism, as the doctrine that "Every law of nature and every particular fact is either physical or to be explained by the physical in such a way as to imply that the nonphysical facts are nothing over and above the physical facts, where the physical facts include the actual distribution of physical properties and the laws of physics," this being the most general and unobjectionable formulation of the position that I am aware of.

The ascendance of physicalism has brought with it many criticisms of mind-body dualism, of which the Exclusion Argument is arguably the most significant. My purpose in what follows is strictly to show that this Argument does not provide sufficient reason for rejecting interactionist dualism. In saying this, I wish to be clear that though at times my manner of expression may suggest otherwise, my aim is *not* to demonstrate that interactionist dualism is true, but rather to show that one of the putative refutations of the position is inconclusive, and that hence, for all the Exclusion Argument shows, interactionist dualism remains a viable theory of mind. My motive likewise stems not from any unshakable conviction that such a position is the only reasonable view to hold, but rather from a desire to check the over hasty dismissal of a theory that could very well be true. Of course, such a project would be pointless if there were no good reason to think

that some form of interactionist dualism *might* be correct. After all, why bother criticizing an argument against a view that is incoherent, unmotivated, or refuted on other, obvious grounds? Chapters 3-6 will hence be dedicated to making the case that interactionist dualism is at least independently plausible. It should, however, be borne in mind that the success of my efforts ought to be judged not on the basis of whether I have shown that such a position is the only, or even the *most* plausible theory of mind, but instead on the basis of whether I have shown that anyone who rejects such a position solely on the grounds provided by the Exclusion Argument is making a leap of faith.

Having defined the class of positions that the Exclusion Argument purports to refute (viz. those forms of dualism committed to both (a) and (b)), and the position that it purports to establish (viz. physicalism), we can now turn our attention to the Argument itself. Before examining the structure of the Argument, though, it may be helpful to have some understanding of its history.

2. A brief history of the Exclusion Argument[3]

While Descartes was certainly not the first philosopher to reflect upon the nature of the mind, it is nevertheless arguably with him that the problem of making one's thoughts about what the mind *is* consistent with what one believes the mind can *do* finds its clearest expression. This is largely because the perceived failure of Descartes' theory of mind is almost entirely attributable to his seeming inability to reconcile these two

[3] Portions of the following historical sketch are drawn from LePore and Loewer (1989) and Kim (2007).

aspects of his position, expressed in his two theses (a) that the mind is an unextended, immaterial substance, and (b), that the mind causally interacts with the extended, physical body to which it is joined, and thence with the rest of the material world. The traditional objection to the conjunction of these theses is that it is incomprehensible how an unextended, immaterial substance could impart motion to a material body that exists in space. The resulting tension between the two basic tenets of Descartes' theory of mind was famously pointed out by Princess Elisabeth of Bohemia, who wrote in a letter to Descartes (dated June 10, 1643) "that it would be easier for me to concede matter and extension to the mind than it would be for me to concede the capacity to move a body and be moved by one to an immaterial thing." Rather than identifying mind with matter, as Princess Elisabeth suggested, other philosophers of Descartes' time sought to remedy the tension she had noted by explaining away the appearance of causal interaction between mind and body as an illusion that arises from close, non-causal correlations between mental and physical events. The Occasionalism of Malebranche, Leibniz's doctrine of Pre-established Harmony, and Spinoza's Parallelism can all be seen as attempts to account for the apparent causal interaction between mind and body in this way.[4] Still others (e.g. Berkeley) took the view that the difficulties surrounding mind-body interaction were best resolved not by reducing mind to matter, as Princess Elisabeth

[4] Things are actually not quite so simple with respect to Leibniz and Spinoza, for while in the 1686/1989 *Discourse on Metaphysics* (sect.33), Leibniz does tout his doctrine of pre-established harmony as providing "unexpected illumination of…how it happens that the passions and actions of the [soul] are accompanied by the actions and passions…of the [body]," in his later *Primary Truths* (1689/1989), he also asserts that material bodies exist only as "true phenomena" and in the scholium to *Ethics* IIp7, Spinoza claims that "a mode of Extension and the idea of that mode are one and the same thing." These statements make it a bit more difficult to read Leibniz and Spinoza as attributing the appearance of mind-body interaction to the existence of correlations between the states of *distinct* mental and physical things.

proposed, but by instead reducing material objects to collections of ideas in the mind.[5] While none of these attempts to resolve the tension between the apparent immateriality and causal efficacy of the mind succeeded in winning the approval of the general philosophical community in the 250 years or so following Descartes' death, the efforts spent by some of the period's most notable figures in grappling with this problem, and the number, variety, and complexity of the solutions they developed in their attempts to solve it attest to the importance that the problem was then viewed as having.

The increased skepticism towards metaphysical speculation occasioned by the rise of Logical Positivism in the early 20th century led many to question whether the alleged difficulty of accounting for the apparent causal interaction between mind and body was indeed as significant as it had previously been made out to be. Along with various other metaphysical notions, the concept of causation (insofar as it was taken to involve more than the mere "constant conjunction" of one type of event with another) came to be viewed as suspect due to its lack of clear observational criteria of application, and longstanding metaphysical problems resting on questions or claims about causation (e.g. how the mind causes bodily motion) were consequently branded as mere pseudo-problems, which were irresolvable only because the sentences in which they were posed had no cognitive meaning. The resulting dismissal of the problem of mental causation was further encouraged by the ascendance of behaviorist psychology, which eschewed the traditional conception of mental states as internal, introspectible events distinct from their outward, behavioral effects, and instead favored an analysis of such states as mere

[5] See *PHK* I.19, where Berkeley cites the unintelligibility of mind-body interaction as grounds for denying the existence of mind-independent material substances.

tendencies or dispositions to respond in certain ways to certain types of stimuli. On such an analysis, the question of how the mind causes behavioral effects rests on a category mistake, for according to the behaviorist, the mind does not *cause* behavior; it just *is* a collection of various patterns of behavior and dispositions to behave in certain ways. Another contributing factor to the general disregard during this time period for traditional problems concerning mind-body interaction was the influential (and behaviorist-leaning) work of the later Wittgenstein and his disciples, who argued that a person's reasons for acting must be distinguished from their actions' causes, and that mental events do not cause, but rather rationalize the behavior they explain. Since, on this view, no causal relations between mental and physical events are needed to justify our practice of explaining peoples' actions in terms of their mental states, those who adopted it came to see worries about how seemingly non-physical mental events could cause bodily motions as pointless at best.

A combination of at least three factors helped set the stage for a revival of interest in the topic of mental causation in the 1960's and 70's. First, the decline and fall of Logical Positivism and its Verificationist constraints on theorizing allowed for the reintroduction of "metaphysical" notions (e.g. causation, viewed as more than mere "constant conjunction") into respectable theoretical discourse. Second, the overthrow of behaviorism and ascendance of functionalism as the new orthodoxy in psychology re-legitimized the conception of mental states as internal states that mediate between, but are nonetheless distinct from, their sensory causes and behavioral effects. And third, Donald Davidson's (1963) paper "Actions, Reasons, and Causes" advanced a compelling rebuttal to the Wittgensteinian's non-causal interpretation of psychological explanations of

behavior that was widely taken to show that mental events can, and indeed *must* cause the behavior they explain, as there would otherwise be no way for us to identify which of an agent's reasons actually accounts for his/her actions in cases where s/he has multiple good reasons for the action s/he performs.[6] Under the influence of these three factors, the intellectual climate was restored to a state wherein questions concerning the mind's ability to causally impact bodily motion could again be viewed as not only intelligible, but worthy of serious philosophical consideration.

Conditions being thus ripe for a renewed interest in mental causation, Davidson's (1970) paper "Mental Events" reintroduced the subject as a central topic of debate in contemporary philosophy of mind. Detailed discussion of Davidson's paper will be reserved for Chapter 6, but the following sketch should suffice for the purposes of the present narrative. In the paper, Davidson argues that since causation requires that causes and effects be connected by "strict" laws, and events are subsumable under such laws only insofar as they satisfy physical descriptions, the fact that mental events cause physical effects entails that the former must themselves be physical events. However, since there are (Davidson claimed) no strict psychophysical laws relating mental properties to physical properties, the mental descriptions that mental events satisfy (by virtue of which they qualify as mental events) cannot be reduced to physical descriptions. This led Davidson to the view that while all events are physical, some events are both physical *and* mental, inasmuch as they fall under both physical and mental kinds. The

[6] Davidson's (1963) argument is discussed further in the following chapter.

causal efficacy of the mind is then ensured by the fact that mental events are also physical events, and as such can be linked to other physical events by way of strict laws.

The position outlined in Davidson's paper provided the blueprint for the "non-reductive" form of physicalism that became accepted as the new "standard" metaphysics of mind. Non-reductive physicalism differs from other forms of physicalism in holding that while mental events (i.e., instances of mental properties) are token-identical with physical events, mental properties are distinct from and irreducible to physical properties. Most non-reductive physicalists joined Davidson in viewing the alleged physical status of mental events as sufficient to stave off any worries about how the mind causes physical effects. Critics of Davidson and the non-reductive brand of physicalism he had helped to create were, however, skeptical of the ability of such theories to fully justify the claim that the mind indeed plays a role in the causation of behavior. The basic worry voiced by these critics was that if it is only by virtue of their satisfying some physical description that mental events are capable of causing physical effects, then the fact that such events are also *mental* events that satisfy certain *mental* descriptions seems *irrelevant* to their having the causal powers that they do. Put simply, on the picture endorsed by Davidson and other non-reductive physicalists, it seems as though it is the physical properties involved in mental events that are really "doing all the work," while the mental properties involved in such events are mere "freeloaders" that contribute nothing to the actual production of any physical effect. If this is so, then the alleged ability of non-reductive physicalism to vindicate the causal efficacy of the mind without reducing mental properties to physical properties is illusory. Upon closer examination, the position actually amounts to a form of epiphenomenalism.

The emergence of the Exclusion Argument as a subject of interest in the philosophy of mind in the late 1980's and early 1990's is attributable primarily to the work of Jaegwon Kim (1989a; 1989b; 1993c; 1998, p.37; 2011, p.214), who first presented the Argument as an articulation of the objection to non-reductive physicalism just described. In its original formulation, the Exclusion Argument was thus intended to serve as an argument for *reductive* (or "type") physicalism, rather than for physicalism in general. Although its success as a refutation of non-reductive physicalism, which grants that mental events are at least *token*-identical with physical events, has been widely contested, the Exclusion Argument is generally viewed as posing an unequivocal threat to all forms of interactionist dualism. The relation between the Exclusion Argument and non-reductive physicalism will be explored further in Chapter 6, but given that our main concern is with the Argument's status as a putative refutation of interactionist dualism, it will better suit our interests to formulate it an argument *against* interactionist dualism and *for* physicalism in general.

3. The Exclusion Problem

Framed in this way, with interactionist dualism as its primary target, the Exclusion Argument can be presented roughly as follows: The problem for dualists is that if (as the explanatory completeness of physics requires) every physical effect has a sufficient physical cause, then any mental causation of physical events seems superfluous, since any physical effect which a mental event might be presumed to cause will already be adequately accounted for in purely physical terms. One might try to

swallow this conclusion by suggesting that perhaps all physical effects with mental causes are causally overdetermined (in that they have distinct, independently sufficient mental and physical causes), but given the fact that mental causes, if existent, are likely abundant, the result would be a kind of systematic causal overdetermination too extravagant to be credible. It appears, then, that if we want to retain a causal role for the mind in the physical world, our only remaining option is to say that mental events *just are* physical events. The Exclusion Argument thus seems to compel us to identify the mind with something purely physical in order to preserve its capacity to impact events in the physical world.

Thus stated, the Argument can be seen to take the following form:

(1) *Causal Efficacy of the Mental:* Mental events (frequently[7]) cause physical effects.

(2) *Causal Self-Sufficiency of the Physical:* Every physical effect has a sufficient physical cause.

(3) *Absence of Systematic Overdetermination:* Causal overdetermination is rare.

(4) *Physicalism:* Therefore, mental events are physical events.

Before looking more closely at the premises and structure of the Argument, two quick remarks regarding its general drift are in order. First, it should be noted that in contrast to the kind of objection raised by Princess Elisabeth, the problem that the Exclusion Argument poses for interactionist dualism is not that the causation of physical effects by

[7] Although this qualification is needed to ensure the validity of the argument, I'll omit it hereafter. Nearly all parties agree that if mental causation of physical events occurs at all, then it happens quite often. It would hence be a rather poor response to the Exclusion Argument to try and avoid the conclusion by maintaining that mental causation of physical events is as rare as or rarer than causal overdetermination.

non-physical, mental events is *incomprehensible* or otherwise intrinsically objectionable, but rather that such causation is simply *redundant*. In short, even assuming that clear sense could be made of how immaterial mental events might do things like alter bodily motions, if the reasoning behind the Exclusion Argument is sound, then the ascription of causal efficacy to such mental events would still be gratuitous, because all the causal work that needs doing can already be done by purely physical events. The contrast between these two different criticisms of dualist accounts of mental causation should be borne in mind, for while a response to the charge of incomprehensibility will be offered in Chapter 9, the bulk of what follows will be devoted to answering the worry raised by the Exclusion Argument, which is that dualism makes mental causes superfluous.

Second, the concern articulated by the Exclusion Argument, that non-physical causes of physical effects are unnecessary or redundant, can also be seen as an expression of a more general view concerning the course of future science that has widespread (though perhaps diminishing) currency in and influence over contemporary thought. This is the view that as science advances, more and more theories in the various "special" sciences will end up being reduced to (or eliminated in favor of) theories in basic physics, until there is ultimately no need to appeal to any "higher level" entities to explain any sort of event that takes place in nature. Motivation for this view is typically drawn from reflection on the history of science, which seems to exhibit a similar trend of reductions and eliminations of higher level laws, forces, and properties as the phenomena they were introduced to explain are found to be better explained in terms of the laws, forces, and properties of lower level disciplines. Awareness of such trends may naturally lead one to expect that the future will be no different, and that as science continues to develop,

physics will gradually usurp the domains of the various other sciences until the "Unified Science" envisioned by the Positivists is finally achieved. Those who are sympathetic to this conception of future science will naturally be more receptive to the idea advanced by the Exclusion Argument that if the mind is non-physical, then it fails to do anything that couldn't be explained just as well without it.

Turning now to the premises and structure of the Exclusion Argument as formulated above, the Argument's strength can be seen to rest primarily on the apparent incontestability of its premises. Few have dared to challenge (2), the Causal Self-Sufficiency of the Physical, as this would be to suggest that if we were to trace the causal history of any physical event, it is possible that we should arrive at a point where no preceding physical state could be cited as a sufficient cause for the physical events that transpired at that moment in time, thereby undermining the ability of physics to provide us with a complete and self-standing account of the physical world. Objections to (3), the Absence of Systematic Overdetermination, have likewise been few and far between, as all parties are generally agreed that true cases of causal overdetermination ought to be rather hard to come by; much harder, at any rate, than they would be if every effect produced by a mental cause was similarly overdetermined. Finally, a commitment to the causal efficacy of the mind is widely considered to be too central to our conception of ourselves as human agents, and to our world-view in general, to be given up, and even assuming that the immediate counter-intuitiveness of rejecting (1), the Causal Efficacy of the Mental, can be legitimately dismissed as a mere psychological fact with no reliable import as to the truth or falsity of (1) itself, the suggestion that the mind is causally inert

still faces serious theoretical difficulties in explaining how we then come to know about our own mental states, and how or why such states evolved.

One person's *modus ponens* is, however, another's *modus tollens*, and as Scott Sturgeon (1998) has noted, one of the more interesting features of the Exclusion Argument is that if we negate its conclusion so as to obtain the following four propositions:

(1) *Causal Efficacy of the Mental:* Mental events cause physical effects.

(2) *Causal Self-Sufficiency of the Physical:* Every physical effect has a sufficient physical cause.

(3) *Absence of Systematic Overdetermination:* Causal overdetermination is rare.

(4*) *Mind-Body Dualism*[8]*:* Mental events are not physical events.[9]

the conjunction of any three of these seems to entail the negation of the remaining one. Thus, in the same way that the Exclusion Argument for physicalism combines (1), (2), and (3) to deny (4*), one could also employ (1), (3), and (4*) to argue against the Causal Self-Sufficiency of the Physical, (1), (2), and (4*) to argue against the Absence of Systematic Overdetermination, or (2), (3), and (4*) to deny the Causal Efficacy of the Mental.

---

[8] Here dualism refers to any position committed to (a), including epiphenomenalism. This applies also to all other occurrences of (4*).

[9] While (4*) is stated in terms of events rather than properties, I'll be relying on a property exemplification view of events of the sort advocated by Kim (1973, p.222) and defended by Goldman (1970, pp.1-10), according to which an event "consists in the instantiation of a property *P* by an object *x* at a time *t*." On such a model, the assertion that mental events are not physical events can be taken as a statement of dualism (in the broadest sense, as defined by (a)), inasmuch as it amounts to the claim that the properties exemplified in the case of mental events are distinct from those exemplified in the case of physical events. States can be understood on this model as including both events and persisting events, so that a state will consist in the instantiation of a property *P* by an object *x* throughout a span of time $t_1 - t_n$, where $n \geq 1$.

Seizing on this point, those who are less wedded to the causal efficacy of the mind than they are to its immateriality may be content to give up (1) in order to avoid the reduction of mental events to physical events that the Exclusion Argument for physicalism appears to require. Those who take this route can then modify the Argument as follows:

(4*) *Mind-Body Dualism*: Mental events are not physical events.

(2) *Causal Self-Sufficiency of the Physical:* Every physical effect has a sufficient physical cause.

(3) *Absence of Systematic Overdetermination:* Causal overdetermination is rare.

(1*) *Epiphenomenalism:* Therefore, mental events do not cause physical events.

The result is what has been referred to by Shapiro and Sober (2007) as the Master Argument for epiphenomenalism, which can be numbered as yet another thorn in the side of interactionist dualism.

Beset by these difficulties, interactionist dualists seem to face a choice between four equally unpleasant options. On the one hand, they can accept the conclusion of the Exclusion Argument for physicalism, and resign their commitment to the claim that the mental cannot be reduced to, identified with, or adequately explained by the physical. To take this route, however, would be to renounce their own position: one can't continue to be a dualist if one gives up on the idea that mind and matter are distinct. Second, they can seek to avoid the Exclusion Argument for physicalism by switching its first premise with its conclusion and negating both, thus transforming it into the Master Argument for epiphenomenalism. The conclusion of this argument is, however, no more amenable to interactionist dualism than that of the former, as it consists in the negation of another of

this variety of dualism's essential tenets; viz. that the mind has a causal impact on events in the physical world. Hence, neither of these two options is viable for anyone who wishes to retain a commitment to interactionist dualism, for either would require giving up one of the position's two most central theses (viz. (1) and (4*), or, correlatively, a. and b. above).

If, however, the impression that propositions (1), (2), (3), and (4*) are incompatible is correct, the only two choices that one who wishes to retain both the immateriality and causal efficacy of the mind has left are to transform the Exclusion Argument for physicalism into an argument either against the causal self-sufficiency of the physical realm ((4*) & (1) & (3) ⊃ ~(2)) or for widespread causal overdetermination ((4*) & (1) & (2) ⊃ ~(3)). Both of these choices would, however, involve contesting a widely accepted metaphysical principle, one of which (2) is purportedly backed by long-standing conservation laws of physics, and the other of which (3) draws support from some of our most basic intuitions about causation, viz. that causes are typically necessary for the occurrence of their effects. Interactionist dualism would appear to be in dire straits if its survival hinges on the refutation of either of these theses.

The Exclusion Argument thus seems to force interactionist dualist into the following double dilemma (hereafter referred to as the Exclusion Problem): they must either reject the completeness of physics or the apparent rarity of causal overdetermination, or else relinquish one of the constitutive tenets of their own position by embracing either the Exclusion Argument for physicalism (thereby securing the causal efficacy of the mind at the expense of its immateriality) or the Master Argument for epiphenomenalism (thereby avoiding the identification of mind with matter by giving up

its causal efficacy). Neither of the first two options looks promising, as each would require going up against a long-standing orthodoxy supported by a number of deeply entrenched theoretical assumptions, but if interactionist dualists balk at the prospect of challenging (2) and/or (3), their only other option is to give up on interactionist dualism altogether, by conceding either that the mind is physical or that it is causally inert.

Viewed in this light, the prospects for interactionist dualism look rather grim, and one naturally grows more sympathetic to the calls of many philosophers to abandon the position in favor of either a physicalist or an epiphenomenalist theory of mind, depending on which of the two central tenets of interactionist dualism (i.e., (1) or (4*)) one takes to be more worthy of salvaging.

While some have resigned themselves to seizing one of these two horns of the above double dilemma, there remain a number of defenders of the beleaguered doctrine who have sought to question the conception of causation in the physical realm (embodied in the conjunction of (2) and (3)) that appears to render a dualist theory of mind so problematic, rather than surrender the field by giving up either (1) or (4*). Others have sought to show that the assumption upon which the Exclusion Problem rests (viz., that the conjunction of (1), (2), (3), and (4*) is contradictory) is false, and that the apparent incompatibility of (1), (2), (3), and (4*) is merely illusory. Each of these ways of responding to the Exclusion Problem will be explored in the following chapters.

4. Summary of the remaining chapters

Our investigation of the Exclusion Problem and the prospects for interactionist dualism will begin in the following chapter with an examination of two further ways of responding to the Problem that have been advocated, respectively, by Tyler Burge (1993) and Lynne Rudder Baker (1993), and Frank Jackson and Philip Pettit (1990a; 1990b). Instead of openly contesting (2), (3), and/or the apparent incompatibility of (1), (2), (3), and (4*), Baker, Burge, Jackson, and Pettit maintain that the proper solution to the Problem lies rather in focusing on how appeals to mental events figure into informative and successful explanations of events in the physical world. After explaining why I find such solutions unsatisfactory, the remaining chapters (Chapters 3-9) will be spent examining the other strategies for resolving the Exclusion Problem noted above, along with the various arguments that seem to me to justify acceptance of (1) and 4* (keeping in mind that as far as (1) and (4*) are concerned, my aim is solely to render them plausible enough to motivate the search for a solution to the Exclusion Problem that does not require the acceptance of physicalism or epiphenomenalism).

The case for (4*), Mind-Body Dualism, is presented in Chapters 3-4. Chapter 3 outlines the structure of the argument for dualism to be filled in by Chapters 4-6, and develops a definition of "realization" that will serve as the proposed relation between mind and body. Chapters 4-6 then argue for the view that mental properties and events are realized by, but distinct from, irreducible to, and incapable of being fully explained in terms of physical properties and events. Chapters 4 and 5 cite the multiple realizability of mental properties as grounds for rejecting the identification of mental properties with physical properties and the reducibility of psychology to neuroscience. Chapter 4 provides some empirical support for the view that mental states involved in perceptions

of color are multiply realizable, and responds to three objections to the alleged multiple realizability of mental properties raised respectively by Kim (1992b) and David Lewis (1980), William Bechtel and Jennifer Mundale (1999), and Lawrence Shapiro (2000). Chapter 5 then reconstructs Jerry Fodor's (1974) argument from the multiple realizability of mental states to the autonomy of psychology and the irreducibility of mental properties, and defends this argument against criticisms raised by Kim (1992b) and John Bickle (1998).

Having given some reasons for denying that mental properties are identical with physical properties in Chapters 4 and 5, Chapter 6 proceeds to argue that mental events (i.e. instances of mental properties) cannot be token-identified with physical events (i.e. instances of physical properties). After noting some potential inconsistencies in the non-reductive physicalist position that maintains that mental events are token- but not type-identical with physical events, Davidson's (1970) argument for the view is examined and rejected. A series of arguments are then offered that point out various obstacles that the distinctive phenomenal and intentional features of mental events present to attempts to identify instances of mental properties with instances of physical properties by analyzing mental properties in strictly functional terms.

Having argued in Chapters 4-6 that there are good grounds for accepting dualism, and that a solution to the Exclusion Problem that does not require the identification of mental properties and events with physical properties and events is thus something worth looking for, Chapters 7 and 8 explore two different strategies for resolving the Problem in a manner consistent with interactionist dualism. Chapter 7 begins by offering some objections to epiphenomenalism to discourage those who are swayed by the arguments

against physicalism advanced in the preceding chapters from thinking that the next best way of resolving the apparent conflict between (1), (2), (3), and (4*) is to simply reject (1), the Causal Efficacy of the Mental. The remainder of the chapter then focuses on (3), the Absence of Systematic Overdetermination, and develops an argument to the effect that, depending on causal overdetermination is defined, (3) is either implausible or else consistent with the conjunction of (1), (2), and (4*). If this is correct, then interactionist dualists are entitled to respond to the Exclusion Problem by either rejecting (3) (if overdetermination is defined in such a way that (3) is implausible) or else maintaining that their position is perfectly consistent with both (2) and (3).

Chapter 8 focuses on (2), the Causal Self-Sufficiency of the Physical. After noting some difficulties that E.J. Lowe (2000) points out in providing a formulation of (2) that is strong enough to rule out dualist interactionism without collapsing into the question begging assertion that *only* physical events cause physical effects, the bulk of the chapter is spent investigating whether (2) is indeed supported by conservation laws of physics, as is often supposed. After examining the issue, I argue that this is not the case, and that (2) does not follow or derive strong inductive support from any scientific law. If this is so, then interactionist dualists can also resolve the apparent inconsistency between (1), (2), (3), and (4*) by denying (2) without thereby setting themselves at odds with current science. I close the chapter by reviewing some challenges that quantum mechanics seems to present to the completeness of physics and Nancy Cartwright's (1994) argument that the laws of physics hold only in such limited cases as can be adequately modeled, both of which provide positive grounds for thinking that (2) is in fact false.

The task of finding a theory of causation that will render (1) both plausible and compatible with (4*) is reserved for Chapter 9. Two candidate theories of causation are presented under which non-physical, mental events may qualify as causing physical effects in a way that is consistent with one of the solutions to the Exclusion Problem proposed in Chapters 7 and 8. The first defines causation as consisting in the transference of energy or some other conserved quantity from one thing to another. A method of ascribing quantities of energy to mental states is proposed that would enable dualists to maintain that mental events cause physical effects by transferring energy to or receiving it from them. The second theory of causation derives from the work of Peter Menzies (2003; 2004), who defines causation as a process that is "picked out" by model-relative relations of counterfactual dependence between causes and effects. Depending on how the relevant "process" connecting mental causes to their physical effects is construed, Menzies' model is shown to allow for the causation of physical effects by non-physical, mental causes in a manner consistent with either of the two solutions to the Exclusion Problem proposed in Chapters 7 and 8. The availability of these accounts of dualist mental causation is taken to answer the lingering worry that even if interactionist dualism can overcome the Exclusion Problem and thereby demonstrate that the non-physical nature of mental events needn't render them causally redundant, it is still incomprehensible how a physical effect could be produced by a non-physical, mental cause.

A final note regarding the contribution that this dissertation aspires to make to the existing literature on the Exclusion Problem: As the Exclusion Problem has been a topic of discussion in the philosophy of mind for over 20 years now, it is difficult to find

anything to say about it that has not already been said. For this reason, apart from a few exceptions, there is very little in what follows that is entirely original. The contribution of this work, such as it is, hence lies not in the proposal of any radically new ideas, but rather in the synthesis of a large body of literature, from which I've picked out those theories, arguments, objections, and replies that strike me as the strongest and most relevant to the point I wish to make, which is that the Exclusion Problem is no reason to give up on dualism. By reconstructing, reorganizing, and evaluating the ideas of others, my aim has been to carve out from the mass of existing views on this issue the most compelling case for, and defense of interactionist dualism available within the current state of the debate.

CHAPTER 2

HOW NOT TO SOLVE THE EXCLUSION PROBLEM

As mentioned in the previous chapter, in addition to the options of rejecting (2) and/or (3), or maintaining that (1), (2), (3), and (4*) are not incompatible, there are two further ways of addressing the Exclusion Problem that have been proposed, respectively, by Tyler Burge (1993) and Lynne Rudder Baker (1993), and Frank Jackson and Phillip Pettit (1990a; 1990b). Rather than grappling directly with the Exclusion Problem in the manner of the above strategies (i.e., by contesting (2) and/or (3), or challenging the apparent incompatibility of (1), (2), (3), and (4*)), these two approaches share the view that a more satisfactory solution to the Exclusion Problem can be achieved by attending to the role that mental events play in our explanatory practices. By doing so, the kinds of psychological explanations of physical effects that are typically accepted as legitimate can, they argue, be seen to exhibit certain features that suffice to safeguard the causal efficacy (or at least, according to Jackson and Pettit, the causal relevance) of mental events against any worries that the Exclusion Problem might be thought to raise.

1. Baker and Burge's argument from explanatory success

Baker and Burge share the view that the apparent *success* of psychological explanations of physical effects suffices to enable such explanations to ensure that mental events are causally efficacious whether they are physical or not. Although neither is entirely clear as to what sorts of standards an explanation must meet in order to qualify as

successful[10], both are adamant that successful explanations do not require any corroboration from metaphysics. As Baker (1993, p.94) puts it: "Systematic explanatory success, in either science or everyday life, stands in need of no metaphysical underpinning." In a similar vein, Burge (1993, p.117) states: "As long as mentalistic explanation yields knowledge and understanding, and as long as that explanation is (sometimes) causal, we can firmly believe that mind-body causation is a part of the world…[M]entalistic explanation and mental causation do not need validation from materialist metaphysics."

Judging from these quotes, and their subsequent rejection of the constraints the Exclusion Problem imposes on the possibility of mental causation, Baker and Burge seem to have something like the following argument in mind:

(a) Metaphysical arguments can give us no reason to believe anything that stands at odds with successful explanatory practices.

(b) Psychological explanations of physical effects are causal explanations.

(c) Such explanations are highly successful.

(d) There is nothing internal to our practice of explaining physical events in terms of mental causes that would suggest that such causes must be identical with or otherwise reducible to physical events.

[10] While Burge (1993, pp.111, 117, 119) speaks of our mentalistic "explanatory scheme" as being "systematic, informative, [and] important," as "yield[ing] knowledge and understanding," and as "work[ing] very well," Baker (1993, p.93) simply characterizes psychological explanations as "explanations that earn their keep." Both do seem, however, to associate explanatory success with an ability to support counterfactuals (Baker, 1993, p.93; Burge, 1993, p.115), and in a later work, Baker (1995, p.122) introduces an interventionist/manipulationist "Control Test," which states that "we know that we have an adequate causal explanation when it affords control over the phenomena of the type explained."

(e) Therefore, no metaphysical argument (e.g., the Exclusion Problem) can give us any reason to believe that the mind is epiphenomenal, or that it is capable of causing physical effects only if it is itself physical.

If this constitutes an accurate representation of their view, then according to Baker and Burge, the apparent success of psychological explanations of physical events is enough to justify our classifying the mental events cited as *explanantia* in such explanations as causes of the physical events they explain, regardless of whether or not these mental causes can be integrated into a network of physical causation and explanation by being identified with or reduced to purely physical events, metaphysical arguments to the contrary notwithstanding. Insofar, then, as the Exclusion Problem places the causal efficacy of mental events in jeopardy by holding it hostage to a physicalist reduction of mind, the Problem and the metaphysical assumptions upon which it is based are to be rejected via *modus tollens*, as standing at odds with explanatory practices whose success renders them impervious to criticism on purely metaphysical grounds.

While this line of thought seems to offer a quick and simple way of dismissing the Exclusion Problem, it rests, I think, on a crucial error. The basic difficulty with Baker and Burge's argument is that in offering an explanation, one incurs an ontological commitment to the existence of whatever one refers to in the course of providing that explanation. This commitment extends not only to the (putative) entities cited in the *explanans* and *explanandum*, but also to the relation between the two that is supposed to explain the latter, even in cases where reference to this relation is only implicit. These commitments can be seen to follow from the fact that explanations are naturally understood as answers to "Why?" questions. As such, their primary aim is presumably to

satisfy some (perhaps hypothetical) audience's request for knowledge about the reason for something's being the way it is. From this it follows that any explanation that appeals to things that do not to exist falls short of its aim, and must be rejected as unsatisfactory[11], for any such explanation will include terms that have no referents, and will hence fail to be true. Consequently, instead of supplying those raising the relevant "Why?" question with the knowledge they've requested, accepting such an explanation would only increase their number of false beliefs.

The upshot of this is that the legitimacy of any causal explanation (psychological explanations of bodily movement included) requires that there be some real causal relation between the events cited in its *explanans* and *explanandum*. Any reason one might have to doubt that there is in fact such a relation between these events is therefore *ipso facto* a reason to doubt that the explanation should be accepted as valid. Seeing, then, as the Exclusion Problem constitutes just such a reason for questioning the existence of any causal relations between mental and physical events unless the former can be reduced to or identified with the latter, the Problem gives us grounds for questioning the autonomy or legitimacy of the psychological explanations whose

[11] An important exception may be presented by explanations whose *explanantia* and/or *explananda* include absences or omissions (e.g. an explanation that appeals to the gardener's failure to water the plants to account for the plants' dying). Such explanations may be satisfactory and successful even though omissions (e.g. the gardener's failure to water the plants) do not exist and hence cannot be referred to. An alternative treatment of such explanations that might render them consistent with the above remark would be to treat statements about omissions as referring to existing *facts* (e.g. the fact that the gardener did not water the plants), and to include facts among the *relata* of the explanatory relation. However they are interpreted, though, explanations of or by omissions do not raise the same problems as other explanations that make use of non-referring terms, for since it is commonly understood that omissions do not exist, audiences will not be prone to form false beliefs in the existence of such things when accepting explanations that cite them as *explanantia* and/or *explananda*.

apparent success Baker and Burge claim is sufficient to establish the causal efficacy of the mind, regardless of whether or not the mind is physical.

These problems generate a dilemma for Baker and Burge's argument that focusing on the notion of explanatory success helps make clear. Either an explanation must be true in order to be successful, or an explanation can be successful without being true. If Baker and Burge accept truth as a necessary condition for explanatory success, then it follows trivially that the success of causal explanations with mental *explanantia* and physical *explananda* entails that mental events do cause physical events. However, if this is what Baker and Burge mean by explanatory success, then to claim that psychological explanations are successful and therefore "do not need validation from materialist metaphysics" is to beg the question against the Exclusion Problem, the main contention of which just is that the truth of such explanations is conditional upon a physicalist reduction of the mind (Burge, 1993, p.117). In other words, the Exclusion Problem offers us some reasons for thinking that the mind must be physical in order for it to cause bodily motions. Those who find these reasons compelling will hence hold that any mental events cited as *explanantia* in causal explanations of physical effects must be physical if such explanations are to be true. And if explanations must be true in order to qualify as successful, then those swayed by the Exclusion Problem will likewise view the success of causal explanations of physical effects in terms of mental events as requiring that mental events be physical in nature. Assuming, then, that successful explanations must be true, to say that the success of psychological explanations is sufficient to establish the causal efficacy of the mind regardless of whether or not the mind is physical is to simply reject the reasons that the Exclusion Problem gives for thinking that such

explanations can only be true if mental events are identical with physical events without offering any grounds for thinking that those reasons are flawed.

On the other hand, if in characterizing psychological explanations as successful, Baker and Burge mean something short of saying they are true, then the success of psychological explanations is no proof that mental events are causally efficacious, for such explanations could very well turn out to be false. The most that the success of these sorts of explanations could offer, then, is evidence in favor of the mind's being causally efficacious. But in that case there seem to be no legitimate grounds for refusing to take metaphysical arguments placing further constraints on the actual truth of psychological explanations into consideration, even if these arguments appear to show that mental events must be physical in order to cause physical effects.

In sum, if the success of an explanation is taken to entail its truth, then the Exclusion Problem suggests that psychological explanations are successful only if mental events are physical, and if an explanation can be successful without being true, then the Problem casts doubt on the inference from the alleged success of psychological explanations to the claim that the mind is causally efficacious whether it is physical or not.

Perhaps the driving motivation behind Baker and Burge's view that metaphysics has no bearing on the legitimacy of psychological explanations or the conclusions we are entitled to draw from their apparent success is the idea that the primary, if not the only evidence we have as to what causal relations there are in the world consists in the regularities we encounter in experience, and that our beliefs about causation ought therefore to be guided solely by those explanatory practices that have proven the most

effective in capturing and systematizing these regularities while leading to the discovery of still more. From this perspective, the metaphysical worries raised by the Exclusion Problem are at first glance apt to appear overly removed from the kinds of empirical considerations that are alone relevant to determining the validity of our causal explanations and ascriptions of causal efficacy.

This appearance, however, is ultimately misleading, for the metaphysical assumptions that give rise to the Exclusion Problem (viz., (2) and (3)) are themselves supported by empirical evidence and the success of certain explanatory practices currently employed in the natural sciences and everyday discourse. Thus, as David Papineau (2001, p.27) argues, the main reasons for accepting (2), the Causal Self-Sufficiency of the Physical derive from reflection on the fact that (a) as physics and the other natural sciences have developed, appeals to "special" non-physical forces have been repeatedly eliminated as new physical mechanisms are discovered that are capable of accounting for all that such "special forces" were originally postulated in order to explain, thereby lending empirical support to the inductive inference that all physical events will ultimately turn out to have sufficient physical causes, and that (b) physiological research into the operations of living bodies has simply produced "no direct evidence" for the existence of any "special forces", thus suggesting that even in the case of organisms, where such non-physical "vital or mental" forces would be most likely to appear, everything can be adequately accounted for in purely physical terms.[12] Due in large part to these very developments, the Causal Self-Sufficiency of the Physical is now

[12] Papineau's views on this topic will be examined and criticized at length in Chapter 8. Here it is enough to note that the considerations he cites provide at least apparent empirical support for (2).

widely accepted as a basic theoretical postulate that guides much of the research and explanatory practices of the contemporary natural sciences, the success of which corroborates Georges Rey's (2001, p.100) remark that "the attempt to fit spatio-temporal phenomena in general into mechanistic and ultimately physical accounts of the world lies at the heart of the most successful explanatory practices we know."

Experience also appears to lend support to (3), the Absence of Systematic Overdetermination. Though cases can be presented in such a way that we sometimes find ourselves unsure of what to say, we are generally quite good at distinguishing instances of overdetermination from those "standard" cases wherein an effect is produced by a single cause. No deep reflection is needed for us to correctly classify the death of a man shot by a firing squad, or the shattering of a vase simultaneously struck by a bat and a hammer as overdetermined effects. When we look into the circumstances surrounding any event that we encounter in our experience, however, it so happens that we are almost always able to isolate a single salient cause that (in conjunction with the requisite background conditions) was sufficient to produce it and without which the effect wouldn't have occurred.[13] Given that we are fairly adept at telling cases of overdetermination and single-cause causation apart, it seems that either overdetermination is indeed rare, as our experience seems to suggest, or else overdetermination is quite frequent, but oftentimes all but one of the overdetermining causes somehow escapes our notice. If the latter were the case, then closer inspection should reveal many effects to be overdetermined that did not at first appear to be so. But

[13] Although, as will be argued in Chapter 9, the event that we pick out as *the* cause of a given effect may differ depending on the context of our inquiry into that effect's occurrence.

this is not the case; even after thorough investigation, a single cause is almost always what we find. Empirical evidence hence weighs strongly in favor of (3).[14]

The support (3) draws from experience also exerts a noticeable influence on our explanatory practices, which is manifested in the fact that although citing only one of two overdetermining causes in the causal explanation of an overdetermined event is objectionable, in that by doing so, one is liable to generate the (in this case misleading) implicature that the *explanandum* wouldn't have occurred if the *explanans* hadn't, we don't generally hedge against this eventuality by offering or accepting causal explanations only on the understanding that they be taken provisionally, pending the discovery of additional, overdetermining causes of the *explanandum*, which if found, would invalidate the explanation given. One does not say "*x*, because *y*, *if, that is, I'm correct in assuming that there was no further event besides y that would have by itself been sufficient to cause x*." The qualification is left out because it is assumed by all parties to be so likely to be true. If, on the other hand, we encountered causal overdetermination much more frequently in our experience than we in fact do, we would expect such qualifications to occur more frequently in our causal explanations as well, since by omitting them, one would run a much greater risk of offering an unsatisfactory explanation. By forgoing such caveats, our explanatory practices hence involve an implicit commitment to the Absence of Systematic Overdetermination, which appears to pose no obstacle to their success.

[14] This empirical evidence for (3) might also be supplemented or corroborated by conclusions arrived at through conceptual analysis, which may tell us, e.g., that it is part of the concept of a cause that causes are typically necessary for their effects.

Reflection on the course of the development of the natural sciences and the continued success of the explanatory scheme that these sciences currently employ thus seems to provide grounds for accepting (2) that are every bit as empirical and rooted in successful explanatory practices as those that Baker and Burge cite in favor of the causal efficacy of the mind. Attention to our facility in identifying cases of causal overdetermination and the results we obtain when we bring this ability to bear on our own experience suggests that the same is true of (3). It is thus the deliverances of experience and an appreciation for explanations that work, not *a priori* reasoning or unconstrained metaphysical speculation, that suggest to us that every physical event has a sufficient physical cause, and that effects are rarely overdetermined. Seeing, then, as these two propositions are all that's needed to generate the Exclusion Problem, the Problem cannot in fairness be dismissed as relying on metaphysical assumptions that have no basis in or relevance to the kinds of explanatory practices that have thus far proven the most successful in making sense of actual experience.[15] This is precisely what makes the Problem so troublesome: while on the one hand, it seems empirically self-evident that mental and physical events are in some significant sense distinct, and that the

[15] Part of the problem may be that Baker (1993, pp.78-9) formulates the Exclusion Problem as relying on "the thesis of the 'causal closure of the physical'," where "a system is causally closed if and only if the elements of the system interact causally *only* with other elements of the system; there is no causal influence from 'outside' the system." It may be that Burge is operating under the assumption that the Exclusion Problem requires a commitment to this thesis as well. If so, it would help to explain why both are so insistent that the assumptions that give rise to the Problem are given too much weight if taken as reasons for questioning the validity or autonomy of psychological explanations, for Baker's "causal closure" thesis is much stronger than (2), (which is entailed by, but does not entail the former), and it can thus be much more plausibly argued that this thesis lacks the empirical support and basis in successful explanatory practices that the ascription of causal efficacy to mental events (without any requirement that they be identified with physical events) has. However, since the Exclusion Problem requires only the weaker premise (2) (which seems to have substantial empirical support), to the extent that Baker and Burge's objections to the Problem presuppose its dependence on the "causal closure" thesis, they commit a straw man fallacy.

state of our minds has a definite causal impact on the movement of our bodies, on the other hand, our experience seems likewise to indicate that there is a sufficient physical cause for every physical effect, and that events are not systematically overdetermined. Since it appears, upon reflection, that these four claims cannot all be true, one seems forced to conclude that our experience is in some way misleading. The difficulty lies in determining how so.

In sum, while Baker and Burge may be right in claiming that experience ultimately gives greater support to the causal efficacy and immateriality of the mind than it does to the causal self-sufficiency of the physical and the rarity of overdetermination, a satisfactory defense of this claim must do more than simply assert that the success of psychological explanations outweighs any argument for physicalism or epiphenomenalism from (2) and (3). Further reasons must be given for thinking that the empirical evidence *for* (2) and/or (3) is somehow inadequate or equivocal, or that the apparent conflict between (1), (2), (3), and (4*) is merely apparent.

2. Jackson and Pettit: psychological explanations as "program" explanations

While sharing Baker and Burge's view that a satisfactory response to the Exclusion Problem requires close attention to the role mental events play in explanatory practices, Jackson and Pettit develop this idea in a very different direction. Rather than trying to maintain, like Baker and Burge, that the success of psychological explanations is sufficient to establish the causal efficacy of mental events, Jackson and Pettit hold the view that mental events are epiphenomenal. They are led to this conclusion by a

generalized version of the Master Argument for epiphenomenalism, which reasons that since all events presumably supervene on, or are in some other way determined by "how things, including the laws, are at the most fundamental micro-physical level," the ascription of causal efficacy to any events aside from those that consist in the instantiation of "certain properties in fundamental Physics" is superfluous, "[f]or we do not need to believe in any fundamental efficacies over and above those between properties at the micro-level in order to explain the regularities, actual and counterfactual, all the way up, because supervenience tells us that they are fixed by how things are at the bottom (*if* there is a bottom[16])" (Jackson and Pettit, 1990b, p.209).[17]

This argument can be presented more precisely as follows:

(i) Events at the micro-physical level are distinct not only from mental events but also from macro-physical events.

(ii) All non-micro-physical events supervene on events at the micro-physical level.

(iii) Every micro-physical event has a sufficient micro-physical cause.

(iv) If an event (or collection of events) e_1 supervenes on another event (or collection of events) e_2, then any direct causation of e_1 is redundant.

(v) There is no such redundant causation.

---

[16] As will be seen below, this qualification represents a somewhat blithe recognition of a rather serious objection to Jackson and Pettit's proposal.

[17] Jackson and Pettit (1990b, p.209) note that this argument rests on the (I think quite reasonable) "view of causation which does not reduce it to nothing more than nomological sequences," for if one reduces causation to law-governed sequences in this way, then the existence of such sequences of macro-physical or mental events and physical events would be enough to establish the causal efficacy of mental and macro-physical events. Reasons for rejecting such Humean "regularist" theories of causation will be offered in Chapter 6.

(vi) Causal overdetermination is rare, and, among physical events, occurs only among events of the same "level" (e.g., no micro-physical effect can have two independently sufficient causes, one of which is also micro-physical, and the other of which is not).[18]

(vii) Therefore, all non-micro-physical events are causally inert.

Thus, if m_2 is any mental or macro-physical event, (i) and (ii) entail that m_2 supervenes on some micro-physical event p_2, and that $m_2 \neq p_2$. This supervenience relation is represented in the diagram below by the double arrow (⇑) from p_2 to m_2. (iii) entails that there is some micro-physical event p_1 that causes p_2. This is represented by the single black arrow (→) from p_1 to p_2. (iv) and (v) entail that no mental or macro-physical event m_1 directly causes m_2. This is represented by the single red arrow from m_1 to m_2 (→). (This holds regardless of whether m_1 supervenes on the micro-physical cause (p_1) of the micro-physical event on which m_2 supervenes (p_2), or instead on some other micro-physical event, which fact is represented by the parenthetical double arrow ((⇑)) from p_1 to m_1.) (iii) and (vi) entail that no mental or macro-physical event causes m_2 by way of causing the subvenient micro-physical base of m_2 (again, regardless of whether that event supervenes on the micro-physical cause of p_2, or on some other micro-physical event instead). This is represented by the single red arrow (↘) from m_1 to p_2. As (vii) states, this

[18] Premises (i), (iii), and (vi) are analogous to (4*), (2), and (3), respectively. Note that (v) differs from (vi) in that since supervenience is a non-causal relation, the type of redundancy ruled out by (v) is not a form of *causal* overdetermination, which concerns only those cases wherein an effect is produced by two independently sufficient causes. In contrast, the kind of case (v) excludes is one wherein an event (or collection of events) simultaneously supervenes on one event (or collection of events), while also being caused by another.

leaves causal relations between micro-physical events (e.g. p_1 and p_2) as the only kind of causation there is.

$$\begin{array}{ccc} m_1 & \rightarrow & m_2 \\ (\Uparrow) & \searrow & \Uparrow \\ p_1 & \rightarrow & p_2 \end{array}$$

Unlike Baker and Burge, who maintain that the Exclusion Problem does not give us sufficient reason to question the mind's ability to cause physical effects, Jackson and Pettit thus see the above extension of the Exclusion Problem as supplying adequate grounds to reject not only the causal efficacy of mental events, but also that of all physical events aside from those that occur at the most basic physical level.[19] Jackson and Pettit insist, however, that this conclusion does not in any way threaten the legitimacy and success of psychological explanations, nor does it entail that mental and macro-physical events have no causal *relevance* to the occurrence of events in the physical world. This is because they deny that "[t]he only way for a property to be causally relevant to the production of a certain effect is by being causally efficacious in the process of production" (Jackson and Pettit, 1990a, p.111). Hence, even though mental and macro-physical properties are, according to Jackson and Pettit, causally inert, they may yet have causal relevance to the occurrence of certain physical effects.

In saying this, Jackson and Pettit are not suggesting that just any property can be construed as causally relevant to a given effect. They instead hold that "we can distinguish among properties in respect of their causal relevance to the obtaining of some

[19] The question whether the Exclusion Problem generalizes in this way so as to render all non-micro-physical causation suspect will be discussed further in Chapter 7.

effect or other," and that "[a] causal explanation of something must direct us to a causally relevant property as opposed to a causally irrelevant property of the factor it identifies as explanatory: a property relevant to the causal production of the effect explained" (Jackson and Pettit, 1990b, p.197; 1990a, p.108). Thus, to borrow an often-cited example of Fred Dretske's (1988, p.79), if a soprano sings a high-C, thereby causing a glass to shatter, it is the acoustic properties of the sound she emits, and not the meaning of whatever word(s) she happens to be singing, that are causally relevant to the glass' shattering and consequently suitable for use in a causal explanation of that event. Jackson and Pettit's point, then, is that among those properties that are causally relevant to the production of a given effect, there may be some that themselves played no part in bringing that effect about.

One might naturally wonder how this can be so. How can an event[20] be causally relevant to or figure into a satisfactory causal explanation of an effect it didn't cause?[21] The primary if not the only way that Jackson and Pettit claim this can happen is for an event to figure as an *explanans* in what they call a "program explanation" of a certain effect. The distinguishing feature of such explanations is that "[t]he realization of the property [cited in the *explanans*] ensures…that a crucial productive property is realized and, in the circumstances, that the [*explanandum*] event, under a certain description,

[20] Although Jackson and Pettit usually speak in terms of properties, since I'm employing a property-exemplification model of events, event and property language can be used more or less interchangeably here.

[21] Kim (1998, p.75) simply rejects this possibility, insisting instead that "a causal explanation of an event that invokes another as its cause can be a correct explanation only if the putative cause really is a cause of the event to be explained. Any weaker conception would merely cheapen the idea of causal explanation."

occurs" (Jackson and Pettit, 1990a, p.114). In other words, the *explanans* of a program explanation does not itself cause the *explanandum*, but is rather such that its occurrence ensures that some *other* event that *is* causally sufficient (in the relevant situation) for the *explanandum* also occurred and thereby caused the *explanandum* to occur as well. Program explanations are thus distinguished from "process explanations," which cite as their *explanantia* events that actually caused the event to be explained, by the fact that the *explanantia* of program explanations explain their *explananda* not by causing them, but rather by ensuring the occurrence of other events that do. As Jackson and Pettit (1990a, p.114) put it: "The property-instance [that serves as the *explanans* of a program explanation] does not figure in the productive process leading to the [*explanandum*] event but it more or less ensures that a property-instance which is required for that process does figure." Though inefficacious in the production of the events they explain, events that are suitable for use as *explanantia* in program explanations are nevertheless said to retain causal relevance by virtue of ensuring or "programming for" the occurrence of some event that is sufficient to cause the *explanandum* (thereby "programming for" the occurrence of the *explanandum* as well) in the same way that "a computer program…ensures that certain things will happen – things satisfying certain descriptions – though all the work of producing those things goes on at a lower, mechanical level" (Jackson and Pettit, 1990a, p.114).

With this distinction between program and process explanations in place, Jackson and Pettit (1990a, p.115) claim that "there are at least two distinct ways in which a property can be causally relevant: through being efficacious in the production of whatever is in question, or through programming for the presence of an efficacious

property." In light of their view that all causal efficacy resides at the micro-physical level, the only true process explanations will, according to Jackson and Pettit, be those given in purely micro-physical terms. The vast number of the causal explanations we make use of (including psychological explanations) will hence turn out to be program explanations. Thus, though they contend that mental properties are epiphenomenal, by allowing that psychological explanations are nonetheless perfectly legitimate program explanations, Jackson and Pettit leave such properties in good and quite plentiful company, which will include, notably, all the properties of the special sciences as well as many of those that we commonly cite in everyday descriptions of macro-physical objects.

Given, however, that all real causal relations are captured by process explanations given solely in terms of micro-physical events, what use can the remaining multitude of program explanations have? The importance of program explanations is held by Jackson and Pettit to consist in the fact that they convey certain modal information about an *explanandum*'s causal history that a process explanation of that *explanandum* may fail to provide. For this reason, Jackson and Pettit (1990a, p.116) maintain that "a program explanation can have a significance that remains in the presence of an explanation invoking the corresponding efficacious property…[I]t may be an explanation which the process explanation does not supersede." Jackson and Pettit's (1990a, p.117) idea is that whereas "[t]he process story tells us about how the [causal] history [of the *explanandum*] actually went…[a] program account tells us about how that history might have been." For instance, making use of a famous example of Hilary Putnam's, Jackson and Pettit (1990a, pp.110, 115) state that whereas a process explanation of one's inability to insert a square peg through a hole whose diameter is equal to the square's side will cite only the

molecular structure of that part of the peg that falls outside of the hole's circumference that is responsible for its impenetrability, a program explanation of this same *explanandum*[22], given in terms of the geometric properties of the peg and the hole, will enable us to know that any of the indefinite ways of realizing those same geometrical properties in any impenetrable substances would produce the same result.

One immediate difficulty with Jackson and Pettit's proposal as presented thus far is that it seems to lead to an overly permissive distribution of causal relevance. Recall that any event that can serve as an *explanans* in a program explanation is, according to Jackson and Pettit, causally relevant to the occurrence of the *explanandum*. This is what enables them to offer us the consoling assurance that mental and macro-physical events, though epiphenomenal, are nonetheless still causally relevant to the production of physical effects. The problem, however, is that given Jackson and Pettit's (1990a) definition of program explanations as explanations whose *explanantia* are themselves causally inefficacious, but which by their presence ensure the occurrence of some other event that is sufficient to cause the *explanandum*, many events will qualify for use as *explanantia* in program explanations that are clearly irrelevant to the occurrence of the effects they are adduced to explain.

[22] See, however, Walter (2005, p.36), who notes that in claiming that process explanations explain what *actually* produced a given event, whereas program explanations explain how that event *might have been* produced, Jackson and Pettit thereby assign these explanations different *explananda*. This stands in tension with their claim that program explanations can provide information about a given *explanandum* that process explanations fail to supply, as this claim is only non-trivial if the explanations share the same *explanandum*.

As an instance of such a counterexample, Jackson and Pettit themselves ask us to consider the case of a person who places an aluminum ladder against some power lines, and is thereupon electrocuted. As they note:

> The categorical basis in metals of the different dispositional properties of electrical conductivity, thermal conductivity, ductility, metallic lustre and opacity is essentially the same, namely, the nature of the cloud of free electrons that permeates the metal. Nevertheless, the person who dies because she allows her aluminum ladder to touch power lines does not die because her ladder is a good conductor of heat, or because it is lustrous or ductile or highly opaque; she dies because her ladder is a good electrical conductor. Although one and the same property is the categorical basis of all these dispositions[23], out of these dispositions it is only being a good electrical conductor which is causally relevant to her death…[T]he fact that there is one categorical basis for the various dispositions does not mean that the various dispositions are alike in causal relevance. (Jackson and Pettit, 1990b, p.204)

The trouble is that while the cloud of free electrons that permeates the ladder is, given Jackson and Pettit's restriction of causal efficacy to micro-physical properties, the sole property of the ladder that is actually efficacious in causing the person's death, *any* of the various dispositional properties of the ladder that Jackson and Pettit mention would suffice to ensure the presence of that property, since it serves as the categorical basis of each of them. Hence, according to Jackson and Pettit's (1990a) definition of program explanations and causal relevance, all of these dispositional properties will turn out to be causally relevant to, and suitable for use in a program explanation of the person's death by electric shock, for each of them ensures, by its presence, the presence of some property (viz., the cloud of free electrons) sufficient to produce that effect. But surely we

[23] One might wonder whether these various dispositions indeed all have the same categorical basis. Even if they don't, the example will still illustrate its intended point so long as the categorical basis of the ladder's being a good electrical conductor is nomologically linked with that of one of its other dispositions in such a way that it can't have the latter disposition without also having the micro-physical property that makes it a good electrical conductor.

don't want to say that the ladder's ductility or thermal conductivity are equally relevant to the person's death as its being a good conductor of electricity. To do so would stretch the notion of causal relevance beyond all recognition.

The problem is quite general. As Sven Walter (2005, p.36) points out, it "arises whenever two or more properties have the same realizers but are not all causally relevant for the occurrence of an effect."[24] Jackson and Pettit's (1990a) proposal is thus far too liberal in its ascription of causal relevance. Indeed, if it were correct, then "any property which *supervenes* upon a causally relevant property" would likewise count as causally relevant as well (Walter, 2005, p.36). Since this is clearly not the case (e.g., while the evenness or oddness of my weight in pounds supervenes on my actual weight, the former clearly does not have the latter's causal relevance to the splash I make when I jump into a pond), it follows, *pace* Jackson and Pettit's (1990a) thesis, that "[i]t can…not suffice for a property's being causally relevant for an effect that it ensures the instantiation of a property which does the 'causal work' required to bring about that effect" (Walter, 2005, p.36).

To avoid counterexamples of the sort just described, in their 1990b article, Jackson and Pettit introduce a different definition of program explanations and causal relevance which states that a property is causally relevant to a given effect and therefore suitable for use as an *explanans* in a program explanation of that effect if and only if it ensures what they call "invariance of effect under variation of realization" (Jackson and Pettit, 1990b, p.204). The idea is that those properties that Jackson and Pettit maintain are

[24] Walter (2005, p.37) and Jackson and Pettit (1990b, pp.201-3) both note that similar difficulties beset Kim's (1984) theory of supervenient causation.

causally inert yet relevant to the production of a given effect are so by virtue of the fact that, of all the various micro-physical states that realize or in some other way suffice for the presence of the causally inert yet relevant property, each, if present, would have been sufficient for the production of the effect, and one of these was in fact present. This is meant to exclude counterexamples like the ladder case by classifying as causally irrelevant such properties as could have been realized by micro-physical states that would have been unable to produce the effect under consideration. Thus, while the ladder's being a good conductor of electricity remains causally relevant to the person's death, since any of the various ways in which this property could have been realized would have been equally capable of producing that effect, "the reason opacity, say, is not causally relevant to her dying is that it might easily have been realized without her dying – as would, for instance, have been the case had the ladder been wooden" (Jackson and Pettit, 1990b p.205).

Jackson and Pettit's (1990b) modified definition of causal relevance does not fix the problem. Equating causal relevance with "invariance of effect under variation of realization" still forces us to ascribe causal relevance to many events that clearly do not have it.[25] One can thus still construct counterexamples to Jackson and Pettit's (1990b) proposal consisting of properties that most would agree are of no causal relevance to a given effect but whose possible realizers are a subset of the possible realizers of some

[25] Indeed, one might already doubt whether this definition of causal relevance even gives an adequate response to the ladder case, for while it is easy to "imagine a realizer of *being opaque* that does not *ipso facto* realize *being a good electrical conductor*," one might find it much more difficult to imagine "a possible realizer of *being a good thermal conductor* that does not *ipso facto* realize *being a good electrical conductor*," especially if possibility here refers to *physical* or *nomological* possibility (Walter, 2005, p.39).

property that is. As an instance of such a case[26], consider the properties of *having a belief that there are apples in the cupboard* and *having a belief that there are shiny red apples in the cupboard*. Since the latter property is a more determinate instance of the former, the various ways in which it might be realized constitute a proper subset of the possible realizers of the former. Imagine, then, that I exemplify both these properties, and someone asks me if there are any apples left in the cupboard. I reply, "Yes, there are." It would seem that in this instance, it is my belief that there are apples in the cupboard, and not my belief that those apples are red and shiny, that is causally relevant to my response. Yet since the possible realizations of *having a belief that there are shiny red apples in the cupboard* constitute a proper subset of the possible realizations of *having a belief that there are apples in the cupboard*, anything exemplifying the former property must exemplify the latter property as well. And since my exemplifying the latter property is (*ceteris paribus*) sufficient to ensure my emitting an affirmative response in answer to the above query, exemplifying the former property would also ensure that same effect, regardless of how it is realized. In the imagined case, then, the property of *having a belief that there are shiny red apples in the cupboard* satisfies the "invariance of effect under variation of realization" condition, even though it is seemingly irrelevant to my saying "Yes, there are."

The point is again quite general. The fact that the possible realizers of a given property are a subset of those of a property that is causally relevant to the occurrence of a certain effect does not entail that the former property is causally relevant to that effect as

[26] See Walter (2005, pp.43-4) for another example illustrating the same point.

well. *Pace* Jackson and Pettit's (1990b) proposal, "invariance of effect under variation of realization" is therefore insufficient for causal relevance, as there are many properties that satisfy this requirement but which are nevertheless irrelevant to the production of the effect concerned. Our conclusion must be that Jackson and Pettit's attempts to bestow causal relevance upon events that lack causal efficacy fail by leading to an overproliferation of causal relevance.

In addition to the preceding criticisms, there also are two further problems for Jackson and Pettit's proposal. The first concerns the possibility of "causal drainage," raised by Ned Block (2003). This is in fact a problem for any generalized version of the Master Argument for epiphenomenalism, and is therefore something I'll revisit in Chapter 7, when discussing whether or not the Exclusion Problem indeed generalizes in the way that Jackson and Pettit seem to think it does. For the present, however, it is enough to recall the conclusion of the generalized Master Argument, which is that all causation occurs "at the most fundamental micro-physical level" (Jackson and Pettit, 1990b, p.209). The problem of causal drainage can then be raised in the form of a question: What if there is no such level? I.e., what if matter should turn out to be infinitely divisible, and the properties exemplified by physical entities of any given size should always turn out to supervene on some more basic set of properties exemplified by even smaller bits of matter? If this happens to be the way the world is, then following Jackson and Pettit's reasoning (according to which only events at the most basic micro-physical level possess causal efficacy, and the only events that are causally relevant to the occurrence of any effect are those that are either themselves efficacious in its production or else somehow "program for" the presence of some other event that is), we would be

forced to conclude that there is in fact no causal efficacy or causal relevance to be found at any level. As Kim (1998, p.81) puts it, "causal powers would drain away into a bottomless pit and there wouldn't be any causation *anywhere*."

Insofar, then, as the absence of any absolute bedrock in nature remains a real possibility, the parenthetical qualification at the end of Jackson and Pettit's (1990b, p.209) assertion that "we do not need to believe in any fundamental efficacies over and above those between properties at the micro-level in order to explain the regularities, actual and counterfactual, all the way up, because supervenience tells us that they are fixed by how things are at the bottom (*if* there is a bottom)" is far from insignificant, for if there should turn out *not* to be any bottom, then, according to Jackson and Pettit, we'd have no reason to believe in causal efficacy at all. But surely, as Block (2003, p.139) puts it, while "[i]t is an open question whether there is or is not a bottom level,…it is not an open question whether there is any causation." In short, we should be suspicious of any argument that entails that the existence of causation is something that has yet to be determined.

It might be thought that even in the face of causal drainage, Jackson and Pettit could still secure causal relevance for some properties by relativizing their program-process distinction, "with an arbitrary level of explanation being designated as involving causal process[es] and…higher levels being cast as programming," so that explanations given in terms of any one level may count as process explanations relative to "higher," supervening levels, and program explanations relative to "lower," subvenient levels

(Jackson and Pettit, 1990a, p.16fn).[27] This modification would, however, require significantly weakening the notion of process explanation to allow such explanations to be given in terms of events that are causally inert. And once this is permitted, it becomes difficult to see how the fact that a given event e_1 "programs for" the presence of some event e_2 that figures into a suitable process explanation of another event e_3 suffices to show that e_1 is causally relevant to e_3 even though it was inefficacious in producing it, for if e_2 can be causally inert as well (which, if causal efficacy drains away, it most certainly will be), then it would surely be an abuse of language to describe whatever relevance e_1 might be said to have to e_3 by virtue of programming for e_2 as causal in nature.

Jackson and Pettit (1990a, p.116) make note of this point, but do not seem to appreciate the problem it raises for their proposal, remarking simply that "if there is an infinite progression of levels downward and therefore no efficacious properties…then the program story will have a different significance, bearing on relations between equally non-efficacious levels." Again, though, even if this story should continue to have some significance in, say, helping us pick up on various non-causal correlations between certain supervenient properties, the properties upon which they supervene, and types of events that regularly attend their instances, the point is that (unless one simply reduces causation to constant conjunction) one could no longer describe the program story as having the *causal* significance that Jackson and Pettit need it to have if it is to provide us

[27] In addition to the objections to this proposed response to the possibility of causal drainage raised below, it is also worth noting that if "an arbitrary level of explanation" could be "designated as involving causal process[es]," then the dualist's problems with mental causation would be effectively (though trivially) solved, since one could then arbitrarily select the level of psychological explanation as the designated level where "causal" processes occur without having to identify mental properties and events with physical properties and events.

with a way of safeguarding the causal relevance of mental and macro-physical events. In other words, if Jackson and Pettit's aim is to find some way to retain potential causal relevance for all those properties, which, falling outside of the domain of fundamental physics, must, to their minds, lack causal efficacy, then given that their strategy for achieving this aim is to contend that inefficacious events can still be causally relevant to a given effect by programming for some other event that *is* efficacious in bringing that effect about, once causal efficacy goes down the drain, it's difficult to see how Jackson and Pettit's proposal can have any hope of success. For by their own account, inefficacious events can be causally relevant only if at least *some* events *are* causally efficacious.

As a final objection to Jackson and Pettit's proposal, one might wonder whether mental events can do the sort of explanatory work that Jackson and Pettit assign to them if they are in fact epiphenomenal as Jackson and Pettit take them to be. For following Davidson (1963), one might reasonably think that in order for a mental event to be capable of providing an informative explanation of, say, some bit of molar behavior, it must actually play some causal role in producing it. Support for this view might be drawn from Davidson's (1963, p.691) observation that "a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it. Central to the relation between a reason and an action it explains is the idea that the agent performed the action *because* he had the reason."

Davidson's point is that the only way to determine which among a person's mental states actually explains their actions when more than one of these seems capable of doing so is to identify which of them actually *caused* the person to act the way they

did. Since Jackson and Pettit accept psychological explanations as legitimate, but hold that the mind is nonetheless epiphenomenal, they must give some alternative account as to how we are to distinguish those mental states of a person that explain their behavior from those that perhaps *could* have accounted for it, but do not. Their suggestion is that while a person may have any number of beliefs, desires, or other mental states that qualify as potential explainers of their behavior on a given occasion, out of these candidate *explanantia*, the ones that actually explain why they behaved the way they did will be those that either (according to their 1990a paper) ensure the presence of some efficacious event necessary for the production of the behavior, or else (according to their 1990b paper) would ensure the occurrence of the same behavior throughout variation in their manner of realization.

The problem, however, is that this will likely still leave us with too many *explanantia*, for as argued above, there will typically be many different mental events that meet these criteria, most of which will not seem well characterized as reasons that give us the correct explanation of the person's behavior. Hence one is again left with the task of finding out what, among those psychological states that meet Jackson and Pettit's conditions, distinguishes the one(s) that actually account(s) for the person's behavior from those that don't, and here it is difficult to find any distinguishing feature to mark out the actual from the merely potential *explanantia* aside from the one that Davidson suggests: viz., that the mental state of the person that actually explains their behavior is the one that causes it.[28]

[28] Note that nothing in this argument requires that we accept Davidson's coarse-grained theory of events or his view that causation requires subsumption under strict laws.

In sum, while Jackson and Pettit partly accommodate Davidson's insistence on a causal interpretation of psychological explanations by conceding that mental events must at least have some kind of causal *relevance* to physical effects if the psychological explanations they figure into are to do the explanatory work we take them to, by excluding mental events from the actual causation of behavior and reserving this role solely for the micro-physical events that mental events "program for," they leave us with a surplus of mental *explanantia,* all of which program for some micro-physical event that was efficacious in producing some bit of behavior, but only some of which seem to actually account for it. The only plausible way to determine which of these putative explainers provides the correct account of the behavior appears to be to reinstate a direct causal relation between the mental event(s) that actually explain the behavior and the behavior itself.

None of this is to say that causal explanations are the *only* legitimate kinds of explanations, or that causation is the *only* relation that can provide the requisite explanatory link between an *explanans* and *explanandum*. Other forms of dependence can of course serve an explanatory function as well. Consider, e.g., the following explanations: "The table must be 3' tall, because the legs are 2'10" long, and the top is 2" thick"; "He's signaling for help, because the way he's waving those flags means S-O-S in semaphore"; "These can't be the same thing, because they have different properties"; "She's now a widow, because her husband just died."[29] While all of these are perfectly good explanations, none of them are causal. In this they differ, however, from

[29] For more on the non-causal forms of dependence or "generation" that underwrite these sorts of explanations, see Kim (1974) and Goldman (1970, ch.2).

psychological explanations of behavior, which as we have just seen, must be causal if they are to explain what they purport to (and, if our criticisms of Jackson and Pettit's proposal hold good, causal not merely in the sense that their *explanantia* "program for" some non-mental cause of the *explananda*, but in the more direct sense that their *explanantia* themselves cause the behavior they explain).

3. Conclusion

There is thus no quick and easy solution to the Exclusion Problem to be obtained by appealing to the role that mental events play in our explanatory practices.[30] Our discussion of Baker and Burge showed that the Problem cannot be justifiably dismissed by simply claiming that any metaphysical worries about, or materialist constraints on the mind's causal efficacy are outweighed by the apparent success of psychological explanations. For (a) the success of our explanatory practices depends in part on whether they can be brought into accord our metaphysical commitments regarding the basic kinds of entities that populate the world and the manner in which these are related to one another, and (b) the metaphysical assumptions that give rise to the Exclusion Problem (viz. (2) and (3)) have a good claim to be numbered among these commitments, as they enjoy a fair amount of empirical support and provide the basis for an explanatory program that has thus far met with tremendous success. The task of determining whether

[30] See also Kim's (1998, p.59) critical remarks on such "'free lunch' solutions" as he calls them.

our standing models of psychological explanation can be made consistent with (2) and (3) (and if so, how) hence cannot be so easily shirked.

On the other hand, we ought not to let Jackson and Pettit's assurance that events with no causal efficacy can still have causal relevance lead us into thinking that we can resolve the Exclusion Problem by accepting epiphenomenalism without thereby having to sacrifice the legitimacy of psychological explanations. For as argued above, once we give up the causal efficacy of mental events, we are no longer entitled to view such events as providing us with informative explanations of behavior. This, I take it, is a result we should seek to avoid, for while the apparent success of psychological explanations doesn't carry so much weight as to enable us to dismiss any potential metaphysical constraints on or challenges to their validity, we ought not to accept that this success is merely illusory without first seeing whether any more redemptive account can be given.

The idea that we should seek a metaphysical account of mental causation that can validate our standing model of psychological explanation is part of a more general, and I think quite healthy presumption in favor of realist interpretations of explanations that work (and metaphysical theories that support such interpretations). Such a view accords with what seems to be the natural interpretation of such explanations, as well as with the attitudes and practices of actual scientists, and may also be supported by an inference to the best, or at any rate the simplest explanation for why a given explanatory paradigm has proven successful (viz., that it works well because it accurately represents the way the world actually is). To adopt this stance is, however, to make one's explanatory practices and metaphysical commitments mutually dependent upon one another, for while metaphysics has a certain priority over explanatory practice in that our explanations have

a claim to validity only insofar as they cohere with the metaphysical facts, in making realism our default attitude towards successful explanations, we propose, in effect, that our metaphysical commitments themselves ought to be guided in part by an attempt to substantiate those explanations that work.[31]

In declaring a preference for metaphysical theories that support realist interpretations of successful explanations, we thus set ourselves the ongoing task of adjusting our explanatory practices and metaphysical commitments so as to bring the two into harmony with one another. Viewed from this perspective, the Exclusion Problem can be seen to stem from the fact that the metaphysical commitments of the standing explanatory paradigms in psychology and the physical sciences seem to be at odds with one another. As these explanatory paradigms appear to be equally successful, it is consequently unclear which set of commitments and/or explanatory practices should be retained, and which revised or eliminated.

Returning, then, to the views discussed above, our conclusion must be that *contra* Baker and Burge, the legitimacy of psychological explanations depends upon their being made consistent with a respectable metaphysics (i.e., one that is scientifically informed and at least consistent with explanatory success not only in psychology, but in all subject areas), and *contra* Jackson and Pettit, the legitimacy of psychological explanations cannot be made consistent with epiphenomenalism. In short, we can't argue directly from the success of psychological explanations to the causal efficacy of mental events, nor can we maintain the legitimacy of such explanations once we accept a set of metaphysical theses

[31] See Kim (1998, p.62).

that render mental events causally inert. The Exclusion Problem therefore can't be resolved by simply appealing to the success of psychological explanations, or by arguing that this success is compatible with epiphenomenalism.

The only way to validate the success of our current model of psychological explanation (which, again, is something I think we should at least *try* to do) is hence to ground it in a coherent metaphysical account of mental causation that rejects epiphenomenalism and is either consistent with (2) and (3), or else supported by reasons for thinking that (2) and (3) are false. It may seem as though the most straightforward way of doing this would be to deny (4*), Mind-Body Dualism. The next four chapters will, however, be dedicated to making the case that there are good reasons not to take this route. If correct, this means that a satisfactory solution to the Exclusion Problem must either reject (2) and/or (3), or deny that (1) ,(2), (3), and (4*) are in fact incompatible.

CHAPTER 3

THE CASE FOR DUALISM AND THE NATURE OF REALIZATION

Before pressing further in our pursuit of a solution to the Exclusion Problem, it may be useful to briefly pause and survey the ground covered thus far. Our analysis of the Exclusion Problem in Chapter 1 showed the Problem to consist in the apparent incompatibility of the following four propositions:

(1) *Causal Efficacy of the Mental:* Mental events cause physical effects.

(2) *Causal Self-Sufficiency of the Physical:* Every physical effect has a sufficient physical cause.

(3) *Absence of Systematic Overdetermination:* Causal overdetermination is rare.

(4*) *Mind-Body Dualism:* Mental events are not physical events.

After examining, in previous chapter, the solutions to the Problem proposed, respectively, by Baker and Burge and Jackson and Pettit, we found that the choice the Problem forces upon us between showing that the above four theses are not, in fact, incompatible, or else accepting that one or more of them is false cannot (as Baker and Burge suggest) be avoided by simply appealing to the success of psychological explanations, and that we also cannot (as Jackson and Pettit suggest) reject (1) without jettisoning the success and legitimacy of psychological explanations along with it. While the latter point is by itself certainly no proof that (1) is true (perhaps the success and legitimacy of psychological explanations *is* merely apparent after all), it does at least suggest that we should accept an epiphenomenalist solution to the Exclusion Problem only if no other solution can be

found that does better justice to our practice of explaining behavior in terms of mental events.

If, then, we must either reject (1), (2), (3), and/or (4*) or else argue that the incompatibility of these four propositions is somehow illusory, and rejecting (1) has the untoward consequence of illegitimating a deeply entrenched and, to all appearances, highly successful explanatory scheme, our best option may appear to be to reject (4*). This will of course seem obvious to those with physicalist leanings, who are liable to see the falsity of dualism as a conclusion already well supported, if not conclusively established, on grounds independent of the Exclusion Problem (which, given that it simply doesn't arise for their position, will itself likely strike physicalists as a mere pseudo-problem generated by a misguided theory of mind). Considering the numerous advantages that physicalism already boasts over dualism (e.g., it is ontologically simpler, derives significant inductive support from the past and continued success of physicalist reductions and explanations in the natural sciences, and exudes an air of more rigorous naturalism), the simple and straightforward solution it offers to the dualist's problems with mental causation might be easily interpreted as just another indication that (4*) is false. So, with all that physicalism has to offer, why bother to look any further for a solution to the Exclusion Problem? Why not simply reject (4*), and put our worries about mental causation to rest?

Given the influence that physicalism exercises over contemporary thinking about the mind, this is in fact the response to the Exclusion Problem that I think people working in the Anglo-American philosophical tradition are currently most likely to take. My purpose in this and the following three chapters, however, will be to offer some reasons

to hold off on rejecting (4*) at least until we have determined whether or not there is any viable solution to the Exclusion Problem that doesn't require us to do so. While hesitation on this point may seem to the convinced physicalist like nothing more than hard-headed refusal to admit something whose truth is, if not manifest, at least as well confirmed as any other theory that any reasonable person would accept, I think that in light of the arguments that can offered in its support, tentative acceptance of (4*) is at least as justified as its denial.

The reasons I'll be offering in favor of (4*) are neither exhaustive nor conclusive. My concern will not be to answer each of the many objections that a dualist theory of mind is apt to encounter, nor to provide a comprehensive, knock-down refutation of physicalism, but instead to trace one line of reasoning that, at least to my mind, gives fairly substantial grounds for thinking that a large and central class of mental events, if not all of them, are non-physical. Since my aim is simply to provide some motivation for seeking a solution to the Exclusion Problem that doesn't involve accepting either physicalism or epiphenomenalism, this should, however, be all that my purpose requires.

1. The case for dualism

The argument for dualism I'll be presenting has two components. The first is the thesis, to be defended in the following two chapters, that mental properties are multiply realizable, meaning that for any mental property, there is more than one type of physical state that is sufficient for its instantiation (and therefore no single type of physical state that is necessary for its instantiation), so that it is possible for two things to share the

same mental property *M* even though one has *M* by virtue being in a certain physical state P_1, while the other has *M* by virtue of being in a different physical state P_2. The second component consists of a series of arguments, presented in Chapter 6, which purport to show that (a) at least some mental properties (viz. phenomenal properties) are not entirely functional, and that exemplifying these properties therefore cannot be equivalent to meeting some purely functional description that a physical state might be found to satisfy, and (b) for those mental properties (viz. intentional properties) that are most likely to admit of an exhaustive functional analysis, it seems plausible that the functions they perform are not merely causal or biological in nature, and that exemplifying these properties therefore cannot be equivalent to being in some state that occupies a certain causal role, or that exhibits traits that have been selected for at some point in the phylogenetic history of the species to which one belongs.

These two components of the argument I'll be presenting can be seen as directed, respectively, against type and token varieties of physicalism. The first component raises a difficulty for any form of type-physicalism, whereas the second is directed against token- or "non-reductive" physicalist theories of mind. Since any (non-eliminativist) physicalist theory of mind must be of one of these two sorts, the two components of the argument together present a challenge to any form of mental realism that denies (4*). The challenge can be summarized as follows: If mental events exist, and physicalism is true, then either every type of mental event is identical with a certain type of physical event, or else every token instance of any type of mental event must be identical with a physical event of some type or other, even though different instances of the same type of mental event might be identical with physical events falling under different physical types. The thesis

that mental properties are multiply realizable rules out the first option, for if that thesis correct, then mental and physical event types have different modal properties (for if a type of mental event *M* is multiply realizable by physical events of types P_1 and P_2, then a token of *M* necessarily occurs whenever there is a tokening of P_1 or P_2, but this is not true of P_1 or P_2, since an instance of P_1 could, e.g., occur at times when there is no occurrence of any event of type P_2). Since identicals must be indiscernible, this entails that types of mental events are not identical with types of physical events.[32] Thus, if the multiple realization thesis is true, the physicalist's only other option is to assert that mental and physical events are token-identical (or that the mind doesn't exist).

With type-identity off the table, however, it seems that the only reason one could offer for identifying a particular mental event *m* with any one physical event as opposed to another would be to suggest that to be a mental event of *m*'s type is just to meet a certain "topic-neutral," functional description that the one physical event satisfies, while the other doesn't. Yet if the arguments to be offered in Chapter 6 are sound, then mental properties are incapable of being defined in such terms, for they either escape any exhaustive functional analysis, or else the functions that define them are functions that no entity can possess solely by virtue of bearing some assortment of physical properties.

With the two most direct routes to a physicalist reduction of mind thus blocked, there are, as far as I can tell, only three other arguments that the physicalist might offer

[32] Since the classification of events that is of interest to us is one wherein events are typed according to their "constitutive property" (i.e. the property of which they are the exemplification), talk of types of mental or physical events is for our purposes equivalent to talk of mental and physical properties. More specifically, a type of mental or physical event will for us consist of all those events that are instances of the same mental or physical property. The above statement is hence equivalent to the claim that mental properties are distinct from physical properties.

for identifying mental with physical events. First, one might suggest that even if mental properties are multiply realizable, identifications of token mental and physical events can still be made on purely empirical grounds, by directly manipulating the physical states of a minded being capable of making introspective reports to see what mental events it reports as occurring (or failing to occur) when one causes certain physical events to occur (or fail to occur) in its body and then simply identifying those mental events with whatever physical events they are found to depend upon. However, in the absence of any analysis of mental properties into a suitable non-mental idiom, this strategy will end up begging the question against the dualist, who is entitled to claim that all that such discoveries reveal are certain (possibly law-like) correlations between mental and physical events, whose distinctness from one another is evidenced by our inability to define either in terms of the other. At this point, the debate will likely come to a stalemate over the applicability of Occam's Razor, with the physicalist arguing that the dualist's interpretation of these findings multiplies entities beyond necessity, while the dualist responds that the postulation of non-physical, mental events is necessary to account for those features of the mind that cannot be completely analyzed in non-mental terms, and that by refusing to countenance such entities, the physicalist is allowing their fondness for ontological simplicity lead them into ignoring relevant data.

Second, the physicalist could argue for their proposed identification of mental and physical events on more general grounds, by appealing directly to the Exclusion Argument or, say, Davidson's (1970) argument for Anomalous Monism. Such general arguments for physicalism are, however, apt to lose much of their plausibility if we aren't also given some method for determining *which* physical event any given mental event is

identical with, and in the absence of any type-type correlations between mental and physical events or a suitable analysis of mental properties in non-mental terms, such a method will not be easy to come by. It is, moreover, one of the primary aims of this dissertation to show that even setting this issue aside, there are good reasons to view the Exclusion Argument (which is, by most accounts, the strongest of the more general arguments for physicalism) as inconclusive at best. Grounds for rejecting Davidson's (1970) argument for physicalism will be given in Chapter 6.

Lastly, in place of the classical deductive model developed by Ernest Nagel (1961), one might offer some alternative model of intertheoretic reduction that could allow for the reduction of theories dealing in multiply realizable properties to theories that are concerned with the realizers of such properties, thereby making it possible for psychology to be reduced to neuroscience (and, perhaps ultimately, to basic physics) even if mental properties should turn out to be multiply realizable. This sort of response to the multiple realization thesis is advocated by John Bickle (1998) as an alternative to the standard functionalist reply, which is to assert the autonomy of those sciences (e.g. psychology) whose properties are multiply realizable while identifying the instances of such properties with physical events. In the Chapter 5, however, it will be argued that even after adopting Bickle's "structuralist" model of intertheoretic reduction, the multiple realizability of mental properties continues to pose an obstacle to the reduction of psychology to neuroscience (or any other physical science) and the type-identification of mental and physical events.

The following three chapters will be devoted to substantiating the argument just outlined. If they are successful, then the difficulties facing any attempt to reduce mind to

matter would seem significant enough to oblige us to either accept (4*), or else deny that mental events exist. So far as there are good reasons to resist the elimination of mental states from our ontology, our only remaining option will then be to accept some form of mind-body dualism, and thus to seek a solution to the Exclusion Problem that doesn't rely on the rejection of (4*).

2. Other arguments for dualism

Such, then, is the case for (4*) that I intend to present. While there are other arguments for mind-body dualism, the one sketched in the previous section is I think both the most persuasive, and the one least likely to be accused of begging any questions against the physicalist. In contrast, two of the most well-known arguments for dualism, viz. Saul Kripke's (1980) Modal Argument and the Conceivability Arguments offered by Descartes and David Chalmers (1996), rest on two crucial assumptions that physicalists might justifiably deny: (1) that "zombies" (i.e., beings that are physically and functionally indiscernible from normal humans, yet lack consciousness) and/or disembodied minds are conceivable, or (in the case of a generalized version of Kripke's argument) that given any token physical event and any token mental event[33], it is possible to conceive of either one occurring without the other, and (2) that in this case at least, conceivability entails possibility. The physicalist is entitled to reject the first of these two assumptions, and the second might reasonably be viewed with suspicion regardless of

[33] Some may wish to restrict this to phenomenal events.

one's theory of mind. It is therefore no surprise that a substantial literature has accumulated raising a number of challenges to these assumptions and the arguments for dualism they support.[34] Focusing on the alternative argument for dualism sketched above thus has the added advantage of enabling us to bypass such objections and steer clear of the contentious issues concerning varieties of modality and the relation between conceivability and possibility that the Modal and Conceivability Arguments inevitably raise.

One might wonder, though, whether the second component of the argument for dualism outlined above is really all that different from the Modal and Conceivability Arguments whose questionable assumptions I've suggested one would do best to avoid in arguing for (4*). In particular, one might wonder what difference there is between saying that mental properties cannot be exhaustively analyzed in non-mental terms, and saying that it is conceivable and/or possible for something to share all the same non-mental properties as a normal human, but lack certain mental properties, or for any two mental and physical events to occur independently of one another. The difference is that while the claim that mental properties cannot be analyzed in non-mental terms might indeed be used to *explain* the conceivability intuitions that drive the Modal and Conceivability Arguments and/or the possibility of the independent occurrence of mental and physical events that those arguments purport to establish, the claim itself says nothing about the

[34] For a brief survey, see Yablo (1993), Dennett (1995), Hill (1997), Balog (1999), Block and Stalnaker (1999), Braddon-Mitchell (2003), and Frankish (2007). A popular strategy among physicalists (endorsed, e.g., by Loar (1997), Hill and McLaughlin (1999), Tye (1999; 2003), Block (2007), and Papineau (2007)) is to argue that the conceivability intuitions that drive the Conceivability and Modal Arguments are due merely to certain differences between our phenomenal and physical *concepts*, and are therefore perfectly consistent with the token-identity of phenomenal and physical *events*. This strategy will be discussed further in Chapter 6.

modal relations between mental and physical events or the conceptual relations between our mental and physical concepts, and therefore does not, strictly speaking, *entail* either of these conclusions. One can hence make this claim while remaining agnostic about (or even *rejecting*) the two assumptions on which the Modal and Conceivability Arguments depend. For there is no logical inconsistency in saying that mental events must be distinct from physical events in part because mental properties are unanalyzable in non-mental terms, while at the same time saying that our current mental and physical concepts are nevertheless so tightly linked that zombies and their ilk are inconceivable, or that there are certain metaphysically necessary laws governing the relations between mental and physical events that render zombies and their ilk impossible, whether they are conceivable or not.[35]

Thus, while I do think dualists would be wise to point out that the unanalyzability of mental properties in non-mental terms offers a straightforward explanation of the differences between our mental and physical concepts that could account for the apparent conceivability of zombies and the like, (thereby giving the dualist the advantage over the physicalist, who will likely find these differences much more difficult to explain in purely physical terms), there is no need for dualists to commit themselves to the stronger claim that such things are indeed completely conceivable and that their conceivability is moreover sufficient proof of their metaphysical possibility. Indeed, as will be seen in Chapter 7, one of the more attractive solutions to the Exclusion Problem available to

[35] This point will be revisited in Chapter 6.

dualists actually requires them to *deny* the metaphysical possibility of zombies and occurrences of physical events without the mental events that actually depend on them.

The argument for (4*) that I'll be advancing therefore does not depend on the controversial assumptions of the Modal and Conceivability Arguments, although it can be used to support these assumptions, if one so wishes. Instead of inferring (4*) from the conceivability of zombies and the like, (4*) is, on our account, to be inferred from the multiple realizability of mental properties along with the fact that the physical features of any given physical event are insufficient to classify it as falling under any mental type. The only role that conceivability intuitions might play in the argument is to provide some additional, abductive support for the claim that mental properties are unanalyzable in non-mental terms.

3. The origins of the multiple realization thesis

Having now outlined the basic structure of the argument to be presented over the course of the next three chapters and distinguished it from some other arguments for dualism, all that remains is to fill it in. The remainder of the present chapter will prepare the way for a defense of the first component of the argument (the thesis that mental properties are multiply realizable) by examining the origins of the thesis and offering a more precise formulation of the realization relation that the thesis involves.

The introduction of the idea that mental properties are multiply realizable into the mainstream philosophical discussion is widely credited to Hilary Putnam's 1967 paper "Psychological Predicates." Contrary to what one might expect, however, the multiple

realization thesis in fact plays a fairly minor role in that paper, the primary aim of which is instead to advance and defend the hypothesis that psychological states are not brain states or behavioral dispositions, but instead "functional state[s]" (Putnam, 1967, p.42). In the paper which is considered to be its point of origin, one will thus find no explicit formulation of, or argument for the multiple realization thesis. Putnam introduces the idea merely by way of the suggestion that functionalism should be favored over type-physicalism in part because the cross-species psychoneural identity statements to which type-physicalism is committed seem implausible.

> Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus…, but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this (Putnam, 1967, pp.44-5)

As Putnam makes clear, this "overwhelmingly probable" outcome raises no difficulties for the functionalist hypothesis he advances, because identifying mental with functional properties allows us to remain agnostic about the types of physical events that realize them on any given occasion.

> It should be noted that knowing the Total State of a system relative to a [functional] Description involves knowing a good deal about how the system is likely to 'behave,' given various combinations of sensory inputs, but does *not* involve knowing the physical realization of the [state that the system is in] as, e.g., physical-chemical states of the brain. (Putnam, 1967, p.42)

At this point, the multiple realization thesis thus appears to have been viewed by Putnam merely as an intuitively credible hypothesis whose main interest lay in the support it lent to the functionalist theory of mind he then favored.

Given, then, the purpose for which the multiple realization thesis was originally introduced (viz., to lend credibility to functionalism), the question naturally arises whether commitment to the thesis necessarily goes hand in hand with a commitment to functionalism (the view that mental properties are functional properties, so that that to possess a certain mental property just is to be in a certain type of functional state). The short answer, I think, is no: functionalism doesn't follow from the multiple realization thesis, nor does functionalism entail that the mental properties in fact have multiple physical realizations. The most that can be said is that functionalism is compatible with the multiple realization thesis, whereas one of its main competitors, the type-physicalism traditionally associated with U.T. Place (1956), J.J.C. Smart (1959), and Herbert Feigl (1958), is not. In order to see why this is so, however, it is necessary to take a closer look at the realization relation itself, so that we can better understand what it means to say that a certain property is multiply realizable.

4. The realization relation

While realization is among the relations most frequently used to describe the nature of the relation between mental and physical events, it has failed to receive the same degree of attention as the two other relations that are most often selected for service in that role; viz. identity and supervenience.[36] Consequently, unlike the latter two

[36] A fourth option, endorsed by Yablo (1992a; 1992b), is to say that mental and physical properties are related as determinables to determinates. Against this proposal, Funkhouser (2006, pp.550-2, 563-5) argues that while, e.g., pains can differ in their "pain-ness" by having different degrees of intensity and/or different apparent locations, and beliefs can differ in their "belief-ness" by having different contents and

relations, there is as yet no common consensus as to the precise character of the realization relation, or how it is best defined. A number of proposals have, however, been made. Prominent among these are the definitions offered by Sydney Shoemaker (2001), Carl Gillett (2002; 2003), Andrew Melnyk (2006), and Thomas Polger (2007). In the remainder of the present chapter, I'll explain why I think none of these definitions are adequate, and propose an alternative theory of realization capable of overcoming the failings of each, which involves treating realization as a synchronic dependence relation that can be exemplified in four different ways. Before examining the definitions proposed by Shoemaker, Gillett, Melnyk, and Polger, though, I'll begin by laying down a few minimal conditions of adequacy for any definition of realization, so that we might more easily evaluate their proposals by considering them in light of these constraints.

4.i. Constraints on an adequate definition of realization

The guiding aim of any attempt to provide a non-stipulative definition of a given term should, I take it, be to specify an extension that is at least roughly equivalent to that which the *definiendum* is commonly, if only imprecisely understood to have. This means

---

degrees of confidence, on the assumption that mental properties are multiply realizable, two instances of a given mental property (e.g., *being in pain* or *believing that p*) can be psychologically indiscernible while differing with regard to their physical realizations. This entails, however, that "content, attitude, phenomenology and similar psychological-level features are the only determination dimensions for mental properties." For if the mental properties *were* determinable with respect to their physical realizations, then two instances of belief that shared the exact same content and degree of confidence but had different physical realizations would be different (and not merely numerically distinct) *qua* states of belief. But this does not seem to be the case. Physical properties therefore cannot be determinates of mental properties because mental and physical properties are determinable along different determination dimensions, and in order for two things to be related as determinate to determinable (as, e.g., *scarlet* is a determinate of *red*), they must share the same determination dimensions (as, e.g., *scarlet* and *red* are both determinable only with respect to hue, brightness, and saturation). (See, however, Wilson (2009).)

that an adequate definition of realization ought first and foremost to specify a relation that includes in its extension all the ordered *n*-tuples that are viewed as paradigmatic cases of realization, while excluding those that few if any would be willing to apply the term to. We want a definition that will permit us to say (or at least hypothesize), e.g., that a statue is realized in the matter of which it is composed, that human mental states are realized (at least in part[37]) by the brain states of the individual that has them, and that an algorithm can be realized in the electrical circuitry of a computer, while prohibiting us from saying, e.g., that the Morning Star is realized by the Evening Star, that a baseball bat striking a vase realizes the vase's consequent shattering, or that Oedipus realizes (in the sense at issue) that he's an incestuous patricide. To hit upon a definition of realization capable of satisfying this central constraint, we must attend to what the various uncontroversial cases of realization have in common, and what differentiates them from ordered *n*-tuples that clearly do not fall within the extension of the relation, so that we can see what distinguishing features realization appears to exhibit and what conditions must consequently be included in our definition so as to enable it to carve out the right set of entities.

Applying this method, we can note, first, that realization appears to be a binary, transitive relation that is also unconnected and synchronic. While the relation also seems, in most cases, to be both asymmetric and irreflexive, these conditions should perhaps be left out of the any definition of the relation in order to allow for the possibility that some

[37] Externalists may also wish to include states of the individual's surrounding environment in the total state that realizes those of their mental states that have any kind of representational content.

things (e.g. fundamental entities) might realize themselves.[38] As for the *relata* of realization, it seems that while the things that can be realized are an ontologically heterogeneous lot, including abstract entities (e.g. Turing Machines), concrete objects (e.g. statues), concrete instances or tokens of some property, relation, or type (e.g. events), and collections of such instances or tokens (e.g. states or processes), realizers are typically concrete. The sole potential exception to this rule would seem to be cases wherein the realizing entities are numbers or other mathematical entities, as one might, e.g., think of the function $f(x)=x^2$ as being realized by the set of ordered pairs {(1, 1), (2, 4), (3, 9),…}. Such cases aside, though, realizers appear to be always concrete. Talk of properties being realized or realizable by other properties, while permissible, must therefore be understood to mean that the realized properties or their instances are realized or realizable by concrete instances of certain properties. While it seems that in all *actual* cases that are treated as instances of realization, realizers are also always *physical*, it is unclear whether realizers should be required by definition to meet this further constraint, as it is sometimes said that in other possible worlds, certain entities might be realized in ectoplasm, or some other concrete, non-physical stuff. To allow for such cases, it seems best to treat the fact that all *actual* realizers are mathematical or physical as an important, but contingent truth.

Another essential feature of realization that any suitable definition of the relation must accommodate is that realization is a form of dependence. As such, it differs from

[38] Note that according to Shoemaker's definition (cited below), realization is actually reflexive, since everything bestows a subset of the causal powers that it itself bestows. The same is also true of Melnyk's definition, for reasons noted in footnote 47.

the supervenience relation, which, as many have noted, merely describes a necessary covariance of properties.[39] The kind of dependence that realization involves does seem closely related to supervenience, in that if a certain kind of properties or property instances are said to be realized by the instances of properties of another kind, one is entitled to infer that properties of the former kind also supervene on those of the latter. In contrast to supervenience, however, realization involves the further requirement that the covariance of realized properties with the properties that realize them must be due to the fact that instances of the realized property occur only *by virtue of* the occurrence of instances of the realizer property (perhaps in conjunction with certain laws), and that the instantiation of a realizer property is always both sufficient and in some sense *responsible for* the simultaneous[40] instantiation of the property that it realizes (again, perhaps only in conjunction with certain laws).

Lastly, there is nothing intrinsic to the realization relation itself to prevent the possibility that different tokens of the same type of entity can be realized by concrete entities of different physical types. Any obstacle to the multiple realizability of a certain realizable entity must therefore come from the nature of that entity itself, or from certain laws governing its behavior and/or that of its potential realizers. In short, if the only thing

---

[39] Supervenience is further distinguished from realization by the fact that its only *relata* are kinds of properties. Unlike supervenience, realization also appears to be both non-monotonic, in that it is not the case that if *x* realizes *y*, then (*x* & *z*) also realizes *y*, (whereas if *A* supervenes on *B*, then *A* also supervenes on (*B* & *C*)), and explanatory, in that if *x* realizes *y*, the fact that *x* obtains, occurs, etc. can be used to explain the fact that *y* obtains, occurs, etc. (whereas supervenience, as Kim (1993b, p.167) notes, "is not an explanatory relation…[R]ather, it is a 'surface' relation that reports a pattern of property covariation, suggesting the presence of an interesting dependency relation that might explain it").

[40] Note that if causation is diachronic, then the responsibility or dependence at work here must be non-causal.

known about *x* and *y* is that *x* realizes *y*, then there is no grounds for denying the possibility that *y* is multiply realizable.[41]

Summing up, it appears that if we want a definition of realization that captures all the paradigm cases of the relation without admitting any cases that clearly do not belong in its extension, we must seek a definition that specifies a binary, unconnected, transitive, synchronic relation between certain abstract entities, concrete objects, tokens, or property instances, or collections of such things, on the one hand, and certain concrete or abstract, mathematical entities, on the other, where this relation is such that (a) the entities belonging to the former class depend on the members of the latter in a way that entails (but is not entailed by) the supervenience of realized properties on the properties whose instances realize them or their instances, and (b) it is at least logically possible for different tokens or instances of any realizable type or property to be realized by entities of different types.

Before moving on to consider how the definitions offered by Shoemaker, Gillett, Melnyk, and Polger fare in satisfying these conditions, two further points warrant mention. First, the fact that realization is a technical notion with a specific, highly specialized explanatory role makes it much more plausible to think that a precise definition of the relation can be provided in terms of necessary and sufficient conditions for its instantiation. Such an attempt would, in contrast, be much more dubious if carried

[41] Polger and Shapiro (2008, p.214) argue that it follows from this condition that property instances cannot be the only *relata* of the realization relation, for property instances, being non-repeatable, cannot be "repeatable in different ways," and hence cannot be multiply realized. Gillett (2011), however, responds by noting that if property instances can persist through time, the same property instance might at different times be realized in different ways. (See also Aizawa (2013, p.76fn13).)

out with respect to thicker or more quotidian concepts (e.g. "good" or "person") which are more likely to be infected with vagueness and ambiguity.

The second point concerns a potential objection to the very project of seeking a single definition of realization, which is that while realization is indeed a technical notion, it may nonetheless serve different explanatory roles, and hence have different meanings, within different theoretical traditions. Such a position is taken by Ronald Endicott (2012, pp.42-3)[42], who argues that the term "realization" in fact has three distinct senses, corresponding to its use in three different theoretical traditions: the *representational* tradition ("according to which realization is a *semantic or intentional relation*" between a representation and something that satisfies it), the *mathematical* tradition ("according to which realization is a *mapping or isomorphic relation*" between Turing machines or sets of numbers, e.g., and things that they can be mapped onto in one-to-one fashion), and the *metaphysical* tradition ("according to which realization is a relation of *determination or generation*" between two things, one of which produces or gives rise to the other). Insofar as these various uses of "realization" pick out different kinds of relations that prove explanatorily useful in different theoretical contexts, it is, Endicott maintains, simply misleading to speak of *the* realization relation, as there are a number of distinct kinds of relations that the term can be legitimately used to refer to.

Although the simplest response to Endicott's point would be to concede that the notion of realization for which a definition is being sought here is just one of the term's many technical senses, there are a couple reasons why any definition that satisfies the

[42] See also Gillett (2010, p.167fn5; 2013, pp.166-9).

constraints laid down above might be justifiably viewed as capturing *the* realization relation. First, while these constraints have little in common with the sense of realization that Endicott associates with the "representational" tradition, the "representational" notion of realization strikes me as redundant, for it appears to be more or less identical with the familiar and already well-established notion of one thing's "satisfying" or "being a model (in the mathematical sense) of" another. Given that this notion of realization serves no special explanatory function that would prevent its being eliminated in favor of the other notions just mentioned, it seems unwarranted to view this sense of the term as expressing a distinctive kind of *realization* relation. Second, while the above constraints on an adequate definition of realization lean rather heavily towards the so-called "metaphysical" tradition (as indicated by the requirement that realized entities depend on their realizers), I hope to show that a definition satisfying these constraints can accommodate many of the sorts of cases that Endicott would classify under the "mathematical" tradition as well. Since, then, the "representational" conception of realization does not seem to pick out a distinctive realization relation, and paradigm cases from both the "metaphysical" and "mathematical" traditions can, if I'm correct, be accounted for in terms of a single definition satisfying the constraints listed above, it seems reasonable to view these constraints as marking out the essential features of *the* realization relation. In deference to Endicott, however, the definition of realization I propose below also incorporates some of the plurality he emphasizes by describing the relation as a form of dependence that can be exemplified in four different ways.

4.ii. Evaluation of prominent proposals

Now to evaluate the definitions of realization proposed respectively by Shoemaker, Gillett, Melnyk, and Polger in light of the constraints enumerated above. These definitions are as follows:

> Shoemaker (2001, p.78): "Property X realizes property Y just in case the conditional powers bestowed by Y are a subset of the conditional powers bestowed by X (and X is not a conjunctive property having Y as a conjunct)."[43]
>
> Gillett (2002, p.322; 2003, p.594): "Property/relation instance(s) F_1-F_n realize an instance of a property *G*, in an individual *s*, *if and only if s* has powers that are individuative of an instance of *G* in virtue of the powers contributed by F_1-F_n to *s* or *s*'s constituent(s), but not vice versa."
>
> Melnyk (2006, p.129): "Token *x realizes* token *y* (or: token *y* is *realized by* token *x*) iff (i) *y* is a token of some *functional* type F (i.e., some type whose tokening just is the tokening of some or other type that meets a certain condition, C); (ii) *x* is a token of some type that in fact meets C; and (iii) the token of F whose existence is necessitated (in the strongest sense) by the holding of clause (ii) is numerically identical with *y*."
>
> Polger (2007, p.251): "Property/state instance *P* realizes property/state instance *G* iff *P* has the function $F_G(x)$."

As is the case with most general definitions of realization currently on offer, these four definitions can be divided into two camps, one of which locates the nature of realization in one thing's deriving its distinctive causal powers from those of another, and the other of which conceives the relation as instead consisting in the performance, by one thing, of the distinctive function of another. The definitions proposed by Shoemaker and Gillett

[43] Shoemaker's definition will be taken as representative of other, similar "causal-subset" definitions of realization endorsed by Clapp (2001, p.129) and Wilson (2011). Wilson (1999) seems to have been the first to propose this idea in print. I focus on Shoemaker's presentation of it merely because his formulation is more clearly intended as a definition of realization. Wilson (1999) presents the idea instead as a constraint on adequate formulations of physicalism, arguing that any such formulation must require the causal powers of all supervenient properties to be subsets of the causal powers of physical properties or disjunctions of such properties.

can be seen to belong to the former camp, while those proposed by Melnyk and Polger belong to the latter.

One immediate problem for the causal definitions offered by Shoemaker and Gillett is that not all of the sorts of things that are commonly spoken of as realizable have or bestow a distinctive set of causal powers. As Melnyk (2006, p.146) notes, certain putatively realizable properties, e.g. "*having such-and-such a biological function, being a member of a species*," or (to use an example of Polger's (2007, p.241)) *being a genuine one dollar bill*, "seem to be such that possession of them requires having not just causal *powers* but an *actual causal history* of some particular kind." Gillett and Shoemaker's definitions are unable to treat such properties or their instances (or any entities, e.g., biological organs, to which such properties are essential) as realizable, because there are no unique sets of "powers that are individuative of" or "bestowed by" them. As Polger (2007, p.241) points out, the problem becomes especially noticeable when one considers that some of the most paradigmatic examples of realizable entities, e.g. Turing machines and algorithms, not only fail to have or bestow a distinctive set of causal powers; they do not appear to have or bestow any causal powers at all! As these sorts of things nevertheless seem like perfectly legitimate, and even central examples of realizable entities, any account of realization that limits the class of realizable entities to such things as can be distinguished from one another solely on the basis of the causal powers they have or bestow appears to be far too restrictive.

Now if one accepts Endicott's contention that there is no single realization relation, but instead a variety of distinct relations introduced to serve different

explanatory needs within different theoretical contexts, one can respond to this criticism (as Endicott (2012, pp.47-51) in fact does) by arguing that the sense of realization captured by definitions that require realized entities to be causally individuated belongs to a different theoretical tradition than the sense of realization that would also treat certain non-causally individuated entities as realizable. It is therefore, one might argue, no objection to the causal definitions offered by Shoemaker and Gillett that they fail to treat entities like biological properties and Turing machines as realizable, for one cannot condemn a definition of one sense of realization for failing to satisfy constraints or accommodate cases imported from a distinct tradition. While I am not dead-set against the pluralist conception of realization that this point rests on, and I do think that such a view may turn out to be the only feasible option should no unified definition of realization prove satisfactory, it seems to me that before one accepts any definition of realization that posits an ambiguity in the term in order to justify omitting some of its central uses, one should first see whether there is any common element running through the term's various uses that could provide the basis for a single definition that encompasses them all. The definition offered in section 4.iii. below is meant to show that such a unified definition of the realization relation can indeed be given.

Additional criticisms have been leveled by Polger (2007) and Polger and Shapiro (2008) against Gillett's "dimensioned" view of realization, which allows realization to relate property instances instantiated in different individuals, so long as these individuals are also mereologically related to one another as whole to part(s). Gillett's dimensioned view contrasts with the "flat" view of realization endorsed by Shoemaker *et al.*, which requires realized and realizing property instances to be instantiated in the same

individual. The gist of Polger and Shapiro's criticisms of the dimensioned conception of realization is that by permitting property instances instantiated in a given individual to be realized by property instances instantiated not only in the individual itself, but also in that individual's constituents, "the dimensioned view draws realization too near to material composition in general," and thereby "runs counter to" the purpose for which Putnam (1967) originally introduced the term into the mind-body debate; viz. "to provide an alternative to the view that mental states are identical to or composed of brain states" (Polger, 2007, p.248).

In response to this criticism, proponents of the dimensioned view are I think perfectly justified in arguing that while the term "realization" may have been coined for a specific purpose, the relation it picks out might nonetheless be found to obtain in cases besides those it was originally invoked to explain. We thus needn't be alarmed if certain relations (e.g. material composition) that Putnam first introduced the realization relation as an alternative to should turn out to be special cases of the latter, for there is no reason why we should take the legitimate usage of the term "realization", or the extension of the relation it refers to, to be restricted to the types of cases to which Putnam first applied it. The fact, therefore, that Gillett's dimensioned view seeks to apply the notion of realization to certain compositional relations invoked in scientific theorizing rather than simply elucidating Putnam's use of it to describe the relation between Turing machines and isomorphic physical systems does not mean that the dimensioned view of realization is thereby guilty of "semantic reformation," as Polger and Shapiro (2008, p.219) would

have it, for the relation referred to in both types of cases may in fact be the same.[44] The definition I propose below will aim to bring out this similarity by showing how in both types of cases, as in indeed in all instances of realization, the realized entities depend on their realizers at least partly by virtue of the latter's performance of a certain function.

While his dimensioned view of realization is thus unobjectionable, Gillett's definition of the relation is, like Shoemaker's, still open to the criticism that it is incapable of accommodating realized entities that are not individuated by their causal powers. This problem, which any causal definition of realization must inevitably encounter, does not, however, affect definitions like those offered by Melnyk and Polger, which instead locate the nature of realization in the performance, by realizers, of certain distinctive functions of the entities they realize. The reason for this is fairly clear: Not all functions are causal functions (Polger, 2007, p.251). Certain functions (e.g., biological, computational, or representational functions) are distinguished not (or not merely) by the causal powers they involve, but instead (or also) by the historical properties that the things that have them must exemplify, by the carrying out of certain formal operations on numbers or symbols, or by the possession of certain veridicality or satisfaction conditions. To be or have a function therefore isn't necessarily equivalent to having a certain set of causal powers or occupying a certain causal role. Indeed, some functions

[44] Gillett himself seems much more inclined to simply grant that the dimensioned view indeed articulates a different sense of realization than that developed by Putnam, and to respond to Polger and Shapiro's criticisms by instead arguing (*à la* Endicott) that realization has acquired a number of different technical senses that prove useful in a variety of different explanatory contexts, so there is consequently no reason to treat Putnam's use of realization as exemplifying the only legitimate sense of the term. While I agree with Gillett that the fact that dimensioned realization deviates from the notion of realization developed by Putnam is no reason to reject it, I think the two forms of realization are still similar enough to warrant treating them as different expressions of the same relation.

(viz. computational functions) do not involve the possession of any causal powers at all. Thus, if, as Melnyk and Polger suggest, to be realized by something is just to be or have a certain function that that something is performing, then realized entities needn't be distinguishable from one another on the basis of the causal powers they have or bestow. Unlike the proposals of Gillett and Shoemaker, the definitions of realization offered by Melnyk and Polger are hence able to treat entities like Turing machines, algorithms, and biological, economic, and semantic/representational properties that are not causally individuated as realizable, (so long, of course, as these things can be equated with, or shown to have a certain function).

As the notion of function occupies a central place in the account of realization developed in the following section, it may be worth adding a word or two on the subject here. While it is now fairly uncontroversial to suggest that there are many different kinds of functions, accepting this idea may seem to raise problems for any attempt to provide a unified definition of realization in terms of the performance, by realizers, of certain functions that are related in a certain way to the entities they realize. For if the various types of functions that can figure into instances of realization are so different from one another that they have nothing in common aside from their all being labeled as "functions,"[45] one might again wonder whether the various instances of realization in which these distinct types of functions are involved are indeed all instances of the same relation. In short, if the term "function" is ambiguous, and the notion of function plays a

[45] See, e.g., Godfrey-Smith (1993, pp.196, 206), as well as Gillett (2013, pp.179-80).

central role in the definition of realization, the question naturally arises whether the term "realization" might not, as Endicott argues, be ambiguous as well.

Against this proposal, I suggest that just as the apparent diversity across different instances of realization needn't prevent us from isolating certain features that all instances of realization have in common, and by virtue of which they all qualify as instances of a single relation, so too the diversity among different types of functions needn't prevent us from isolating certain features that all functions share, insofar as they are functions. One such feature can be located in the fact that every function seems to be individuated at least in part by the inputs and outputs it connects, and the means by which it does so. Whereas a specific causal function might thus be distinguished by its production of certain effects in response to certain causes by means of a particular kind of mechanism, a specific computational function might be distinguished by its connection of certain arguments with certain outputs by means of a particular series of calculations, or a specific representational function by the fact that it yields truth or veridicality under certain conditions by way of a particular mode of representation (e.g. pictorial or linguistic). As all functions seem to be individuated along these lines, it would appear that all functions must share at least one distinguishing feature in common; viz. that of connecting certain inputs with certain outputs by certain means.[46] Given the availability of a general notion of function that encompasses functions of all kinds, the fact that

[46] Note that according to this conception of functions, there is no reason why a function couldn't be a fundamental, first-order property "which is neither realized by, [n]or dependent upon, any other property" (Gillett (2013, p.175). As such, the conception of functions here advocated should not be confused with the conception criticized by Gillett (2013, p.170), according to which functions are second-order properties, specified by Ramsey sentences, which can only ever be instantiated "in virtue of some [other] property playing [their] individuating role."

different instances of realization can involve different kinds of functions should consequently not lead us to think that there can be no general realization relation, for while the various kinds of functions may enable us to distinguish different ways in which the realization relation can be instantiated, they still have enough in common to justify our treating the various instances of realization in which they figure as instances of a single relation.

Returning, now, to our assessment of Melnyk and Polger's definitions of realization, the first problem for definitions such as theirs, which hold that for one thing to realize another is for it to perform the latter's distinctive function, is raised by instances of "dimensioned" realization. Since, in such cases, the realized and realizing entities are instantiated in distinct individuals, the functions performed by the realizers (by which they ensure the occurrence or instantiation of the entity they realize) will typically differ from the distinctive function of the thing they realize (assuming that the realized entity in fact *has* any such distinctive function; more on this below). Thus, the hardness of a diamond, which, according to Gillett's dimensioned view, is realized in certain properties of and relations between its constituent atoms, is individuated by a certain causal function (viz. that of scratching, without being scratched by, certain types of surfaces) that none of its realizers itself has or performs. The functional roles occupied by the atomic level realizers of the diamond's hardness are individuated not by an ability to scratch surfaces, but rather by such functional features as the capacity to form certain molecular bonds having a certain kind of structure, which remain intact under certain degrees and types of external stress (See Gillett (2010, p.172)). In similar fashion, the causal roles occupied by the physical constituents of an internal combustion engine might

lead us to say that the engine is realized in its constituents (since by occupying these various roles, the constituents ensure that the system they constitute performs the distinctive causal function of an internal combustion engine), even though the functional property that individuates the engine as a whole is different from the more specialized functions performed by the engine's various constituents (e.g. the function of the combustion chamber). The fact that in such cases the functions performed by the realizers differ from the distinctive function of the thing they realize entails that *pace* Melnyk and Polger, the functional roles occupied by realizers needn't be individuative of the entities they realize.

This result leads naturally into a second objection to Melnyk and Polger's functional definitions of realization, which parallels their own principal objection to the causal definitions of Shoemaker and Gillett. Here the problem is that just as the latter run afoul of the fact that not all realized entities are causally individuated, the former face the difficulty that not all realized entities are individuated in purely functional terms either. Thus, while it seems natural to say that the *Mona Lisa* is realized in certain globs of paint on a wood panel hanging in the Louvre, there does not appear to be any unique function that only the *Mona Lisa* has and that any perfect replica of the painting would lack.[47] This needn't deter us, however, from saying that the *Mona Lisa* is realized in the matter of which it is composed, for given that realized entities needn't be individuated in terms of

[47] Unless, that is, we weaken the notion of function so far as to hold, like Melnyk, that to have a certain function is just to be a "tokening of some or other type that meets a certain condition, C," but such a move leads to the unacceptable result that *everything* realizes itself simply by virtue of satisfying the condition of being identical with itself, since, on Melnyk's view, this would seem to be a function that everything both has and performs.

the functions performed by their realizers, there seems no reason not to go one step further and suggest that they needn't be functionally individuatable at all. Instead of insisting, as Melnyk and Polger do, that realizable entities be functionally individuatable, we might instead deem it sufficient that they have or bestow *some* function that their realizers ensure is instantiated, so long as any further, non-functional properties that are necessary for their individuation are *also* instantiated in virtue of certain properties of their realizers.[48]

Taking this route would enable us to handle the preceding example as follows: While the *Mona Lisa* is realized in the matter of which it composed at least in part because it has the causal function of reflecting incident light in certain ways so as to cause certain visual experiences in its viewers under certain conditions, and this function is instantiated in virtue of the causal functional roles occupied by the painting's various material components, since the *Mona Lisa* shares this function with any perfect replica of the *Mona Lisa*, this particular function is by itself insufficient to individuate the painting; further information concerning its causal history must also be taken into account. If presented with an array of paintings consisting of the *Mona Lisa* and a number of perfect replicas, the only way to determine which physical object realizes the *Mona Lisa* is hence to say that the *Mona Lisa* is realized in that object which, *in addition to* performing the

[48] To be clear, to deny that a certain entity is functionally individuated is *not* to deny that it has *any* function or occupies *any* functional role; it is merely to say that there might be other things distinct from that entity that nonetheless have the same functions or occupy the same functional roles as it does. Put another way, to say that an entity is not functionally individuated is to say that whatever functions that entity has are either not essential to it, or else do not completely *exhaust* its essence, as there are other, non-functional features it possesses that also partially constitute its essence. The above proposal is hence that while a realized entity must have *some* function that its realizers ensure is instantiated, that function needn't wholly constitute its essence, or even be essential to it at all.

Mona Lisa's causal function, also has certain historical properties that are appropriately related to those historical properties of the painting that, together with its causal function, suffice to individuate it. We can therefore say that the *Mona Lisa* is realized in *this* physical object, as opposed to any others that perform the same causal function, due to the fact that the constituents of this object not only collectively perform the *Mona Lisa*'s causal function, but were also arranged by Da Vinci in such a way that they *would* perform that function.

These considerations suggest that in order to identify the realizers of a realized entity, one must sometimes (viz., in certain cases where the realized entity is not individuated in purely functional terms) know more about the realizers than the functional roles they occupy, even though their occupying the functional roles that they do is still *necessary* to their being the realizers of that entity, and the realized entity's having or bestowing *some* function is likewise still necessary for its being realizable at all. Consequently, the performance by one thing or a group of things of some function or functions that ensure the instantiation of some function that another thing has or bestows is necessary but not sufficient for the former to qualify as realizing the latter. The satisfaction of this condition is both necessary *and* sufficient for realization only in cases where the function whose instantiation is ensured is individuative of the thing that is realized.

4.iii. Realization defined

The previous section has led us to the conclusion that neither the causal definitions proposed by Shoemaker and Gillett, nor the functionalist definitions offered by Melnyk and Polger prove adequate as general definitions of realization. Whereas Shoemaker and Gillett's definitions fail to accommodate those realized entities that are individuated in non-causal terms, Melnyk and Polger's definitions are unable to allow for cases wherein the realized entity is individuated by a function that is distinct from the function(s) performed by its realizer(s), as well as cases wherein the realized entity is not functionally individuated at all. A fully general definition of realization will hence have to be more inclusive than those considered thus far.

In working our way towards such a definition, it will be useful start by developing a taxonomy of the various forms that realization can take. Here the distinction introduced above, between realized entities that are, and those that are not individuated in purely functional terms is significant, because it indicates how such a taxonomy might be constructed on the basis of the different ways in which realizers ensure the occurrence or instantiation of the entities they realize. Thus, in the most straightforward cases of realization, the thing realized is functionally individuated and its realizers ensure its occurrence or instantiation either by occupying the very functional role that individuates the thing[49], or else by occupying a range of subsidiary functional roles into which the individuative function of the realized entity can be decomposed (as, e.g., the individuative function of an internal combustion engine can be decomposed into a number of different sub-functions that its various parts perform). In such cases, the

[49] Such cases seem to be the only ones that the definitions of Melnyk and Polger are able to accommodate.

realizers ensure the occurrence or instantiation of the things they realize by being *logically sufficient* for them, since for any functional entity *x* with the individuative function *F*, the existence of some entity that performs *F*, or of some collection of entities performing a set of functions f_1-f_n that together constitute a functional decomposition of *F*, is logically sufficient for the instantiation or occurrence of *x*. As can be seen from the description just given, this form of realization encompasses not only cases that Endicott would likely classify as belonging to the "metaphysical" tradition (e.g. the realization of an internal combustion engine in its physical constituents), but also paradigmatic "mathematical" cases as well, such as the realization of a Turing machine in a physical system that performs a causal function isomorphic to (and hence logically sufficient for the instantiation of) the computational function that individuates it.

Where the realized entity is not fully individuated in functional terms, matters become more complicated, for in such cases, the functional roles occupied by the realizers will not be logically sufficient for the occurrence or instantiation of the thing realized. The realizers of such entities must consequently either (a) exemplify certain other, non-functional properties that somehow ensure the instantiation of the non-functional individuative properties of the thing realized, or else (b) certain of the functions that the realizers perform must be such as to ensure, *without logically entailing*, the instantiation of the additional, non-functional individuative properties of the realized entity.

In cases of the former sort, it appears that the additional, non-functional properties of the realizers will always be logically sufficient for the non-functional individuative features of the thing they realize. Such is the case, e.g., in the *Mona Lisa* example

discussed above, where the additional, historical property of the painting's material constituents, of having been arranged by Da Vinci in a certain way (perhaps with a certain intention in mind), is logically sufficient for the *Mona Lisa*'s further individuative, non-functional property of having been made by Da Vinci. The same would also appear to be the case for any realized entity possessing certain historical properties that are essential to it (e.g. its origins). Assuming, e.g., that the origins of the desk at which I'm currently sitting are essential to it, the matter in which it is realized must, it seems, not only perform certain functions that suffice to ensure the instantiation of whatever causal functions the desk has; it must also have the historical property of having been made part of the desk (or of having replaced certain parts, or replaced certain parts that replaced certain parts…that were originally used as parts of the desk) at the moment of the desk's creation.

By contrast, in cases of the second sort, while the realized entities are also, as in the cases just discussed, not individuated in purely functional terms, the additional non-functional properties that are necessary for their individuation are instantiated not by virtue of any additional non-functional properties of their realizers, but instead solely by virtue of the functions that their realizers perform. Since, however, the mere performance of a function cannot be logically sufficient for the instantiation of a non-functional property, the realizers' performance of certain functions in such cases ensures the instantiation of the non-functional individuative properties of the things they realize *not* by being logically sufficient for them, but instead by virtue of some (non-logical, perhaps non-physical) *law*, which makes it such that the performance of those functions is necessarily attended by the instantiation of the further, non-functional individuative

properties of the entities they realize.[50] Thus, whereas in the former sort of case, the realizers ensure the occurrence or instantiation of the things they realize by being logically sufficient for them, in cases of the sort now under discussion, the realizers ensure the occurrence or instantiation of the things they realize by being not logically, but instead *nomologically* sufficient for them.

An example of such a case might be found in the realization of phenomenal properties. If, as will be argued in Chapter 6, phenomenal properties such as pain or the experience of blue cannot be individuated in purely functional terms (as there is something it is like to experience pain, e.g., that cannot be completely analyzed in terms of the typical causes and effects of that experience), then the physical realizers of such properties cannot logically suffice for their instantiation simply by performing the causal/biological functions that they do. Nonetheless, it seems that the realizers' performance of such functions *is* all that's needed in order to ensure the instantiation of the phenomenal properties they realize. If we were to find, e.g., that under certain conditions, a certain type of neural activity always gives rise to a certain type of experience, it would seem reasonable to infer that any physical state performing the same causal/biological functions as the neurons engaged in that activity would also produce the same type of experience under those conditions. Since, however, the performance of those functions would still not be *logically* sufficient for the occurrence of such an

[50] The modal force of the relevant law might differ from case to case, running anywhere from mere natural or nomological necessity to full-blown metaphysical necessity. Following Fine (2012, pp.38-40), one might also distinguish between the *normative* laws or conventions by virtue of which, e.g., an insult or chess move is realized in a certain gesture, and the *natural* laws by virtue of which, e.g., pain is realized in certain neural activity.

experience (since the performance of a function cannot logically suffice for the instantiation of a property that is not functionally individuated), it remains that there must be some non-logical (perhaps non-physical) *law* correlating the two types of events, which ensures that any physical state performing the appropriate functions will thereby ensure the instantiation of the phenomenal property of having such an experience. While I have yet to argue that phenomenal properties are in fact incapable of being individuated in purely functional terms, the above description shows that the conception of realization being developed can account for the realization of such properties if they *should* prove resistant to functional analysis. This is more than can be said for the functional definitions of Melnyk and Polger, which require that all realized entities be functionally individuatable.[51]

The potential existence of cases of the sort just described suggests that there is also yet another way in which entities that *are* functionally individuated might be realized; viz., by their realizers' occupying certain functional roles that ensure, *without being logically sufficient for*, the instantiation of the functional property that individuates them. In contrast to the aforementioned instances of realization where the entities realized are functionally individuated, the functions performed by the realizers in the sort of case now proposed will be neither identical with nor part of a functional decomposition of the

[51] The same goes for the causal definitions of Shoemaker and Gillett, for if phenomenal properties cannot be functionally individuated then they cannot be causally individuated either. Shoemaker and Gillett would, however, both probably reject the idea that phenomenal properties cannot be exhaustively analyzed in purely causal or functional terms, since they both accept a causal theory of properties, according to which properties are individuated by the causal powers they bestow on their bearers. The fact that the causal theory cannot allow for the possibility that phenomenal properties have certain non-causal individuating features just seems to me further indication of the view's inadequacy. (Biological properties also seem to present another counterexample to the view.)

individuative function of the thing they realize, but will nonetheless ensure the instantiation of the realized entity's individuative function by virtue of some non-logical (perhaps non-physical) law, which makes it such that the instantiation of the functions performed by the realizers is necessarily attended by the instantiation of the distinctive function of the thing they realize.

An example of this type of case might be found in the realization of intentional states, such as beliefs and desires. Whereas the functional individuatability of phenomenal properties is a subject of continued debate, the general consensus in contemporary philosophy of mind appears to be that intentional states likely admit of some type of functional analysis. That said, while many take the functions that serve to individuate intentional states to be certain causal or biological functions that the realizers of such states might logically suffice for in one of the ways described above, following an idea proposed by Tyler Burge (2010), one might think that intentional states are instead individuated at least partly in terms of certain *representational* functions that cannot be reduced to or fully analyzed in terms of any purely causal or biological functions. If this is so, then it follows that the causal or biological functions performed by the physical realizers of intentional states cannot be logically sufficient for them. Since it appears, however, that (as with phenomenal properties) the realizers' performance of such functions nonetheless *is* all that's needed in order to ensure the instantiation of the intentional states they realize, it remains that there must be some non-logical (perhaps non-physical) *law* correlating the two, which ensures that any physical state performing the appropriate causal/biological functions will thereby ensure the instantiation of a certain type intentional state having a certain representational function. While I have yet

to argue that intentional states are in fact individuated in terms of certain representational functions that do not admit of any causal or biological analysis, the above description nevertheless shows that the conception of realization being developed here can account for the realization of such states even if they *should* turn out to be this way. This is, again, more than can be said for the definitions of realization offered by Melnyk and Polger, which require that realizers perform the very same function that individuates the entity they realize.

Another, less speculative example of the type of realization currently under discussion can be found in the dimensioned realization of a diamond's hardness in the properties of and relations between its constituent atoms. As mentioned in our previous discussion of this case, while the diamond's hardness appears to be functionally individuatable, the function that individuates it is distinct from the functions that its atomic-level realizers ensure its instantiation by performing. In contrast, moreover, to the realization of an internal combustion engine in its various components (where the realized entity is also individuated by a function that differs from the functions performed by its various realizers), the function that individuates the diamond's hardness also cannot be decomposed into a range of subsidiary functions that its atomic-level realizers perform. The consequent "qualitative distinctness" of such realized entities from their realizers is often emphasized by Gillett (2010), who I think rightly takes it to be an advantage of his dimensioned view over exclusively "flat" functionalist conceptions of realization that it can accommodate such cases. At any rate, since the functions performed by the atomic-level realizers of the diamond's hardness are not identical with the latter's individuative function, nor do they together constitute a functional

decomposition of it, the realizers' performance of these functions cannot be logically sufficient for the diamond's hardness. As the atomic-level realizers of the diamond's hardness *are* nevertheless able to ensure the instantiation of the latter's individuative function merely by performing the various atomic-level functions that they do, it therefore remains that there must be some non-logical *law* correlating these functions, which ensures that any performance of the functions performed by the diamond's constituent atoms will give rise to an instance of the function that individuates the diamond's hardness. If this interpretation of the case is correct, then are likely to be many cases of dimensioned realization that may be classified as instances wherein the realized entity is functionally individuated, but its realizers are nonetheless merely nomologically sufficient for its occurrence or instantiation.

If all the different varieties of realization distinguished above indeed represent real or at least logically possible ways in which realizers can ensure the occurrence or instantiation of the entities they realize, then we might sum up the various forms that realization might take in the following table:

Table 1

Types of realization

	Realized entity is functionally individuated	Realizers are logically sufficient for the entities they realize
Type 1	Yes	Yes
Type 2	No	Yes
Type 3	No	No
Type 4	Yes	No

Having now distinguished the various ways in which entities can be realized, one might naturally wonder what all these different forms of realization have in common such that we are justified in continuing to speak of them as various expressions of a single relation. Why, one might ask, shouldn't we instead take the variety represented in the above taxonomy as proof of Endicott's contention that there is no single realization relation, but rather a number of distinct senses of realization, each of which corresponds to a different kind of relation? Despite the differences between them, a common thread can still be found running through the four types of realization distinguished above that is substantial enough for us to treat them as various expressions of a single relation: All realized entities, regardless of what type, have some function, and realizers always ensure the occurrence or instantiation of the entities they realize at least partly by virtue of performing certain functions. Putting this together with the various ways in which realizers can ensure the occurrence or instantiation of the entities they realize, the following may serve as a general definition of realization:

> For any entity or collection of entities *A* that performs a certain function *F* (or set of functions f_1-f_n that constitute a functional decomposition of *F*), and any entity *B* that has or bestows some function *G*, *A* realizes *B* iff *B* occurs or is instantiated at *t* in virtue of *A*'s performing *F* (or f_1-f_n) at *t*, either because *G* is individuative of *B* and *G=F* [Type 1], or because *G=F* and *G* is individuative of *B* only in conjunction with some additional property *P* whose instantiation is logically entailed by the instantiation of some non-functional property *Q* of *A* [Type 2], or because there is some non-logical law correlating *B* or its instances with the performance of *F* [Types 3 and 4].

Having hit upon a definition of realization that avoids the inadequacies of the proposals made by Shoemaker, Gillett, Melnyk, and Polger, we can now answer the question that first spurred our investigation of the realization relation: Does commitment to the view that mental properties are multiply realizable entail a commitment to functionalism? Given our definition of realization, we can see that if functionalism is the view that mental properties are wholly functional in nature, meaning that they can all be individuated in purely functional terms, the answer is clearly *no*, for our classification of the different varieties of realization reveals that there are in fact two different ways in which entities that are not functionally individuated might be realized (viz., by falling under Types 2 or 3 in the table above).[52] If any mental properties are of this sort, as I now suggest and will later argue that phenomenal properties are, then there will be some mental properties that are realizable and hence potentially multiply realizable even though they are not entirely functional in nature.

The proposed definition of realization also entails that commitment to the thesis that mental events are realized by physical states is compatible with forms of dualism (e.g. emergentism) that hold that mental events depend upon, but are nonetheless distinct from, irreducible to, and incapable of being fully explained in terms of physical events. For if all mental events are realized by physical events in the manner of Types 3 or 4 in the table above, and the physical realizers of mental

[52] Conversely, commitment to functionalism doesn't *entail* commitment to the multiple realization thesis, as one might hold that mental properties are wholly functional while maintaining that each type of mental state is nonetheless realized by a single type of physical state.

events are hence merely *nomologically* sufficient for the latter, one might also have reason to think that the laws by virtue of which the physical realizers of mental events ensure their occurrence are themselves non-physical. If this is so, then mental events will not be reducible to or fully explainable in terms of their physical realizers, for the laws which render the latter sufficient for the former will not be included in or entailed by the totality of physical facts. According to the definition of realization here proposed, physicalists hence do not have exclusive rights to the idea that the mind-body relation is best understood as a form of realization. Dualists who grant that mental events depend on physical events may accept that idea as well.

On the other hand, those (e.g. Melnyk and Jessica Wilson (2011)) who are interested in using realization to formulate physicalism may also find something of use in the above definition, which in fact enables us to more clearly see how various realization-based formulations of physicalism might differ from one another. Thus, whereas some physicalists may prefer to state their position as the view that everything is either physical or else realized by something physical in the manner of Type 1, others may wish to allow also for instances of Type 2 realization, so long as in all such cases, the additional, non-functional properties of the physical realizers needed to ensure the occurrence or instantiation of the realized entity are themselves purely physical. Still other physicalists may wish to allow for the existence of non-physical entities that are physically realized in the manner of Types 3 and 4, on the condition that in all such cases, the laws by virtue of which the physical realizers ensure the occurrence or instantiation of the entity they realize are deducible from the laws of physics. The definition offered above thus proves useful not only in

capturing a general notion of realization, but also in clarifying the commitments of and differences between various forms of physicalism and dualism.

CHAPTER 4

THE MULTIPLE REALIZABILITY OF MENTAL PROPERTIES

The next two chapters provide a defense of the claim that mental properties are multiply realizable by and consequently both non-identical with and nomologically irreducible to the physical properties they seem to depend on. Since to assess the claim that a given sort of property is multiply realizable, one must, as Ken Aizawa and Carl Gillett (2009b, p.189) point out, know what sort of entities the relevant properties are claimed as being multiply realized by, I should state at the outset that the multiple realization thesis I am interested in defending is the thesis that each mental property is realizable by distinct types of brain, or more specifically neurobiological states. Even if this thesis should turn out to be false, one might still maintain that mental states are multiply realizable, either because one thinks that such properties can be instantiated by non-organic systems (e.g. computers), which have no neurobiological states, or because such properties are, like virtually every macro-level property, multiply realizable by physical events at the quantum level. As, however, it is neurobiological, and not quantum or silicon states that mental properties seem to most naturally and directly depend on, and since the case for dualism outlined in the previous chapter would consequently be difficult to sustain if mental properties were not multiply realizable by such states, my discussion will focus on the claim that mental states are realizable by distinct types of neurobiological states. Henceforth, I shall therefore refer to this claim as *the multiple realization thesis,* and whenever I speak of mental properties as being realizable by

distinct types of physical states, it is primarily neurobiological states that I will have in mind.

The defense of the claim that mental properties are multiply realizable by and consequently both non-identical with and nomologically irreducible to their neurobiological realizers will be presented in two parts, with the present chapter devoted primarily to providing some empirical support for the multiple realization thesis, and the following chapter dedicated to drawing from it the further conclusion that mental properties are also nomologically irreducible to the brain states they seem to depend on. The present chapter is structured as follows: After briefly clarifying the role that the multiple realization thesis plays in the case for dualism, I will offer some empirical evidence for the multiple realizability of mental properties involving perceptions of color. I then provide some reasons to expect that the results of this test case will likely generalize to all mental properties, and for seeing in these results a response to the local reduction strategy advocated by David Lewis (1980) and Jaegwon Kim (1992b). I close the chapter by responding to two further objections to the multiple realization thesis raised, respectively, by William Bechtel and Jennifer Mundale (1999) and Lawrence Shapiro (2000).

1. The role of multiple realization in the case for dualism

One of the more compelling reasons for thinking that mental properties are distinct from the neural properties on which they seem to depend is that mental properties appear to be realizable by distinct types of neurobiological states. If this appearance is

correct, then mental properties cannot be identical with the neural properties they seem to depend on, for the two sorts of properties will themselves have different modal properties, and therefore must be distinct (for as noted in the previous chapter, if a mental property *M* is realizable by distinct neural properties P_1 and P_2, then *M* is necessarily instantiated whenever there is an instantiation of P_1 or P_1, but this needn't be true of P_1 or P_2, since P_1 could, e.g., be instantiated at times when P_2 is not). What reason, though, do we have to accept the thesis that mental properties are indeed multiply realizable in this way? The only reason that Hilary Putnam (1967), who introduced the thesis, offers is that the idea that mental properties are multiply realizable seems intuitively plausible. Consider again the following quote, cited in the previous chapter:

> [I]f we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus…, but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. *It seems to me overwhelmingly probable that we can do this*. (Putnam, 1967, pp.44-5, emphasis added)

Shortly thereafter, Putnam also writes:

> [T]he brain state theorist has to hope for the eventual development of neurophysiological laws that are species-independent, which seems much less reasonable than the hope that psychological laws (of a sufficiently general kind) may be species-independent, or, still weaker, that a species-independent *form* can be found in which psychological laws can be written. (Putnam, 1967, p.45, underline added)

Note first that Putnam is wrong to claim that advocates of type-physicalism (or "the brain state theory" as he calls it) must hope for the development of species-independent neurophysiological laws, for the type-physicalist may simply deny that there are any such laws, and instead hold (like David Lewis (1980) and Jaegwon Kim (1992b)) that the types of species-independent mental states postulated in psychology are not natural kinds,

but instead idealized conglomerations of various similar, yet distinct species-specific physical or functional states, and that the alleged species-independent psychological laws that involve reference to such species-independent mental states are merely simplified generalizations that owe whatever predictive success they have to their rough approximation to the laws that actually hold among the more complex species-specific physical/functional states to which such mental states can be "locally reduced."

Even setting this issue aside, Putnam's argument for the multiple realization thesis is still ultimately unsatisfying. Though we might accept his remarks as an accurate description of our intuitions, if this were all that could be said in favor of the multiple realization thesis, there would be scant reason to accept it. For this thesis is clearly an *empirical* claim, the truth or falsity of which can be decided only through scientific investigation of the way the world actually is, which is at times quite different from the way we expect it to be. In short, considering how often our intuitions lead us astray[53], (and hence how unreliable they are as a method for acquiring true beliefs), justified belief in the multiple realization thesis must be based on something more than its intuitive plausibility. Some empirical support for the thesis needed.

A good place to look for such support is vision science, and in particular the study of the physiological mechanisms responsible for color vision. I think that current knowledge of these mechanisms supports the view that mental properties involving perceptions of color are multiply realizable not only across different species, but also in the same individual human at different times. Of course, even if physiological evidence

[53] See P.S. Churchland (1986, pp.290-1) for a survey of some of the errant intuitions embedded in folk physics.

does show that color perceptions are multiply realizable in this way, the same may not be true of other mental properties. It could, e.g., be that while properties like *having a perception of yellow* are multiply realizable, mental states such as pains and beliefs are not. However, if it can be shown that mental properties involving so specialized a psychological capacity as color perception are multiply realizable, this would, I think, give us a fairly strong reason to expect, in the absence of any clear evidence to the contrary, that other psychological states and capacities whose functions are much more general, and whose realizers would be consequently be under much less selective pressure to converge on a single neurobiological type (since the more general the function, the more ways it can be executed with equal efficiency), will turn out to be multiply realizable as well. So while the conclusive confirmation or disconfirmation of a fully generalized multiple realization thesis awaits the results of future science, I suggest that taken together with certain other empirical findings discussed further below, the evidence for multiple realization in the case of color perception warrants adopting the thesis in its general form as a reasonable working hypothesis.

2. Empirical support for the multiple realizability of color perceptions

The evidence for the multiple realizability of color perceptions appears at a number of different levels and has a number of different causes.[54] I will focus on the level

[54] Aizawa and Gillett (2009a) offer additional evidence for multiple realization in color vision between the atomic and molecular levels, the molecular and cellular levels, and the cellular and tissue levels.

at which mental properties such as *having a perception of yellow* are realized in humans by sequences of physical events involving different patterns of activity across cones in the retina. The cause of multiple realization at this level that I will examine concerns the impact of simultaneous contrast on perceptions of color. The following discussion relies heavily on E. Bruce Goldstein (2014, ch.9) and C.L. Hardin (1988).

The physiological mechanisms responsible for color vision in a normal human involve at least two subsystems, one of which consists of different types of cone receptors in the retina that respond differently to different wavelengths of light, and the second of which involves the processing of the signals sent from these receptors by certain specialized cells in the retina, LGN, and visual cortex whose activity is excited by input from certain types of cones, and inhibited by others. The basic features of the first subsystem were first outlined in the trichromatic theory of color vision proposed by Thomas Young (1802) and Hermann von Helmholtz (1911) to explain the results of certain color matching experiments, which showed that normal humans are generally able to match the color perception produced by any single wavelength of light by mixing and varying the relative intensities of three other wavelengths, but are often incapable of doing so when they are given only two other wavelengths to mix together. The reason why this is so is that a normal human eye contains three different types of cone receptors, which are distinguished from one another by the absorption spectra of the photosensitive pigments they contain. These differences in their pigments make it such that one type of cone (the S cone) is maximally responsive to shorter wavelengths of light (with peak sensitivity around 419 nm), another (the M cone) is maximally responsive to middle range wavelengths (with peak sensitivity around 531 nm), and the third (the L cone) is

most responsive to longer wavelengths (with peak sensitivity around 558 nm). The response curves of these cones types are represented in the following graph:

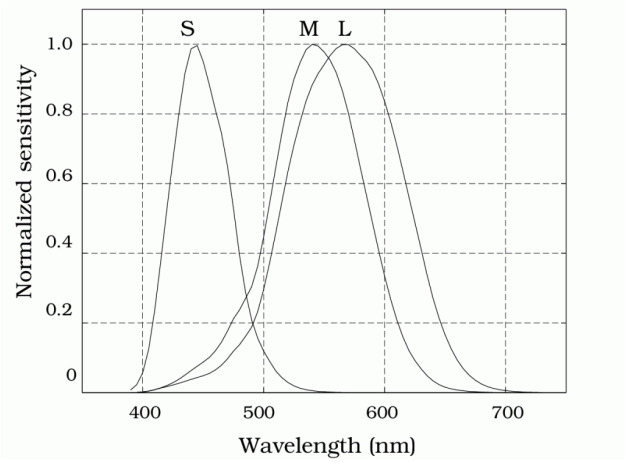


Figure 1. Spectral sensitivities of the three types of cone receptors in the human eye (Wandell, 1995).

The varying sensitivities of these three types of cones makes it such that each particular wavelength of visible light will produce its own unique pattern of activity across the cone receptors in the retina. (E.g., a wavelength of 600nm will produce a fairly high level of response in the L cones, a moderately low level of response in the M cones, and no response in the S cones.) This pattern of activity, consisting of the relative response rate of each cone type in proportion to the others, determines (in part) what color we see.

The crucial point here is that since the various cones in the retina are not each specifically tuned to respond to only one particular wavelength of light, the information that the cone receptors transmit regarding the wavelength distribution of incident light depends not on the response of any single type of cone, but rather on the ratio of their respective response rates considered in relation to one another. It is this feature of the

physiology of color vision that makes for the possibility of *metamers*, i.e. lights with different wavelength distributions that produce perceptions of the same color under the same viewing conditions. Since the colors we perceive depend (primarily) on the pattern of activity that incident light produces in our cone receptors, so long as two lights viewed under the same conditions generate the same pattern of response, they will give rise to similar color perceptions, even if they have different spectral distributions. An example of two such lights is presented in the following graph:

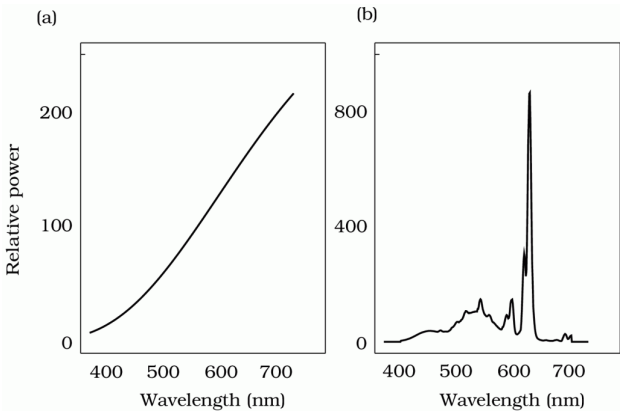


Figure 2. Wavelength distributions of two metameric lights (Wandell, 1995).

The existence of metamers shows that the type-physicalist who wishes to identify perceptions of color with certain types of physical states in the perceiver will not be able to find a unique type of response in any single type of cone with which the perception of each color can be identified, but will instead have to either (A) identify each type of color perception with a type of physical state involving a certain distinctive pattern of activity across the various cones types, or else (B) identify color perceptions with certain types of

physical states "further downstream," and treat patterns of cone activity as merely the biologically normal *cause* of such states.

Due largely to certain problems with strategy (A) that will be discussed shortly, most type-physicalists are apt to favor strategy (B). Rather than seek a single type of cone response pattern to associate with each type of color perception, such theorists will instead argue that different patterns of cone activity must converge on a single type of physical state further along in the visual system if they are to generate the same type of color perception, and that it is this single type of physical state (which will likely consist in the firing of a certain type of neuron in the visual cortex) that the perception of that color should be identified with, not any event or pattern of events in the retina. On this proposal, then, color perceptions are not realized (nor, *a fortiori*, multiply realized), but instead *caused* by cone response patterns, inasmuch as these patterns are the normal, non-deviant causes of the types of neural events in the cortex with which each type of color perception is to be identified.

While such an approach may perhaps prove viable for *experiences* of color, which can be credibly viewed as individual events that occur at a single point in time[55], reason to resist the exclusion of cone response patterns from the realization base of color *perceptions* might be drawn from the fact that perceptions are more plausibly construed as *processes*, i.e. chains of causally related events. Such a conception of perception is already embedded in the common notion that perception involves both (a) a causal interaction between the perceiver's sense organs and the external environment, and (b)

[55] More on this in the following section.

the subsequent formation in the perceiver of a representation of its environment, the content of which is determined by the information about the environment that is encoded through the environment's impact on the perceiver's sense organs.[56] As the most natural explanation for how the information captured by the perceiver's sensory apparatus is preserved and taken up into the content of the ensuing perceptual representation seems to be to say that the former stands in some causal relation to the latter, it would appear (at least according to the conception of it embodied in (a) and (b)) that perception involves at least two distinct, causally related events; which is to say that perception is itself not an event, but a process. Viewing color perceptions in this way gives us the option of treating them as extending in time from the moment the brain's response to incident light begins in the retina to the moment when the resulting visual experience of color is taken up into the content of a representational state attributing the phenomenal quality involved in that experience to some surface or medium in the perceiver's environment. This conception of color perceptions enables us to include the pattern of cone activity that occurs at the beginning of the process with which a given color perception is to be identified in the realization base of that process as a whole. Assuming, then, that perceptions of color are indeed processes that are realized in part by patterns of cone activity, if it can be shown that different patterns of cone activity can give rise to the same type of color perception, it will follow that such perceptions are multiply realizable.

The most obvious objection to the proposal that color perceptions are process whose initial stages are realized by the response of cone receptors to incident light is that

[56] Some qualifications may need to be added to (a) and (b) to accommodate instances of proprioception, wherein the things sensed and represented are parts or states of the perceiver's own body.

perceptions of color may presumably be produced by "deviant" causal chains that bypass the retina entirely and instead act directly on some part of the visual cortex (e.g. by way of electrodes). If color perceptions can be produced in such abnormal ways, then there may seem to be little reason to include patterns of cone activity in the realization base of color perceptions, since such perceptions could be generated independently of any retinal stimulation. Two things can be said in response to this objection. First, in suggesting that perceptions of color can be produced not only by light striking the retina but also by more direct action upon the visual cortex, the objection may actually end up supporting the thesis that color perceptions are multiply realizable. For given that perceptions, as standardly conceived, are processes, if the same type of perception can be initiated in two different ways so that its initial stages are realized by two distinct types of states, then there will consequently be more than one way in which that type of perception can be realized. Hence, if the perception of a certain color can, e.g., be realized both by a sequence of physical events beginning with the response of cone receptors to certain wavelengths of incident light, as well as by a sequence of events that starts with the response of certain neurons in the visual cortex to direct stimulation by electrodes, then that type of perception will *ipso facto* qualify as multiply realizable.

Second, rather than treating each type of color perception as a type of process that can be physically realized by both deviant and non-deviant causal chains, one might instead introduce a more fine-grained classification of types of color perceptions, which distinguishes between, e.g., the property of *having a normally produced perception of green*, and the property of *having an abnormally produced perception of green*. While these properties may be phenomenologically indiscernible from the perspective of

subjects who instantiate them, the differences between them may nonetheless be viewed as holding sufficient importance for biology and epistemology to justify treating them as truly distinct. Focusing, then, on those color perceptions that are produced by normal, non-deviant causes, given that such perceptions are plausibly construed as processes that begin with the response of cone receptors to incident light, it seems that the initial stages of any such perception *will* be realized by some pattern of cone activity in the retina. If this is so, then patterns of cone activity will at least figure into the realization base of any mental property involving a *normal* perception of color. Consequently, if it can be shown that different patterns of cone activity can help realize the same type of normal color perception, then it will follow that mental properties such as *having a normally produced perception of green* are multiply realizable.

Summing up, the plausible conception of color perceptions as processes that begin, at least in normal cases, with the response of cone receptors to incident light gives us grounds to reject the type-physicalist strategy (B), which excludes cone activity from the realization base of color perceptions entirely. This leaves us with type-physicalist strategy (A), which maintains that each type of color perception can be identified with a single type of physical state or process involving a single type of cone response pattern. To rebut this strategy, it will suffice to show that different types of cone response patterns can give rise to the same type of color perception. The remainder of the present section seeks to support this idea by drawing upon some recent studies of the effects of simultaneous chromatic contrast. In order to understand these effects, though, we must first examine the second component of the human chromatic visual system noted briefly above.

The general features of this second component were first described by Ewald Hering (1920) in the opponent-process theory of color vision that he developed to explain certain phenomenological observations (e.g., that the afterimage of green is red, whereas that of red is green, and likewise for blue and yellow) indicating that red and green share a distinctive connection that neither has with any other color, and likewise for blue and yellow. Hering accounted for these phenomena by postulating the existence of certain opponent processing mechanisms that respond in opposite ways to different wavelengths of light, one of which is excited by wavelengths that produce perceptions of red and inhibited by wavelengths that produce perceptions of green (or *vice versa*), and another of which is excited by wavelengths that produce perceptions of blue and inhibited by wavelengths that produce perceptions of yellow (or *vice versa*).[57]

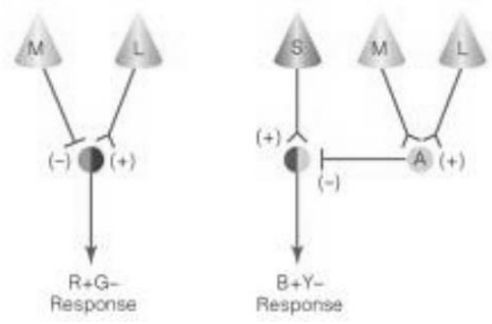


Figure 3. Example of potential neural circuitry for opponent processing (Goldstein, 2014).

These mechanisms were thought to explain the unique pairing of red and green, on the one hand, and yellow and blue, on the other, as a phenomenological effect resulting from

[57] Here and elsewhere, when I speak of wavelengths that produce perceptions of a certain color, I mean those wavelengths that produce perceptions of that color when viewed by themselves, at moderate intensity, independently of any other wavelengths of visible light.

the opposing responses of the visual system to wavelengths associated with these four colors.

The opponent responses of the visual system to these paired colors is depicted in the following graph, generated from data collected by Leo Hurvich and Dorothea Jameson (1957) through a series of hue cancellation experiments, wherein they recorded (relying on Hering's postulate that the red-green and blue-yellow systems, when stimulated with an equal amount of red and green or blue and yellow, would yield an achromatic response) how much of an opponent color had to be added to various wavelengths of light spanning the visible spectrum (e.g., how much blue had to be added to a monochromatic test light that generated a perception of yellow) to cancel out the hue of the color perception it produced and reduce it to white.

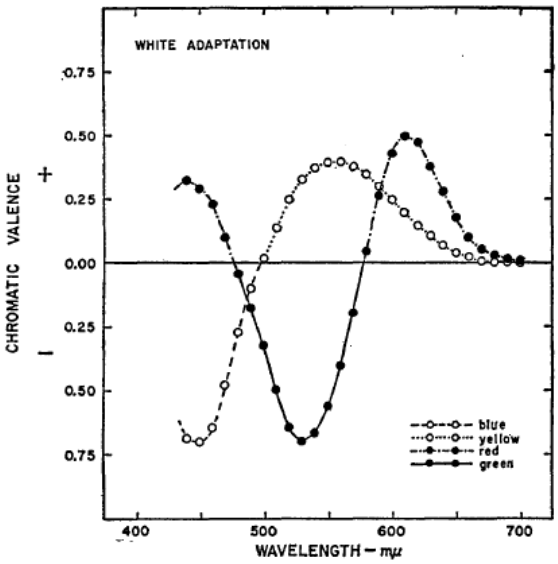


Figure 4. Chromatic response curves for standard observer (Hurvich and Jameson, 1955). The points where the curves cross the horizontal line at chromatic valence 0 represents the wavelengths at which the red-green and blue-yellow systems are stimulated with an equal amount of red and green or blue and yellow, and thus yield an achromatic response.

Groups of neurons meeting the description of these opponent processing mechanisms have since been discovered by Gunnar Svaetichin and Edward MacNichol (1958) in the retinas of fish, and by Russell DeValois (1960) and Margaret Livingstone and David Hubel (1984) in the LGN and visual cortex of monkeys, thus lending physiological support to Hering's theory.

As stated above, the interest these opponent processing mechanisms have for us lies in the part that they appear to play in the production of simultaneous chromatic contrast effects. These effects consist in the change that the color of a surface seems to undergo when viewed adjacent to surfaces of different colors under otherwise identical viewing conditions. The existence of such effects means that the colors that the surfaces in a normal setting appear to have are determined not only by the spectral distribution of light they reflect, but also by the apparent colors of other surfaces in the visual field. A surface that appears to have one color when viewed in such a way that it takes up the entire visual field may thus take on an entirely different color when viewed adjacent to other, differently colored surfaces, and there are in fact certain colors (e.g. brown) that can *only* be seen in juxtaposition with other, contrasting colors, and which hence cannot be seen isolation at all (Hardin, 1988, p.70).[58]

[58] Experiments conducted by Delk and Fillenbaum (1965), Bloj, Kersten, and Hurlbert (1999), and Boyaci, Doerschner, and Maloney (2004) indicate that apparent surface color is influenced not only by chromatic contrast but also by memory and the perceived spatial orientation of and relations between different surfaces in the visual field. The effect of memory and scene geometry on color perception lends further support to the conclusions that I'll be drawing from the effects of simultaneous chromatic contrast.

Figure 5. Example of simultaneous chromatic contrast (Ekroll and Faul, 2012a). The two circles are physically identical, but look different (the left one looks more bluish, and the right one more yellowish) due to the difference in their respective surrounds.

The existence of simultaneous chromatic contrast effects presents a problem for type-physicalists who would pursue option (A). Given that the spectral distribution of the light reflected from a surface "is a product of the spectrum of the illuminant and the selectivity – normal reflectivity – of the surface," the same surface viewed under the same illumination will produce the same pattern of response in the cone receptors that are struck by light reflected from that surface, because the spectral distribution of that light will be the same (Hardin, 1988, p.46). Since, however, the color a surface appears to have may be altered without any change in illumination simply by altering the colors of other surfaces in the visual field, it follows (a) that the same pattern of cone activity can give rise to a number of different color perceptions, and (b) that the same color perception can be generated by a number of different patterns of cone activity, for there may be a number of different ways that non-metameric[59] surfaces with different reflectance spectra

[59] Matches in the perceived color of two surfaces produced in this way cannot be explained on the grounds that while the spectral distribution of the light reflected from them is different, it differs in such a way as to generate the same pattern of cone activity (as is the case, e.g., with the metameric lights represented in Figure 2 above). For if the contrasting colors responsible for the apparent match are removed from the visual scene, the surfaces will no longer appear to be the same color, even though the illumination remains the same.

can be made to appear to have the same color under the same illumination by placing these surfaces adjacent to other surfaces of the appropriately contrasting colors. Assuming, then, that (normal) color perceptions are *realized* rather than merely *caused* by retinal activity, the effects of simultaneous chromatic contrast seem to show that the (normal) perception of a particular color can be realized by multiple types of cone response patterns.

This conclusion, however, is challenged by the traditional explanation of simultaneous chromatic contrast effects, which appeals to the fact that cone receptors become less responsive under prolonged stimulation by the same wavelengths of light. Thus, if one puts on a pair of glasses with rose tinted lenses, the L cones in one's retinas will quickly become desensitized by the constant influx of longer wavelengths of light, causing their response levels to decrease. As these adjustments are made, the redness in the visual field will fade as the activity of one's L cones slows. This process is called chromatic adaptation. In instances of simultaneous chromatic contrast, chromatic adaptation is thought to cause one's cone receptors to grow more desensitized to the wavelengths of light reflected by surfaces occupying larger portions of the visual field, thereby becoming less responsive to similar wavelengths reflected by other surfaces as well. As the sensitivity of the cone receptors to the predominant wavelengths of incident light decreases, the colors associated with these wavelengths of light are then thought to be "cancelled out" of the apparent colors of other surfaces in the visual field, thereby causing the corresponding opponent colors to become more pronounced in their appearance (Ekroll and Faul, 2012a, p.1246). This is thought to explain why, e.g., a grey surface viewed in a blue surround will appear slightly yellowish; viz. because, as one's

cone receptors adapt to the influx of light reflected by the surround, they become desensitized to wavelengths that produce perceptions of blue. Of the wavelengths of light reflected by the grey patch, one's cone receptors consequently end up being less responsive to those that produce perceptions of blue than to those that produce perceptions of yellow. This pushes the blue-yellow opponent channel in the direction of yellow, thereby resulting in the patch's yellowish appearance.

The objection that type-physicalists might raise at this point is that if the traditional explanation of simultaneous chromatic contrast just given is correct, then there is no reason to conclude from the existence of simultaneous chromatic contrast effects that the same type of color perception is realizable by different types of cone response patterns. For the sameness of color perception under different stimuli might be explained by the cone receptors' *adapting* to variations in the incident light in such a way as to yield the same pattern of activity even in response to lights with different spectral distributions.

Against this proposal, however, the results of experiments conducted by Vebjørn Ekroll and Franz Faul (2012a, 2012b) strongly suggest that chromatic adaptation in fact plays no role in the production of effects properly attributable to simultaneous chromatic contrast, and that the traditional explanation of these effects is therefore incorrect. As evidence for this conclusion, Ekroll and Faul note that there are certain laws that simultaneous contrast effects should, on the traditional account, obey, which they turn out not to. In particular, if the traditional explanation is correct, then two things should follow: (a) the shift in a surface's perceived color under simultaneous contrast should always be complementary to the color of the contrasting surface that produces the effect

(e.g., surfaces viewed in yellow surrounds should always appear more blue), and (b) the more one increases the saturation of the adjacent, contrasting color, the greater the effect of simultaneous contrast should be, since more of the adjacent color will be canceled out of the colors placed next to it. Contrary to these predictions of the traditional account, however, Ekroll and Faul found (a) that shifts in the color of a surface due to simultaneous contrast are not always complementary to the adjacent color, and that one in fact cannot know in what direction the shift in color will be without knowing the colors of both the surfaces placed next to one another, and (b) that the effect of simultaneous chromatic contrast is, beyond a certain point, inversely related to the distance between the two juxtaposed colors in color space, so that the effect that simultaneous contrast has on the perceived color of a surface will actually begin to decrease once the saturation of the surface adjacent to it exceeds a certain level.[60]

What all this shows is that while opponent processing undoubtedly plays some role in the production of simultaneous chromatic contrast effects (for it can't be a mere

[60] Ekroll and Faul (2012a, p.1248) arrived at these findings through color matching experiments, wherein "subjects adjusted the chromaticity of a central disk embedded in a violet surround…in order to make it appear as similar as possible to a disk embedded in a gray surround." The effect of simultaneous contrast on perceived color was then determined by measuring the difference in the actual chromaticity of the two disks when the adjustable one had been set to a point where the disks appeared to match in color. Of considerable interest is the fact that the effect of simultaneous chromatic contrast can be cancelled by introducing a spatial barrier in the form of a black line between the disks and their surrounds, thus suggesting that unlike the effects of chromatic adaptation, simultaneous chromatic contrast effects are heavily dependent upon local contrast. This difference leads Ekroll and Faul to propose that these effects are produced by two distinct mechanisms in the visual system that serve two different functions. They suggest that while chromatic adaptation is a temporal process whose function is to provide for color constancy (i.e., the ability to identify surfaces as having the same color under changes of illumination), simultaneous chromatic contrast is an atemporal, spatial effect, whose function is to facilitate perception of transparent objects (Ekroll and Faul, 2012a, p.1254). Further evidence for the distinct nature of the mechanisms underlying simultaneous chromatic contrast and chromatic adaption has also been found by Rinner and Gegenfurtner (2000, pp.1823-4), who report (a) that while "about 50% of total adaptational effects" occur within 40-70 ms, the effects of simultaneous contrast occur virtually "instantaneously," taking "no more than 25 ms," and (b) that while chromatic adaptation affects both the appearance of and the ability to discriminate colors, simultaneous contrast only affects color appearance.

coincidence that grey targets viewed in yellow surrounds appear more bluish, whereas grey targets viewed in blue surrounds look more yellow), chromatic adaptation evidently does not. The type-physicalist therefore cannot appeal to chromatic adaptation to maintain that the pattern of cone activity is the same for all perceptions of the same color, for matching color perceptions can be produced under conditions of simultaneous contrast, and the effects of simultaneous contrast cannot be explained in terms of chromatic adaptation. The objection to the type-physicalist strategy (A) raised earlier thus still stands: Since light from the same illuminant reflected from non-metameric surfaces with different reflectance spectra will produce different cone response patterns, and non-metameric surfaces with difference reflectance spectra can be made to appear to have the same color under the same illuminant due to the effects of simultaneous chromatic contrast, (and these effects cannot be attributed to a process of chromatic adaptation that preserves cone response pattern despite differences in stimulus), it follows that if color perceptions are indeed processes whose initial stages are (normally) realized by patterns of cone activity, then the same type of (normal) color perception can be realized by distinct types of physical states involving different types of cone response patterns.

The results of the preceding discussion can be summed up in the following argument against type-physicalism from the multiple realizability of mental properties involving perceptions of color:

(i) Perceptions of color are processes whose initial stages are normally realized by patterns of activity across cone receptors responding to incident light.

(ii) The effects of simultaneous chromatic contrast show that different types of cone response patterns can give rise to the same color perception.

(iii) Therefore, mental properties such as *having a normal perception of green* are multiply realizable.

Premise (i) is supported by an understanding of the physiological basis of trichromacy and reflection on the nature of perception. Acceptance of (i) entails the rejection of the type-physicalist strategy (B). Premise (ii) is supported by the fact that simultaneous chromatic contrast enables matching color perceptions to be produced in response to different spectral distributions, and such perceived matches in color cannot be attributed to the spectral distributions' differing in such a way as to produce the same pattern of activity across cone receptors (as is the case, e.g., with the metameric lights in Figure 2), or to the cone receptors' adapting so as to respond to the different spectral distributions in the same way. Acceptance of (ii) entails the rejection of the type-physicalist strategy (A). Since strategies (A) and (B) are, for reasons mentioned above, the only viable options for a type-physicalist treatment of mental properties involving perceptions of color, premises (i) and (ii) together entail the falsity of any such treatment. If these premises are true, then there is no single type of physical state with which the normal perception of a particular color can be identified, but at most a disjunction of such states, involving a different pattern of cone activity for each of the different ways in which non-metameric surfaces with different reflectance spectra may be made to appear to have that color by being viewed adjacent to the right contrasting colors under the same illuminant. Taken together with an analysis of color perceptions as processes that normally begin with the response of cone cells to light striking the retina, current scientific knowledge of the physiological mechanisms underlying color vision thus seems to support the conclusion that mental properties involving normal perceptions of color (e.g. the property of *having a normal*

perception of red) are multiply realizable, due to the fact that their instances are processes whose initial stages are realizable by different types of cone response patterns.

3. The multiple realization thesis generalized

As it seems reasonable to expect there to be fairly severe constraints on the range of physical states capable of carrying out the highly specialized functions performed by the chromatic visual system, mental properties involving normal perceptions of color are among those for which type-physicalism initially seems most plausible. The reasons just offered for thinking that such properties are instead multiply realizable thus suggest that, pending any evidence to the contrary, the same may likewise be true of all other mental properties whose associated functions are no more specialized (and therefore no more likely to be implemented by a single type of physical event) than those performed by normal color perceptions.

The major potential counterexample to any such generalization of the multiple realization thesis is presented by mental properties involving *experiences* (as opposed to *perceptions*) of color and other phenomenal qualities. In contrast to perceptions, which I have suggested are aptly treated as processes terminating in the formation of a representation of the perceiver's environment, phenomenal experiences are plausibly construed as events that (considered in themselves) have no representational content.[61] As the physical events that realize such experiences presumably occur further downstream

[61] The claim that experiences are non-representational is contentious. It will be defended in Chapter 6.

than the initial physiological response to the stimulus that produces them, the type-physicalist strategy (B) may consequently prove viable in the case of phenomenal experiences, even if it fails in the case of perceptions. Thus, in the case of color vision, one might suggest that even if normal perceptions of color are processes whose initial stages are realizable by distinct types of cone response patterns in the retina, *experiences* of color are *events* that are type-identical with certain types of physical events in the visual cortex. Support for this proposal might be drawn from experiments conducted by Semir Zeki (1983a, 1983b), who found certain "color coded" neurons in the V4 area of monkeys that show the same response to light reflected from a surface that appears to have the same color under different illuminants, despite the fact that opponent and "wavelength selective" cells located in the V1 respond to the different illuminants in different ways. These results might be taken to suggest that the V4 contains certain highly specialized cells that respond only during the experience of a particular color, thereby allowing for the identification of each type of color experience with the firing of a certain type of neuron in the V4.

While there are a number of reasons to question the idea that color experiences are type-identical with the activity of certain neurons in the V4[62], none of them provides conclusive evidence for the opposing thesis that experiences of color are instead multiple realizable. At this point, not enough is known about the physiological basis of phenomenal experience to enable us to say for certain whether or not different types of

[62] For one thing, the receptive fields for the V4 neurons studied by Zeki are "largely confined to the central 30 degrees" of the visual field, but humans can perceive color outside of this range (albeit not as well) (Abramov and Gordon, 1994, p.474).

neural states can give rise to the same type of experience. That said, there are still some general considerations that may be adduced in favor of an unrestricted formulation of the multiple realization thesis. The most notable of these is the brain's ability to form new neural connections so as to enable different regions to perform the functions necessary to realize a certain type of mental state when an area of the brain that previously performed those functions is damaged.[63] As this ability appears to be a general feature of the nervous system, there is no reason to think that only a select few mental states are such that the neural pathways that play a part in their realization can be rerouted so as to enable them to be realized by different neural states in different areas of the brain. If this is correct, then it should be possible at least in principle for any mental state to be realized by different types of physical states in the same individual at different times.

Aizawa and Gillett (2009b) also identify two further pieces of empirical evidence that seem to weigh in favor of a fully generalized multiple realization thesis. The first is the fact that neuronal dendritic spines change their shape and structure over time. Inasmuch as the ability of a group of neurons to realize a given mental property depends in part on the shape and structure of their dendritic spines, Aizawa and Gillett (2009b, p.202) suggest that "the implication of these findings…is that we will likely have multiple realization of psychological properties at [the dendritic spine] level." The second piece of evidence comes from the observation that the various types of proteins that help to realize mental properties in organisms are each themselves realizable by different sequences of amino acids. On the basis of this observation, Aizawa and Gillett (2009b,

[63] See Murphy and Corbett (2009). Doubts about whether such findings support the multiple realization thesis have however been raised by Shagrir (1998, pp.448-9) and Shapiro (2004, pp.59-65).

p.200) argue that “insofar as any given psychological property is realized, in part, by the properties and relations of such proteins then that property will evidently be multiply realized at the biochemical level. This appears to hold for both cognitive and qualitative properties, suggesting that all psychological properties may be multiply realized across a range of species at the biochemical level.” Type-physicalists may object to the use of these findings as evidence for multiple realizability on the grounds that the relevant realizers in the cases noted by Aizawa and Gillett are not different enough to qualify as multiple realizations of a given mental property. Criticisms of this sort will be considered and addressed below. For now, it is enough to note that even if the evidence cited by Aizawa and Gillett does not constitute conclusive proof of the multiple realizability of all mental properties, the physical differences they note among the neural realizers of mental properties can at least be taken as providing a *pro tanto* reason to view the generalized multiple realization thesis as plausible and to think that more definitive evidence in its favor may soon be discovered.

4. The local reduction strategy refuted

While the final word on the multiple realizability of phenomenal properties and the fully generalized multiple realization thesis thus awaits the results of future science, the conclusions drawn in section 2 do at least indicate that mental properties involving normal perceptions of color are realizable not merely by different types of physical states in different species, but by different types of physical states *in the same individual*. This point is significant because it forestalls a certain response that type-physicalists often

offer to claims of multiple realizability across species, which is to argue that mental properties that appear to be multiply realizable do not, in fact, pick out single natural kinds, but rather arrays of similar, but distinct kinds, each of which is unique to a particular species or domain of individuals, and identifiable with (or "locally reducible" to) a single type of physical property exemplified by the members of that species or domain.[64]

The typical response to this argument is that it does not do justice to the cross-species generalizations that we commonly make about those mental properties that the type-physicalist suggests we should divide and locally reduce in this way. For if we were to replace *pain*, e.g., with a number of distinct, domain-specific properties, such as *Human pain, Dog pain, Cat pain*, etc., as Lewis (1980) and Kim (1992b) propose, then we would also have to replace the various generalizations we are accustomed to make about pain *simpliciter* with narrower, domain-specific generalizations pertaining to the various distinct pains that are unique to each species. To do so, however, would render the predictive success of our customary generalizations about pain (which seem to hold regardless of the species they are applied to), and the evident similarities between human, dog, and cat pains, mysterious. As Jerry Fodor (1974, p.112) puts it:

> [W]e could, if we liked, *require* the taxonomies of the special sciences to correspond to the taxonomy of physics by insisting upon distinctions between the natural kinds postulated by the former wherever they turn out to correspond to distinct natural kinds in the latter…But [doing so]

[64] In support of this argument, Enç (1983, p.289), P.S. Churchland (1986, pp.356-8), and Bickle (1998, pp.121-2) note that temperature, while multiply realizable, is nonetheless locally reducible to different physical properties in different domains, the suggestion being that the same may be true of mental properties as well. Kim (1989b, p.39) argues that "'local reductions' of this sort are the rule rather than an exception in all of science."

> would…loose us precisely the generalizations which we want the special sciences to express.

In sum, if we are to accommodate the apparent validity of the broader, cross-species generalizations about mental states that psychology makes, then we must hold that instances of such states in different species are not, *ipso facto*, instances of distinct properties, but rather different instances of a single mental property that happens to be multiply realizable.[65]

The fact that the mental properties involving normal perceptions of color seem to be realizable not merely by different types of physical states across different species, but moreover by different types of physical states in the same individual makes the strategy of local reduction even more unfeasible, for if mental properties are "massively" multiply realizable in this way, then one can no longer expect to locally reduce such properties to a single type of physical state for each species. One will instead have to divide each species-specific property even further, into a range of different physical properties that a single member of that species might instantiate at different times. But now the objection raised above seems overwhelming, for if mental properties are further broken up in this manner, then the domains of psychological generalization will have to be further restricted as well. Instead of a single domain of psychological generalization covering all minded creatures, or even a single domain for each species, there would instead be a vast multitude of extremely narrow domains, each covering only those members of a given species that are currently in a specific type of physical state. But if restricting psychological generalizations so as to be applicable only to members of a certain species

[65] See also Pereboom and Kornblith (1991, pp.135-7.)

makes psychology unduly myopic, by making it insensitive to relevant similarities among instances of mental properties across species, then further restricting such generalizations so as to be applicable only to members of a certain species that are currently in a specific type of physical state makes it infinitely more so. It appears, therefore, that if mental properties are multiply realizable by different types of physical states in the same individual at different times, as mental properties involving normal color perceptions appear to be, then the strategy of local reduction advocated by Lewis and Kim cannot be pursued without raising serious doubts as to whether psychology is sufficiently general to even qualify as a legitimate science. Since, however, the scientific status of psychology is, by most contemporary standards, far less questionable than the domain-specific type-physicalism that the local reduction strategy enjoins us to accept, the evidence indicating that at least some mental properties are multiply realizable in the same individual at different times is evidence also that the local reduction strategy is impracticable.

Our investigation into the physiology of color vision has thus yielded three substantial conclusions. First, current empirical evidence seems to support the view that mental properties involving normal perceptions of color are multiply realizable. Second, since such properties serve highly specialized functions, and are consequently among those mental properties that would be most likely to be identical with a single type of physical state, given the evidence in favor of their multiple realizability, it seems reasonable to assume (pending any evidence to the contrary, and taken together with the other empirical findings noted in section 3 above) that most if not all other mental properties are likely multiply realizable as well. Third, since mental properties involving

normal color perceptions appear to be realizable by different physical states in the same individual, (and since, given our second conclusion, the same seems likely to hold of most if not all other mental properties), the local reduction strategy cannot provide an adequate solution to the problems that the multiple realizability thesis raises for type-physicalism.

5. Bechtel and Mundale's objection: The multiple realization thesis is at odds with practices and methodologies in contemporary neuroscience

The remainder of the present chapter will be spent addressing two objections to the multiple realization thesis raised, respectively, by William Bechtel and Jennifer Mundale (1999) and Lawrence Shapiro (2000). These objections come from opposite directions: Whereas Bechtel and Mundale reject the thesis on the grounds that it conflicts with certain practices and methodologies in contemporary neuroscience, Shapiro rejects the thesis as conceptually incoherent. I'll start by responding to Bechtel and Mundale's objection, and then turn to Shapiro's in the following section.

Bechtel and Mundale (1999, pp.176-7) argue that the multiple realization thesis has implications that are incompatible with certain "working assumptions" of neurobiology and cognitive neuroscience. More specifically, if the multiple realization thesis were true, then, they claim, "brain taxonomy would have to be carried out both independently of psychological function and without comparative evaluation across species," and "information about the brain [would be] of little or no relevance to understanding psychological processes." As these putative implications of the multiple

realization thesis are, Bechtel and Mundale claim, at odds with practices and methodologies of modern neuroscience, they reject the thesis *via modus tollens*.

In support of this charge of inconsistency between the multiple realization thesis and modern neuroscience, Bechtel and Mundale cite (a) the brain maps produced by Korbinian Brodmann (1909) and David Ferrier (1886), which, Bechtel and Mundale argue, classify brain states in a way that makes essential reference to psychological functions and interspecific similarities, (b) the more recent use of positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) to identify brain areas in terms of the psychological functions they correlate with, (c) the assumption neuroscientists typically labor under that discoveries linking a certain psychological function to a certain brain region in one species can be at least provisionally projected to other anatomically similar (and perhaps phylogenetically related) species as well, and (d) the guiding influence that our growing knowledge of various neural mechanisms in the visual system has had on theories of visual processing.

Bechtel and Mundale (1999, pp.177-8) also argue that the multiple realization thesis has only been made to seem plausible by the fact that "philosophers appeal to different grain sizes in the taxonomies of psychological and brain states, using a coarse-grain in lumping together psychological states and a fine grain in splitting brain states." They claim, however, that closer attention to the taxonomies used by neuroscientists reveals that the various highly determinate neural states that advocates of the multiple realization thesis typically identify as the realizers of a putatively multiply realizable mental property are "a philosopher's fiction," and that "the scientifically operative notion of a 'brain state'…is more coarse-grained and linked to an equally coarse-grained notion

of psychological state." A notion of brain state that is at least "closer to what neuroscientists…use" is, Bechtel and Mundale suggest, that of "activity in the same brain part or conglomerate of parts." They then claim that once brain states are conceived in these terms, so that mental states and their putative realizers are individuated, "as [they are] in scientific practice," at the same level of grain, "the plausibility of multiple realizability evaporates."

In response to these criticisms of the multiple realization thesis, we might first note, as Ken Aizawa (2009, p.497) points out, that Bechtel and Mundale's "claim that brain taxonomy makes essential use of psychological function should seem implausible. One would expect it to be possible to identify visually distinctive, regularly occurring parts of the mammalian brain, even without understanding what functions, psychological or otherwise, those parts might have." Given the visibly apparent structural differences among many of the major components of the brain, it does not seem as though distinguishing these components from one another should require knowledge of what they do. In light of this observation, a more charitable reading of Bechtel and Mundale might be to interpret them as claiming that reference to psychological functions is essential to brain taxonomy only in certain areas of the cortex where such visual distinctions are less easily made (Aizawa, 2009, p.497). Even this weaker claim, however, fails to follow from the arguments that Bechtel and Mundale mount on the basis of PET and fMRI studies and the brain mapping projects of Brodmann *et al*.

Ironically, certain of Brodmann's statements about his own methodology seem unequivocally at odds with the view that Bechtel and Mundale cite him in support of.

Take, e.g., the following two quotations that Aizawa produces from Brodmann's Introduction to his 1909 *Localisation in the Cerebral Cortex*:

> The subject of the following treatise is histological localization in the cerebral cortex, that is to say localization which uses *exclusively anatomical features* as the basis for investigation, in contrast to physiological or clinical aspects. (p.3, emphasis added)

> Those who find it to their taste can dress up the individual layers with terms borrowed from physiology or psychology, such as '*sensitive*' or '*perceptive*' layers, *association* or *projection* layers, '*memory*' or '*psychic*' layers, but they should not claim to be serving scientific progress in so doing. These, and all similar expressions that one encounters repeatedly today…are utterly devoid of any factual basis; they are purely arbitrary fictions and only destined to cause confusion in uncertain minds. (p.8)

Thus, according to Brodmann at least, (who has some claim to be treated as an authority on the subject), reference to psychological functions does not play *any* role, let alone an essential one, in neuroanatomical brain mapping.

This, as Aizawa (2009, pp.498-500) points out, is only what we should expect, given the methods that such brain mapping projects typically employ in developing their taxonomies, which consist primarily in the application of various "staining techniques that highlight different features of brain cells" to tissue samples taken from various parts of the brain. As these staining techniques "do not involve any knowledge of the psychological function of the tissues" to which they are applied, their use in brain mappings yields taxonomies based solely on the anatomical differences that they help render discernible. Consequently, "it simply does not appear to be the case that knowledge of function…is essential to brain mapping." As brain states can be individuated, then, on purely anatomical grounds and without reference to psychological functions, the multiple realization thesis cannot be reasonably rejected on the grounds

that being in the same mental state entails being in the same brain state because it is essential to brain taxonomies that brain states be individuated on the basis of the mental states they correlate with.

Equally unsuccessful is Bechtel and Mundale's claim that the multiple realization thesis is inconsistent with the results of PET and fMRI studies linking certain psychological functions with activity in specific areas of the brain. In this case, Bechtel and Mundale's error lies in a confusion of the notions of *realization* and *localization*. As Aizawa (2009, p.501) puts it: "unique localization means something like always occurring in the same place, [whereas] unique realization means something like always constructed in the same manner. They are entirely separate distinctions." These distinctions are indeed so separate that for a given property to be multiply realized, it is neither necessary nor sufficient that it be multiply localized. A mental property can be realizable by different types of neural states, even if the instances of these different types should always occur in the same area of the brain, and conversely, from the fact that a mental property is realized by neural states occurring in different areas of the brain, it does not follow that that mental property is multiply realized, for those different neural states, while occurring in distinct locations, might all be of the same neurobiological type (Aizawa, 2009, p.502).[66]

[66] One might see this point as providing the basis for a potential response to the argument for multiple realization from brain plasticity, for the fact that a given mental property can be realized in different brain areas through the formation of new neural pathways does not show that that mental property, when realized in a different location, is also realized by a different type of neural state. (See Shagrir (1998, pp.448-9) and Shapiro (2004, pp.59-64).)

With the distinction between realization and localization in mind, it becomes clear that while PET and fMRI studies may be taken to indicate that certain psychological functions are uniquely localized in certain regions of the brain, in that the neural activity that realizes them always takes place in the same brain areas, they do not demonstrate that psychological functions are uniquely *realized*. This is because such studies only measure levels of positron emission or blood oxygenation throughout the brain (on the assumption that increased radioactivity or oxygen usage in a given brain area is a consequence of increased neural activity in that area), yet it is entirely possible that type distinct realizations of a certain psychological function might be localized in the same brain region, involve the same (or an as yet undetectably similar) amount of overall neural activity, consume or emit the same (or an as yet undetectably similar) amount of oxygen or positrons, and hence look indistinguishable under a PET or fMRI scan, even though the type and circuitry of the neurons whose activity realizes that psychological function are sometimes different enough to warrant our classifying them as different types of brain states. In short, "sameness of [blood oxygenation] and sameness of positron emission is not sufficient to establish sameness of realization" (Aizawa, 2009, p.505). The fact that PET and fMRI studies show that similar levels of blood oxygenation or positron emission are found in the same areas of the brain when certain psychological processes occur hence does not entail that the multiple realization thesis is false.

That said, one might suggest that while the results of PET and fMRI studies do not *entail* that the mental properties they show to be uniquely localized in specific brain areas are also uniquely realized by certain neural states in those areas, sameness of localization nevertheless provides strong enough evidence for sameness of realization

that the results of such studies at least cast doubt on the claim that mental properties are multiply realized. In short, the burden of proof seems to be on defenders of the multiple realization thesis to supply some positive grounds for thinking that those mental properties identified by PET and fMRI studies as being uniquely localized are *not* also uniquely realized. Two things might be said in response to this criticism. First, it should be borne in mind that the multiple realization thesis states only that mental properties are multiply realiz*able*, not that they are actually multiply realiz*ed*. The role that the multiple realization thesis plays in the case for dualism requires only the former, weaker claim, for the potential multiple realiz*ability* of mental properties is all that is needed to establish that mental properties and their physical realizers have different modal properties, and are therefore distinct. Defenders of the multiple realization thesis hence needn't show that those mental properties that PET and fMRI studies identify as being localized in specific brain regions are also multiply realized in those areas, so long as they can offer some grounds for thinking that there are other neural states that *could* realize those same properties (even if those properties have never *actually* been realized in any of these alternative ways).

Second, defenders of the multiple realization thesis might also appeal to the empirical findings noted by Aizawa and Gillett (2009b) regarding the plasticity of dendritic spines and multiple realizability of neural proteins as evidence for thinking that even mental properties that are uniquely localized are likely to be multiply realizable by neural states involving different sequences of amino acids and/or dendritic spines with different shapes and structures. If such evidence is substantial enough to outweigh the support that sameness of localization lends to sameness of realization, then proponents of

the multiple realization thesis are entitled to maintain that the results of PET and fMRI studies fail to provide adequate grounds for denying that mental properties are multiply realizable.

Turning now to Bechtel and Mundale's assertions that the multiple realization thesis entails that psychological functions and comparative anatomy are completely irrelevant to brain taxonomy, and that knowledge of brain organization has little or no relevance to understanding psychological processes, we might first note, again following Aizawa (2009, p.506), that the uncontroversial multiple realizability of functional artifacts, such as engines, mousetraps, or screwdrivers, does not preclude us from classifying their realizers in functional terms or from engaging in comparative evaluation of the different ways they are realized "across types." Thus, while the property of being an internal combustion engine is multiply realizable if anything is, we can still functionally classify the various parts of a physical object that realizes an internal combustion engine as the combustion chamber, the exhaust, or the ignition system, and compare and classify the parts of the physical realizers of different types of internal combustion engines (e.g., 2-stroke vs. 4-stroke engines, or reciprocating vs. rotary engines) on the basis of similarities in their constitution and structure. By the same token, the proposal that mental properties are multiply realizable seems perfectly compatible with our ability to classify their realizers in terms of the psychological functions they most regularly correlate with (even if such functional considerations are not *essential* to any and all classification of brain states), and to develop brain taxonomies that take into account neuroanatomical similarities and differences across species. There is thus no reason to think that the multiple realization thesis entails that the only legitimate brain

taxonomies are those that are constructed "both independently of psychological function and without comparative evaluation across species" as Bechtel and Mundale (1999, p.177) suggest.

In attributing these implications to the multiple realization thesis, Bechtel and Mundale seem to be operating under an inflated conception of what those who endorse the thesis are committed to. To this extent, many of their remarks appear to commit a straw man fallacy, by attacking a position that no reasonable advocate of the multiple realizability of mental properties would be likely to hold. This is most evident in their declaration that their "primary concern…is with the implication drawn from the multiple realizability argument that information about the brain is of little or no relevance to understanding psychological processes" (Bechtel and Mundale, 1999, p.176). Whether or not anyone who accepts the multiple realization thesis actually draws this implication from it, there is nothing in the multiple realization thesis itself that compels them to do so. On the contrary, given that realized entities depend on their realizers, the claim that mental properties are multiply realized by distinct types of neurobiological states is not only consistent with the view that neuroscience is relevant to psychology, it actively supports it. Indeed, if mental properties in fact depend on brain states in the way the thesis affirms, the direction of relevance is likely to go both ways; i.e., both psychology and neuroscience will have some influence on the construction of theories and delineation of kinds in each other's domains, resulting in what Patricia Churchland (1986, pp.284-5) calls as a "coevolution" of the two sciences as the theories and taxonomies of each are

continually refined in response to information the other provides.[67] As Aizawa and Gillett (2009a, p.574) emphasize, this is something we can expect to find in all cases involving two sciences, one of which studies properties that are realized by certain entities studied by the other:

> [I]f one has a very well-confirmed theory of the nature of some realized property…then this theory can be used top-down to guide and even constrain research about the realizers of this property given other information about them. These realizers must result in the known powers [and any other individuating features] of the realized property, so one can exclude certain hypotheses about the realizers or prioritize others, depending on whether these hypotheses make claims about the realizers…that together allow them to…result in…the realized property. In the reverse direction, working bottom-up, if one has a well-confirmed account of the nature of the realizer properties of some realized property, this constrains theories of the realized property in various ways. For instance, precise knowledge of the realizers…can exclude or prioritize certain hypotheses about the individuating [features] of the realized property. Theories of the realized property's nature are in part plausible to the degree to which we can see that the [features] the hypothesis accords to the realized property are such that they can…result from the [features] attributed to the realizers by our well-confirmed account of the latter.[68]

In short, since realization is a dependence relation, we should expect that understanding the realizers of a realized property should be of immense usefulness in understanding the property itself, and conversely, that understanding the nature of a realized property

---

[67] While Enç (1983, pp.297-8) and P.S. Churchland (1986) seem to think that such coevolution generally tends toward and terminates in the reduction of the higher level science to its lower level counterpart, this needn't be true in all cases. Psychology and neuroscience might hence coevolve without the former being reducible to the latter.

[68] The quote has been modified to omit any commitment to the view (rejected in the previous chapter) that realized entities must be causally individuated and that it is only the causal properties of realizers that enable them to "result in" the entities they realize. Also omitted is the claim that realizers *noncausally* result in their realizers. This claim seems to be premised on the assumption that since realization is synchronic, the dependence of realized entities on their realizers cannot be causal, because causation is diachronic. Certain features of quantum entanglement seem to me to cast doubt on the idea that causation is necessarily diachronic. I hence do not see the synchronic nature of realization as precluding the possibility that the dependence of realized entities on their realizers is in some sense causal. These modifications to the above quote do not seem to detract from the force of the argument it presents.

should give us an indication as to what sorts of features its realizers are likely to exhibit. For these reasons, "it is simply false that multiple realization entails that there is no intertheoretic constraint between the sciences studying realizer and realized properties. In fact, the reverse is true" (Aizawa and Gillett, 2009a, p.574). Contrary, then, to what Bechtel and Mundale suggest, those who endorse the multiple realization thesis need not, and indeed should not deny that neuroscience is relevant to psychology.

With the above criticisms of the multiple realization thesis thus answered, Bechtel and Mundale's claim that its plausibility rests merely on an illicit asymmetry in the levels of grain at which mental properties and brain states are individuated collapses as well. For while Bechtel and Mundale are right to point out that the assessment of the multiple realization thesis depends heavily on how mental states and brain states are individuated, it seems that the most relevant, or at any rate the least arbitrary taxonomies we can appeal to in evaluating the thesis are those that are employed within the associated sciences (viz., psychology and neuroscience). As argued above, though, standard neuroscientific taxonomies classify brain states on purely anatomical grounds, and with respect to such anatomically distinct brain states, the kinds of psychology do seem to be multiply realizable. Put simply, when we assess the claim that mental properties, as individuated in psychology, are realizable by distinct types of brain states, as individuated in contemporary neuroscience, the claim seems well supported by the evidence we have. If, then, the types of brain states with respect to which mental properties are multiply realizable seem to be individuated at a finer level of grain than the mental properties themselves, their being so is not the consequence of some deceptive stratagem employed

by advocates of the multiple realization thesis, but rather because that's how mental and brain states are classified by our best scientific theories.[69]

6. Shapiro's objection: the multiple realization thesis is conceptually incoherent

The next objection to be discussed is a criticism of the multiple realization thesis raised by Lawrence Shapiro (2000). Shapiro's objection is a profound one, in that it suggests that the claim that a given entity is multiply realizable imposes a conflicting, and thus unsatisfiable set of requirements on its realizers. If correct, this means that the multiple realization thesis is not merely false, but incoherent. One might immediately protest that by making it impossible for *anything* to be multiply realizable, this objection violates the requirement, laid down in Chapter 3, that any adequate definition of realization must be such as to make it at least logically possible for any realized entity (about which everything is disregarded except its being realized) to be multiply realized. However, while it is, I think, true that the fact that Shapiro's objection renders multiple realization impossible suggests that his conception of multiple realization must be flawed in some way, it would be tendentious to dismiss the objection on these grounds alone. More must therefore be said about Shapiro's conception of multiple realization so that we may more clearly understand how it goes wrong.

[69] It is also, I think, disputable whether mental properties are in fact, as Bechtel and Mundale suggest, always less fine-grained than the brain states with respect to which they seem to be multiply realizable. After all, intentional states with highly specific content are *extremely* fine-grained, arguably more so than any reasonable taxonomy of brain states, yet it seems very likely that they are multiply realizable.

Shapiro (2000, p.643) starts by imposing the following constraint on the sorts of things that can be multiply realized: "[The multiple realization thesis], to the extent that it is true, is true of kinds that are defined by reference to their purpose or capacity or contribution to some end." Note that this statement amounts to the claim that all multiply realizable entities must have functional essences, a claim which was rejected in Chapter 3 on the grounds that entities that are not functionally individuated can nonetheless be realized, and thus potentially multiply realized. As, however, the basic thrust of Shapiro's objection does not essentially depend on his restriction of multiple realizability to functional kinds, I merely flag the contentious nature of this restriction and proceed with the reconstruction of Shapiro's argument.[70]

The key step in the argument is Shapiro's claim that the only properties of a realizer that are relevant to its being a realizer of a given kind are those that make some difference to how the distinctive function of that kind is carried out. He then argues, on the basis of this claim, that a kind only qualifies as multiply realized if certain of its realizers differ in the properties that are relevant to their being realizers of that kind; i.e., if they differ in some way that pertains to the performance of the kind's distinctive function. In his own words:

> To say that a kind is multiply realizable is to say that there are *different* ways to bring about the function that defines the kind. But, if two particulars differ only in properties that do not in any way affect the achievement of the defining capacity of a kind, then there is no reason to say that they are tokens of different realizations of the kind…[Thus,] multiple realizations count truly as *multiple* realizations when they differ in causally relevant properties – in properties that make a difference to

[70] As will be seen below, however, the central flaw in Shapiro's objection follows naturally from this dubious starting assumption, though neither entails the other.

> how they contribute to the capacity under investigation [i.e., the one that is individuative of the kind they realize]. (Shapiro, 2000, p.644)

More succinctly: "[T]wo realizations of a kind *T* are in fact different kinds of realizations of *T* only when they differ in their causally relevant properties, that is, the properties by which they contribute to the capacity, purpose, goal, and the like that serves to individuate *T* as the kind that it is" (Shapiro, 2000, p.646).

Shapiro thus offers us a conception of multiple realization according to which numerically distinct realizers of a given realized entity multiply realize that entity iff they ensure the instantiation of the distinctive function of the entity they realize by different means. Without seriously altering Shapiro's proposal, we can free it of the contentious assumption that all multiply realizable entities are functionally individuated by replacing it with the broader requirement that in order for something to be multiply realizable, certain of its realizers must ensure its occurrence or instantiation by the possession of different properties. Making this substitution gives us a constraint on multiple realization that is in keeping with the spirit of Shapiro's original idea without having to rely on the supposition that multiply realized entities must be definable in purely functional terms.

With this sort of condition on multiple realizations in place, Shapiro claims that a dilemma arises whenever we attempt to determine whether the potential realizers of a certain realized kind differ in such a way as to render that kind multiply realizable:

> The problem is this…Either the realizing kinds truly differ in their causally [or otherwise[71]] relevant properties, or they do not. If they do not, then we do not have a legitimate case of multiple realizability, and [the multiple realization thesis], in this given instance, is false. If the realizing

[71] Under the broadened version of Shapiro's constraint on multiple realization suggested above, there is no reason to stipulate that the only relevant properties of realizers are causal ones; any properties of realizers by which they ensure the occurrence or instantiation of the entities they realize may qualify as relevant as well.

> kinds do genuinely differ in their causally [or otherwise] relevant properties, then, it seems, they are different kinds. But if they are different kinds, then they are not the same kind, and so we do not have a case in which a single kind has multiple realizations.[72] (Shapiro, 2000, p.647)

Shapiro's basic objection is thus that in order to qualify as distinct *qua* realizations of a given kind, the different realizers of a putatively multiply realizable kind will have to be so markedly different in precisely those respects that pertain to the instantiation of that kind that it becomes questionable whether they are in fact realizations of the same kind at all. Put simply, there seems to be a tension within alleged cases of multiple realization between the differences that must obtain among the realizers of a multiply realized kind and the similarities that must obtain among the instances of the kind itself. If Shapiro is right, then this tension is so strong as to constitute a form of incoherence, for realizers cannot, he argues, satisfy the condition for being distinct types of realizations of the same kind unless the kinds they realize are distinct.

Why, though, should the fact that two realizers ensure the occurrence or instantiation of the entities they realize by different means entail that the entities they realize belong to different kinds? As Shapiro notes, one might e.g. hold, following Fodor (1968, p.119), that "[w]hat justifies a taxonomy, what makes a kind 'natural', is the power and generality of the theories that we are enabled to formulate when we taxonomize in that way." If, then, a certain collection of entities with distinct types of realizers can be grouped together in a way that enables us to formulate more powerful

[72] While Shapiro's point is, I think, clear enough, this last sentence is a bit infelicitous. Since two things that belong to two different kinds might also both belong to another, third kind, the fact that two realizers are of different kinds with respect to a taxonomy that classifies realizers according to the properties by which they ensure the occurrence or instantiation of the entities they realize does not *ipso facto* entail that they cannot also be of the same kind with respect to a taxonomy that classifies realizers according to the entities they realize.

and general theories, why should the fact that their realizers ensure their occurrence or instantiation by different means prevent us from treating them all as belonging to the same kind?

Shapiro's answer is that since any laws that a theory affirms about multiply realized kinds will be true only in virtue of certain regularities that obtain among their realizers, given that the realizers of a realized kind must, in order to qualify as distinct *qua* realizers of that kind, ensure the instantiation of that kind by different means, the different realizers of any multiply realized kind will have to be so heterogeneous that the only regularities likely to be found among them are those that make true certain trivial and uninteresting statements about the entities they realize, which merely attribute to the latter certain features that are contained in their definition (e.g., to use an example of Shapiro's (2000, p.650): All eyes "have the function to see"). While such trivial statements about realized kinds will indeed be supported by regularities among their realizers, (as nothing can realize a kind if it does not correlate with the things that the kind by definition correlates with), they do not qualify as laws, for "laws about functional kinds must do more than simply state the capacity in virtue of which a functional kind is the kind that it is" (Shapiro, 2000, p.649). Thus, on Shapiro's view, since all we can justifiably affirm of allegedly multiply realized kinds are these sorts of trivial statements, and such statements are not laws, or at least not the stuff of which powerful and general theories are made, it seems that according to Fodor's criteria for kindhood, by which real kinds are those that serve the formulation of powerful and general theories, distinct types of realizers are incapable of realizing the same kind. In Shapiro's (2000, p.650) words:

> [R]ealizations ought to be distinguished in kind only if they differ according to how they achieve the capacity that serves to individuate the kind they realize [or, under the broadened criterion proposed above, according to how they ensure the instantiation of the kind they realize]…Paradoxically, realizations that meet the above criterion are not, after all, realizations of the same kind precisely because there are no interesting laws that unify them.

Such, then, is the objection that Shapiro raises to the multiple realization thesis. In response, we might first take issue with Shapiro's claim that due to the differences that must obtain between realizers in order for them to qualify as distinct realizations of the same kind, the only general statements about allegedly multiply realized kinds that are likely to be supported by regularities among their realizers are trivial statements that merely affirm of such kinds features that are contained in their definition. Viewed in light of Ned Block's (1997) account of the projectibility of mental kinds (discussed at length in the following chapter), this claim seems false. As Block argues, the fact that certain multiply realizable kinds are the products of homogenizing forces (e.g. natural selection and conscious design) operating under the constraints imposed by the laws of nature gives us reason to expect that all the realizers of such a kind (and hence the kind itself) will regularly correlate not only with those features that are essential to that kind, but also with any features that have consistently played a role in its production and preservation under some homogenizing force, as well as any features that all its potential realizers must correlate with as a matter of natural law. If Block is correct, then it is simply not true that the only true generalizations we can make about multiply realizable kinds are trivial and uninteresting, because such kinds may also regularly correlate with features that are not essential to them, for the reasons just noted. Consequently, despite the differences among their realizers, we do have reason to view mental and other multiply

realizable properties as unified kinds, for only those theories that taxonomize them as such will be able to provide the most powerful and general formulations of the various non-analytic, empirical laws they figure into.[73]

However valid, this response does not yet shed full light on what I take to be Shapiro's basic mistake, which is to make the individuation of realized kinds so dependent upon that of their realizers that the fact that two realizers are of different kinds entails that the things they realize are of different kinds as well. The error of this idea becomes apparent once one allows that the job of identifying and distinguishing between kinds of realizers and that of identifying and distinguishing between the kinds of things they realize are two separate tasks, the former of which is properly assigned to the sciences that study the realizing entities, while the latter is properly assigned to the sciences that study the entities they realize. In cases where the sciences to which these tasks are allotted make use of different criteria in developing their respective taxonomies, the differences relevant to whether or not two realizers are classified as being of the same kind may be independent of those that are relevant to whether or not the things they realize are of the same kind.[74] If this is so, then the differences among realizers that determine whether or not they are distinct kinds of realizers of the same kind of thing are not what determines whether or not the entities they realize are of the same kind. If this *were* the case, then Shapiro's contention that claims of multiple realizability are

[73] Block (1997, pp.124-5) and Antony and Levine (1997, p.93) provide some examples of such non-analytic laws about mental kinds.

[74] Again, though, this doesn't entail that the taxonomy of a science that studies certain realized entities cannot be influenced by discoveries and developments in the science that studies their realizers. See Aizawa and Gillett (2011) for further discussion of the different ways in which the individuation of kinds at a given scientific level can be influenced by discoveries at lower levels.

incoherent would be correct. For if two realizers could not be different in kind unless the things they realized were too, then differences between realizers that would lead us to classify them as distinct kinds of realizations of the same kind would likewise also force us to say that they do *not* realize the same kind, but rather two different kinds. To claim that a given kind or property was multiply realized would thus be to make a self-refuting statement. Since, however, the criteria for the individuation of realized kinds may be independent of the individuation of their realizers, no such consequence need ensue. The failure to take note of this point is, I think, the major flaw in Shapiro's conception of multiple realization. Any theory of multiple realization that makes the same mistake will likewise be driven to the conclusion that multiple realization is impossible.

Looking back, one can see how Shapiro may have been led to this conclusion through the combination of his original assumption that all multiply realizable entities must be functionally individuated with an exclusively "flat" conception of realization, which holds that realized entities must be instantiated in the same individual as their realizers (so that if *F* realizes *G*, *F* and *G* must both be instantiated in the same individual *x*). One who believes that all multiply realizable entities must be functionally individuated is first of all likely to favor a functionalist conception of realization of the sort represented by Melnyk (2006) and Polger (2007) in Chapter 3, according to which for one entity to realize another is for the former to perform the latter's individuative function. If on top of this, one favors an exclusively "flat" conception of realization, then one will also hold that the function that is essential to a given realized kind and the properties of its realizers that enable them to carry out that function must both be instantiated in the same individual. The combination of these two ideas leads naturally to

the view that the features of a realizer that make it a realizer of a certain kind of thing are in fact identical with the features that individuate the kind that it realizes, or more precisely, that the properties by virtue of which a given entity realizes a certain kind are what defines the function that individuates that kind. On such a view, there is hence no "space" between the properties of realizers that are relevant to their being realizers of a certain kind and the individuative functions of the kinds they realize, because the former are what define the latter. Consequently, one could not alter the causal, historical, or representational properties of a realizer that enable it to carry out the individuative function of a certain kind without thereby depriving it of the ability to carry out that function and thus realize that kind. At best, one will leave it with a set of properties that amount to the performance of some different function, by virtue of which it realizes some different functionally individuated kind. Assuming, then, that realizers are to be type individuated on the basis of differences in the properties by virtue of which they perform the individuative function of the kinds they realize, two realizers could not, on such a view, be of different types and realize the same kind.

Thus, by the above reasoning, the combination of the idea that multiply realized entities must be functionally individuated with the idea that realization is an exclusively "flat" relation might easily lead one to think that multiple realization is impossible. As, however, we have no more reason to accept these two ideas now than we did in Chapter 3, when we rejected them both, we likewise have no reason to accept Shapiro's claim that the multiple realization thesis is incoherent simply because it seems naturally to follow from their conjunction. *Pace* Shapiro, we therefore cannot dismiss the empirical evidence in favor of the multiple realization thesis on the grounds that the thesis contradicts itself.

In conclusion, since current empirical evidence suggests that mental properties are realizable by distinct types of neurobiological states, and objections to the multiple realization thesis (or at least those discussed above) have thus far proven unsuccessful, we seem to have good reason to believe that mental properties cannot be identified with types of neurobiological states. This, however, leaves open the possibility that every token *instance* of any mental property is nonetheless identical with a token instance of some type of physical state. Should this be the case, then even if mental *properties* are distinct from the physical properties on which they seem to depend, mental *events* (i.e. instances of mental properties) would nevertheless be token-identical with physical events (i.e. instances of physical properties). The results of the present chapter are hence insufficient to establish mind-body dualism, as there is a form of physicalism (viz. token/non-reductive physicalism) which is compatible with the multiple realization thesis. The refutation of this alternative form of physicalism will be the task of Chapter 6. There also remains the possibility, equally incompatible with dualism, that mental properties, though multiply realized by neurobiological states, may nonetheless be nomologically reducible to the latter through a more general reduction of psychology to neuroscience. The following chapter will make the case that no such reduction is likely to occur.

CHAPTER 5

THE NOMOLOGICAL IRREDUCIBILITY OF MENTAL PROPERTIES

Having offered, in the previous chapter, some empirical support for the thesis that mental properties are multiply realizable, the present chapter is devoted to the defense of an important conclusion that is often drawn from that thesis, which is that mental properties are consequently also *nomologically irreducible* to their physical realizers, meaning that they cannot be identified with the physical properties they seem to depend on as part of a more general reduction of psychology to neurobiology or some more basic physical science. After providing a summary, in the first section of the chapter, of Jerry Fodor's (1974) attempt to derive this conclusion from the multiple realization thesis, the remaining sections of the chapter will be spent evaluating two objections to Fodor's argument that have been raised, respectively, by Jaegwon Kim (1992b) and John Bickle (1998).

1. Fodor's argument for the nomological irreducibility of mental properties

According to the classical account of intertheoretic reduction developed by Ernest Nagel (1961), the reduction of one scientific theory to another consists in the deduction of the laws of the reduced theory from the laws of the theory to which it reduced. In "heterogeneous" cases, where the reduced theory contains terms that the reducing theory does not, the laws of the reducing theory must be conjoined with certain "bridge laws"

identifying[75] the properties that such terms refer to with certain properties of the reducing theory in order for the deduction of the reduced theory to go through. Whether or not intertheoretic reductions in fact have the deductive form Nagel ascribes to them, I will call reductions of properties that are achieved through the reduction of the theory they belong to *nomological* in reference to the fact that such reductions of properties are accomplished by reducing the laws that the reduced theory posits as governing them to the laws that the reducing theory posits as governing the properties to which they are reduced. As a certain type of property is thus nomologically irreducible iff the science that studies it is *autonomous*, in the sense that the proprietary laws of that science cannot be reduced to those of any other science, the claim that mental properties are nomologically irreducible can be seen to go hand in hand with the claim that psychology is an autonomous science.

The idea that these two claims can be derived from the multiple realization thesis is customarily credited to Jerry Fodor (1974). Fodor's argument for this conclusion proceeds from two plausible assumptions: (1) that "the natural kind predicates of a science are the ones whose terms are the bound variables in its proper laws," and (2) that, conversely, "a necessary condition on a universal generalization being lawlike is that the predicates which constitute its antecedent and consequent should pick out natural kinds"

[75] Richardson (1979, pp.548-50) notes that Nagel (1961, p.355fn) also allows for bridge laws that have the form of biconditionals, or even one-way conditionals from the terms of the reducing theory to those of the reduced theory, rather than statements of identity. As, however, the potential reduction of psychology to neuroscience that is at issue here is one that would show mental properties to be nothing more than, rather than merely biconditionally correlated with or conditionally dependent on neurobiological properties, the requisite bridge laws must, in the case we are interested in, be identity statements. (See Fodor (1974, pp.99-100), as well as Hooker (1981b, pp.201-4), who makes a strong case for "employing identities in reductions rather than correlations or nomic connections.")

(Fodor, 1974, pp.102, 108). Put simply, a predicate of a science refers to a natural kind iff it figures into a law of that science. The second step in the argument consists in the observation that if mental properties are multiply realizable (as the results of the previous chapter suggest), then there can be no bridge laws identifying or correlating any mental property with any one physical kind. One will instead have to try and identify or correlate each mental property with the disjunction of all the distinct physical kinds that seem capable of realizing it. Thus, if a certain mental property M is realizable by a range of distinct physical kinds P_1, P_2,… then the bridge law linking M to these kinds will have to be one that identifies or correlates M with the disjunction of P_1, P_2,…. Such a bridge law will then have the form: $M = P_1 \vee P_2 \vee \ldots$, or, alternatively (if not equivalently), $(\forall x)(Mx \equiv (P_1x \vee P_2x \vee \ldots))$.[76]

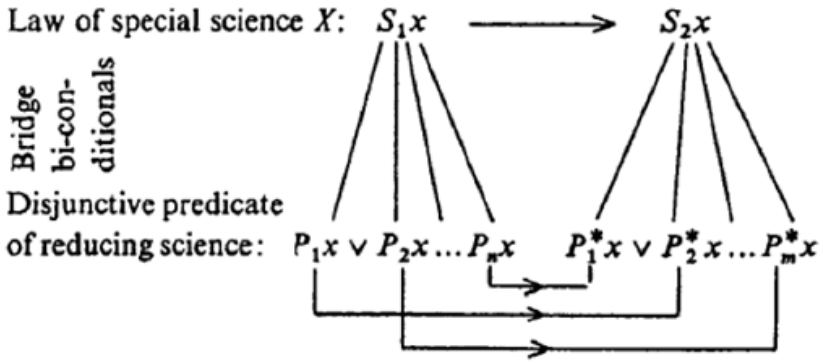
At this point, a problem arises for any attempt to reduce the laws and properties of psychology to those of any physical science. For while P_1, P_2,…, inasmuch as they are each physical kinds, will figure into various physical laws correlating their respective instances with other physical kinds, the same is unlikely to hold of their disjunction, as there is presumably no physical law stating, for any physical kind P^*, that $(\forall x)(P^*x \supset (P_1x \vee P_2x \vee \ldots))$, or that $(\forall x)((P_1x \vee P_2x \vee \ldots) \supset P^*x)$. Given, then, that "a necessary condition on a universal generalization being lawlike is that the predicates which constitute its antecedent and consequent should pick out natural kinds," it follows that

[76] While the bridge laws linking mental properties with their realizers must, again, take the form of identity statements if the reduction they facilitate is to yield a form of type-physicalism, since biconditionals are weaker than identity statements, one need only show that there can be no biconditional bridge laws between mental properties and their realizers in order to demonstrate that no such reduction is forthcoming.

while P_1, P_2,… each individually qualify as natural kinds, their disjunction $P_1 \vee P_2 \vee\ldots$ does not (Fodor, 1974, p.108). The point is quite general: a disjunction of natural kinds is not itself a natural kind. As Fodor points out, though, "this is tantamount to allowing that at least some 'bridge laws' [viz., those identifying the kind predicates of one science with a disjunction of kind predicates of another science] may, in fact, not turn out to be laws," since some of the predicates that such bridge statements equate with one another (viz., those that consist of disjunctions of kind predicates) will fail to pick out natural kinds (Fodor, 1974, p.108).

For these same reasons, the statements in the language of any physical science that a certain set of bridge statements are used to deduce the laws of psychology from will also fail to be laws. For if it is a psychological law, e.g., that $(\forall x)(M_1x \supset M_2x)$, (where M_1 and M_2 are psychological predicates that refer to distinct mental kinds), then assuming that M_1 is multiply realizable by physical kinds P_1, P_2,…, and M_2 is multiply realizable by physical kinds P^*_1, P^*_2,…, the bridge statements equating M_1 and M_2 with predicates of the science concerned with such physical kinds must state that $(\forall x)(M_1x \equiv (P_1x \vee P_2x \vee\ldots))$ and $(\forall x)(M_2x \equiv (P^*_1x \vee P^*_2x \vee\ldots))$. Using these bridge statements to translate the psychological law $(\forall x)(M_1x \supset M_2x)$ into the language of that physical science, one then ends up with the statement $(\forall x)((P_1x \vee P_2x \vee\ldots) \supset (P^*_1x \vee P^*_2x \vee\ldots))$. But since neither $P_1 \vee P_2 \vee\ldots$ nor $P^*_1 \vee P^*_2 \vee\ldots$ pick out natural kinds, the statement $(\forall x)((P_1x \vee P_2x \vee\ldots) \supset (P^*_1x \vee P^*_2x \vee\ldots))$ cannot be a physical law, even though the predictive success of the statement $(\forall x)(M_1x \supset M_2x)$ may, and likely will be supported by the existence of certain physical laws stating, e.g., that $(\forall x)(P_1x \supset P^*_1x)$, $(\forall x)(P_2x \supset$

P^*_2x), and $(\forall x)(P_3x \supset P^*_3x)$. Fodor (1974, p.109) represents this situation by way of the following diagram:



The upshot of all this is that according, at least, to the classical, Nagelian model of intertheoretic reduction, no science that studies lawfully related multiply realizable properties can be reduced to a science that studies the properties by which those of the former science are multiply realizable. For without the requisite one-to-one "correspondences between the natural kinds of the reduced and the reducing science," the bridge statements connecting the predicates of the two sciences, and the statements in the language of the would-be reducing science into which the laws of the science targeted for reduction are translated, will both include disjunctive predicates that fail to pick out natural kinds, and both will therefore fail to qualify as laws (Fodor, 1974, p.110). From this it follows that the laws of psychology will not be deducible from those of any physical science that studies properties by which mental properties are multiply realizable. For while the following reasoning (using the terms of our previous example) seems compelling[77]:

[77] Since the antecedents of 1., 2., and 3. are different from that of 4., the inference from 1., 2., and 3. to 4. may not be strictly valid, but I take it that the point is clear enough.

1. $(\forall x)(P_1x \supset P^*_1x)$
2. $(\forall x)(P_2x \supset P^*_2x)$
3. $\underline{(\forall x)(P_3x \supset P^*_3x)}$
4. $(\forall x)((P_1x \vee P_2x \vee P_3x) \supset (P^*_1x \vee P^*_2x \vee P^*_3x)$
5. $(\forall x)(M_1x \equiv (P_1x \vee P_2x \vee P_3x)$
6. $\underline{(\forall x)(M_2x \equiv (P^*_1x \vee P^*_2x \vee P^*_3x)}$
7. $(\forall x)(M_1x \supset M_2x)$

since 4., 5., and 6. are not laws, the nomological force of 1., 2., and 3. is not preserved across the inference to 7., and the status of 7. as a psychological *law* therefore cannot be deduced from the physical laws 1., 2., and 3. Thus, if mental properties are indeed multiply realizable, and the generalizations that psychology makes about them indeed have the status of laws, then psychology is irreducible to any physical science that studies the properties by which mental properties are multiply realizable, and mental properties likewise cannot be nomologically reduced to the proprietary kinds of any such science.

If sound, Fodor's argument hence shows that the multiple realization thesis has major implications regarding the relationship between psychology and the physical sciences. The remainder of the present chapter will be dedicated to defending the argument against two noteworthy objections.

2. Objections to the multiple realization argument for the nomological irreducibility of mental properties

Since its publication, Fodor's argument for the autonomy of psychology and the nomological irreducibility of mental properties has won widespread acceptance and played a major part in establishing token- or "non-reductive" physicalism as the

longstanding orthodoxy in contemporary philosophy of mind. Though written over 20 years ago, Jaegwon Kim's (1992b, p.1) remark that "it is part of…conventional wisdom in philosophy of mind that psychological states are 'multiply realizable,'…[and that this] refutes psychophysical reductionism once and for all" largely still holds true today. The past two decades have, however, produced a number of challenges to this piece of conventional wisdom, which have subjected the non-reductive implications that Fodor draws from the multiple realization thesis to increasing critical scrutiny. As space does not permit an exhaustive survey of these challenges[78], I will focus solely on two that to me seem the most trenchant and representative of the various sorts of objections that have been raised. These two challenges are presented, respectively, by Kim (1992b) and John Bickle (1998).

2.i. Kim's objection: Multiply realizable kinds are unprojectible, and hence unsuited to participate in laws

One of the better known objections to Fodor's argument originates with Kim (1992b). Kim's objection is surprising in that he argues that the autonomy of psychology, which Fodor claims follows from the multiple realization thesis, is in fact *inconsistent*

[78] For a brief sampling, see Richardson (1979; 1982), Enç (1983), P.S. Churchland (1986), and Sober (1999). One of Richardson's central criticisms was addressed in footnote 75 above. (Kitcher (1980; 1982) offers a more comprehensive response.) Most of Enç and P.S. Churchland's objections are addressed in the discussion of Bickle (1988) below. And while Sober is, I think, right to emphasize the compatibility of multiple realization with the continued relevance and legitimacy of lower level explanations, this point does not, I think, pose any serious threat to Fodor's conclusions, for the claim that a psychological kind is nomologically irreducible is consistent with the claim that instances of any properties that are nomically correlated with that kind are also nomically correlated with, and hence explainable in terms of, its various realizers.

with the multiple realizability of mental properties. He arrives at this conclusion by way of an attempted diagnosis of what it is that prevents disjunctions of kinds from figuring into legitimate laws or serving as legitimate properties to which mental properties might be nomologically reduced. While crucial to the implications that non-reductivists wish to draw from the multiple realization thesis, the idea that disjunctions of kinds are unsuited for such roles is, as Kim notes, more asserted than argued for by Putnam and Fodor. Putnam (1967, p.45, emphasis added) simply dismisses out of hand the idea that multiply realizable properties might be reduced to disjunctions of kinds: "Granted, [if mental properties are multiply realizable,] the brain-state theorist can save himself by *ad hoc* assumptions (e.g. defining the disjunction of two states to be a single 'physical-chemical state'), *but this does not have to be taken seriously*." And while Fodor argues that disjunctions of kinds cannot figure into the bridge laws needed to effect a nomological reduction of multiply realizable properties because such disjunctions are not natural kinds, their failure to count as kinds is then said to consist in the fact that they do not figure into any legitimate laws, which leads us around in a circle. Kim's pursuit of a more satisfactory explanation for the exclusion of disjunctions of kinds from participation in laws thus seems well motivated.

Kim (1992b, p.11) suggests that the reason why disjunctions of kinds cannot figure into laws is that such disjunctions are not *projectible*[79]; i.e., non-trivial general

[79] The notion of projectibility was first introduced by Goodman (1955). While Goodman thought that the projectibility of a predicate depends solely on its degree of "entrenchment" in a language, and that projectibility is hence a language-relative matter, I'll follow general consensus in taking the projectibility of a predicate to be an indication that the property it picks out is a natural kind that figures into objective, language-independent laws. (See Hooker (1981b, pp.215-8, esp. fn32) for a forceful defense of this view.)

statements involving such disjunctions are not confirmed by observation of positive instances. If true, this would explain why such disjunctions are unsuited for use in laws, for it is a hallmark of genuinely *nomological* generalizations that they *are* confirmed by, and thus derive inductive support from, their positive instances.

In support of this proposal, Kim (1992b, p.11) cites the example of jade, which, "as it turns out, is not a mineral kind,…[but is instead] comprised of two distinct minerals with dissimilar molecular structures, *jadeite* and *nephrite*." Kim argues that this discovery should lead us to conclude that generalizations about jade (e.g. "Jade is green") are in fact not projectible, and therefore cannot have the status of laws.

> For we can imagine this: on re-examining the records of past observations, we find, to our dismay, that all the positive instances of [the generalization "Jade is green"]…turn out to have been samples of jadeite, and none of nephrite! If this should happen, we clearly would not, and should not, continue to think of ["Jade is green"] as well confirmed…But all the millions of green jadeite samples are positive instances of ["Jade is green"]…[H]owever, ["Jade is green"] is not confirmed by them…And the reason, I suggest, is that jade is a true disjunctive kind, a disjunction of two heterogeneous nomic kinds which, however, is not itself a nomic kind. (Kim, 1992b, p.12)

The equivalence of jade with the disjunction of two distinct kinds is thus taken to show that jade is not a natural kind capable of figuring into real laws, because (non-trivial) generalizations about jade are not projectible. Positive instances of such generalizations can only confirm, at best, more restricted generalizations pertaining solely to either jadeite or nephrite, and it is hence only jadeite and nephrite that are eligible to participate in laws.

I now introduce a notion that will prove useful for the remainder of the present discussion. A property *P* will be said to be *nomically equivalent* with a disjunction of

kinds *D* iff "in all possible worlds that are compatible with laws of nature," nothing can instantiate or be an instance of *P* without instantiating or being an instance of some member of *D*, and *vice versa* (Block, 1997, p.110). According to this definition, a multiply realizable property may hence be nomically equivalent with the disjunction of its potential realizers even if, as Fodor argues, there can be no bridge law identifying the two. Thus, even if it is not a law, e.g., that $(\forall x)((\text{Jade})x \equiv ((\text{Jadeite})x \vee (\text{Nephrite})x))$, jade may nonetheless be said to be nomically equivalent with the disjunction (Jadeite $\vee$ Nephrite), since in all nomologically possible worlds, nothing can be jade without being jade or nephrite.

The crucial step in Kim's argument consists in his claim that what goes for jade goes also for all properties that are nomically equivalent with disjunctions of kinds. Thus, if the multiple realization thesis is true, and mental properties are hence nomically equivalent with disjunctions of distinct physical kinds, then such properties, Kim argues, are, like jade, ineligible to participate in laws, because unlike genuine laws, non-trivial generalizations about such properties will not be confirmed by their positive instances.[80] As Kim (1992b, p.15) puts it:

> If pain is nomically equivalent to…[a] wildly disjunctive and obviously nonnomic [disjunction of distinct physical kinds], *why isn't pain itself equally heterogeneous and nonnomic*…? Why isn't pain's relationship to its realization bases…analogous to jade's relationship to jadeite and nephrite? If jade turns out to be nonnomic on account of its dual "realizations" in distinct microstructures, why doesn't the same fate befall pain?

[80] This line of reasoning leads naturally to the local reduction strategy rejected in the previous chapter, for if it is only generalizations about the individual realizers of multiply realizable properties that turn out to be projectible and hence eligible to participate in laws, it would seem to make sense to divide such properties up and reduce them to the more restricted kinds about which such law-like generalizations hold.

In short, Kim argues that anything that is nomically equivalent with an unprojectible disjunction of kinds must itself also be unprojectible and thus incapable of participating in laws. So if mental properties are, as the multiple realization thesis entails, nomically equivalent with disjunctions of physical kinds, and disjunctions of kinds are unprojectible, then it seems that mental properties must be unprojectible as well.

If Kim is right, then it looks as though Fodor's argument backfires, for in arguing that psychology cannot be reduced to any physical science because the disjunctions of physical properties with which mental properties are nomically equivalent are not kinds and therefore cannot participate in laws, he invites the question why the same does not hold of mental properties as well. Far from ensuring the autonomy of psychology and nomological irreducibility of mental properties, the multiple realizability of mental properties thus instead seems to entail that there are no (general, species-independent) psychological laws, no (general, species-independent) psychological kinds, and no (general, species-independent) science of psychology.

2.i.a. Fodor's response

Two different responses to Kim's objection to Fodor's argument for the autonomy of psychology have been offered, respectively, by Ned Block (1997) and by Fodor (1997) himself. Fodor's approach is to concede that generalizations about jade are, as Kim claims, unprojectible, while rejecting the analogy that Kim draws between properties like jade, which Fodor labels "(merely) disjunctive," and those, e.g. mental properties, which are truly multiply realizable.

According to Fodor (1997, p.153), the reason why generalizations about jade are unprojectible is that jade, like other "(merely) disjunctive" properties, is simply *identical* with the disjunction of its realizers. Hence, "being jade *just is* being" either jadeite or nephrite, and any truths about jade likewise *just are* truths about jadeite or nephrite (or both). Generalizations about jade are thus not confirmed by their positive instances simply because "*there are no general empirical truths about jade as such*…for samples of jade, as such, to confirm."

Mental and other multiply realizable properties differ from jade in this respect, for it is, Fodor claims, a defining characteristic of multiply realizable properties that there *are* laws about them as such; i.e., there are general empirical truths about them (of which the special sciences give us numerous examples) that are not mere abbreviations for conjunctions of truths about their realizers.[81] For this reason, unlike merely disjunctive properties, multiply realizable properties *cannot* be identified with the disjunctions of their possible realizers, because such disjunctions are not eligible to participate in laws. If multiply realizable properties indeed differ from merely disjunctive properties in this way, then there is of course no grounds to infer that since the latter are unprojectible, the former must be so as well, for the very distinction between the two sorts of properties rests on the idea that there *are* projectible generalizations about multiply realizable

[81] Fodor (1997, p.153) actually claims that the difference between merely disjunctive and multiply realizable properties consists more fundamentally in the fact that a "(merely) disjunctive" property like jade "has no realizer in *any* metaphysically possible world that it lacks in the *actual* world," whereas multiply realizable properties "have different bases in different worlds." This, however, strikes me as inessential to the distinction between the two sorts of properties, which I think may instead be said to consist simply in the fact that multiply realizable properties are distinct from the disjunctions of their realizers and suited to participate in laws, whereas merely disjunctive properties are not.

properties, whereas there are no statements at all, whether projectible or unprojectible, about merely disjunctive properties as such.

In claiming that multiply realizable properties differ from merely disjunctive properties like jade in this way, Fodor may seem to beg the question against Kim, for the projectibility of multiply realizable properties and thus their ability to participate in laws is precisely what Kim disputes. Fodor argues, however, that there are in fact good reasons to believe that multiply realizable properties can and do participate in laws, and that they are therefore not merely disjunctive. These reasons, it turns out, are closely related to what makes the disjunctions of the realizers of multiply realizable properties *un*suited to participate into laws. Put simply, Fodor's suggestion is that we need a place for multiply realizable properties in laws for the very reason that general statements about the disjunctions of their realizers do not adequately account for certain regularities across the members of such disjunctions that our evidence often reveals.[82]

In support of this idea, Fodor first points out that what prevents disjunctions of the realizers of multiply realizable properties from participating in laws is different from what prevents disjunctions of kinds that are identifiable with merely disjunctive properties. The problem with disjunctions of the latter sort, e.g. *jadeite or nephrite*, is simply that "the nomic properties that a thing has qua F or G are either properties that it has qua F or properties that it has qua G" (Fodor, 1997, p.157). In short, there can be no laws about such disjunctions as such, because any truths about them *just are* truths about their disjuncts. In contrast, the problem with general statements involving disjunctions of

[82] See also Pylyshyn (1984, ch.1).

the realizers of multiply realizable properties is that such statements "*suggest missed generalizations*" (Fodor, 1997, p.158). As Fodor (1997, p.158) puts it:

> To offer a law of the form $R1 \vee R2 \vee ... \rightarrow Q$ is to invite the charge that one has failed correctly to identify the property in virtue of which the antecedent necessitates the consequent…Someone who offers such a law undertakes a burden to provide positive reason that there isn't a *higher level* but *non*disjunctive property of things that are R1 ∨ R2…in virtue of which they bring it about that Q.

What is it, though, that explains our preference in such cases for laws stated in terms of higher level, nondisjunctive properties to those that are stated in terms of disjunctions of all the states that are capable of realizing such properties? Why should we accept the former while rejecting the latter? Fodor (1997, p.159) offers the following answer:

> [W]e [are] prepared to buy…laws [with nondisjunctive antecedents] at the cost of reifying high level properties…[because] this policy complies with an injunction that all of our inductive practice illustrates: *Prefer the strongest claim compatible with the evidence, all else equal*.

Fodor's suggestion is thus that in cases where a number of distinct types of states R_1, R_2,… are found to be regularly correlated with the same type of state Q, our inductive practice *requires* the hypostatization of a higher level, nondisjunctive property that captures whatever it is that R_1, R_2,… have in common, by virtue of which they all correlate with Q. This is not because a law stated in terms of the disjunction of R_1, R_2,… would be incompatible with the available evidence, but rather because the regularities that the evidence reveals in the correlation of these distinct types of states with Q warrants a stronger claim, (viz., that they all correlate with Q because they all share or realize a certain higher level, nondisjunctive property that is nomologically linked with Q), and, as Fodor (1997, p.158) puts it, "[a]ccepting the strongest generalizations that

one's evidence confirms is *what induction is about*."[83] In short, the conjunction of (a) the inductive axiom to opt for the strongest hypotheses that our evidence confirms with (b) the existence of evidence indicating the regular correlation of distinct types of states with the same type of state gives us reason to believe that there are multiply realizable properties that are distinct from the disjunctions of their realizers and suited to participate in laws.[84]

The same, however, cannot be said of merely disjunctive properties, for when dealing with disjunctions of kinds like *jadeite or nephrite*, any overlap in the correlations that the various disjuncts enter into can be adequately explained in terms of the lower level sciences that study such kinds. Evidence for such overlap hence does not warrant the postulation of any higher level, nondisjunctive property, for with a satisfactory lower level explanation of these common correlations already in hand, no stronger claim involving commitment to some higher level, nondisjunctive property that each of the disjuncts realize is required. We thus remain content in such cases to frame our laws in terms of the individual disjuncts themselves, treating any true general statements about

---

[83] There actually seem to be two further conditions that these higher level generalizations must meet in order for Fodor's argument to go through. First, as Antony and Levine (1997, pp.91-2) and Shapiro (2000, p.649) point out, it seems essential to Fodor's argument that the relevant generalizations be "non-analytic"; otherwise, there would be no reason to view laws about multiply realizable kinds as making stronger claims compatible with the evidence than corresponding laws stated in terms of the disjunctions of their realizers, for such laws would be true by definition, and hence trivial. Second, it seems that the higher level generalizations must have some independent motivation, for only in such cases does it seem warranted to prefer a stronger claim to a weaker one that is just as compatible with current evidence. (If "$(\forall x)(Fx \supset Gx)$" is confirmed by current evidence, then so is the stronger claim that "There is an undetectable entity in my nose and $((\forall x)(Fx \supset Gx))$," but that doesn't seem to give us any reason to confer nomological status on the latter statement instead of the former.) (See Sober (1999, p.557fn).)

[84] Further reasons for distinguishing multiply realizable properties from the disjunctions of their realizers might, in the case of mental properties, be drawn from certain features of the properties themselves (e.g., phenomenal or intentional features), which their realizers, being physical, may be thought incapable of having. These features will be discussed at length in the following chapter.

the disjunction as a whole as equivalent to conjunctions of such laws. The same is not true of multiply realizable properties, for while there will typically be laws of the lower level sciences correlating each of the realizers of a multiply realizable property with those states that the multiply realizable property is itself nomologically correlated with (or with states that realize those states), the lower level sciences will not be able to provide an adequately unified explanation for why these realizers all happen to correlate with the same group of states (or with states that realize the same group of states).[85] The most these sciences will be able to do is enumerate each of the various laws that relate each type of realizer to each of the various states (or some realizer of each of the states) that the realizers are all collectively correlated with.[86] This is why the evidence of common correlations in such cases supports the postulation of a higher level, nondisjunctive, multiply realizable property that is nomologically correlated with those states that all of its realizers are nomologically correlated with.

If the above reasoning is sound, then contrary to one of the central assumptions of Kim's objection, genuinely multiply realizable properties can be nomically equivalent with disjunctions of physical kinds and yet be suited to participate in laws even though the disjunctions with which they are nomically equivalent are not. This is because laws stated in terms of such properties make stronger claims than, while being equally

[85] Sober's (1999, p.551) response that "there is no objective reason to prefer the unified over the disunified explanation" strikes me as inadequate. While he is I think right to claim that the disunified lower level explanations are not incompatible with or invalidated by the unified higher level one, the latter is nonetheless preferable to the former inasmuch as it explains something that the former do not; viz., why these different lower level kinds all correlate with the same things. When the lower level sciences cannot explain why this is so, postulation of some higher level property seems necessary to account for this fact. (See Antony (2003, pp.16-18).)

[86] See also Antony and Levine (1997 p.91).

consistent with the available data as general statements about their nomically equivalent disjunctions, and it is part of our inductive practice to favor the strongest claims that our evidence confirms. In this respect, multiply realizable properties differ from merely disjunctive properties, which our inductive practices in conjunction with the available evidence give us no reason to distinguish from the disjunctions of their realizers. Barring a major overhaul of our use of induction, there thus seems to be no way to assimilate multiply realizable properties to merely disjunctive properties without disputing the evidence that our inductive practices put to use in distinguishing these types of properties from one another. So long as evidence continues to confirm general statements about mental and other multiply realizable properties, then given that such statements make stronger claims than corresponding statements about disjunctions of physical kinds, we have reason to believe that mental and other multiply realizable properties are, unlike jade, real natural kinds that are distinct from the disjunctions of their realizers and capable of participating in laws. As Fodor (1997, pp.161-2) puts it:

> Science postulates the kinds that it needs in order to formulate the most powerful generalizations that its evidence will support. If you want to attack the kinds, you have to attack the generalizations. If you want to attack the generalizations, you have to attack the evidence that confirms them. If you want to attack the evidence that confirms them, you have to show that the predictions that the generalizations entail don't come true. If you want to show that the predictions that the generalizations entail don't come true, you have actually to *do the science*…So far…when the guys in the laboratories actually do the science, they keep finding that mental kinds are typically [multiply realizable], but that the predictions that intentional psychology entails are, all the same, quite frequently confirmed.

In sum, empirical data and standard axioms governing our use of induction suggest that the analogy Kim draws between mental properties and jade, and between multiply

realizable and merely disjunctive properties more generally, is a false one. The fact that merely disjunctive properties are unprojectible and thus unsuited to participate in laws hence gives us no reason to believe that the same holds true of mental or other multiply realizable properties.

2.i.b. Block's response

Block (1997) takes a markedly different approach in responding to Kim's objection than that pursued by Fodor (1997). Instead of attacking the analogy Kim draws between mental properties and jade, Block focuses his critical attention on an assumption common to both Kim and Fodor, viz. that the disjunctions of properties with which multiply realizable properties are nomically equivalent are "wildly disjunctive." In opposition to this idea, Block (1997, pp.120-1, underline added) argues that "special science kinds are typically <u>not</u> nomically [equivalent] with completely *heterogeneous* disjunctions of physico-chemical properties." There are two reasons why Block thinks this must be so. The first Block calls "the Disney Principle: that laws of nature impose constraints on ways of making something that satisfies a certain description." The second "is that there are *forces* at work [e.g. natural selection, learning, and conscious design] that can be expected to produce similarities." The homogenizing influence of these forces operating under the limitations that the laws of nature impose on the different possible ways of achieving the same result make it highly likely, Block argues, that there will be strong similarities in the way that products of these forces that are similar enough to be classified by the special sciences as belonging to the same kind are realized in different

individuals and species. The basic idea being that, at the same time as natural selection, e.g., leads to the development of physiological mechanisms that realize the same function in different species with similar needs and resources, the laws of nature ensure that there are only a certain limited number of physical structure types that are capable of realizing that function. As a result, the different ways in which that function is realized across different species or individuals are bound to be, in some important respects, *homogeneous*. As Block (1997, p.122) puts it:

> The power of natural selection [e.g.] to produce similarities in realizations derives from the fact that the forces [of natural selection, conscious design, etc.]…can only move in certain *channels*, the ones provided by the restrictions mentioned in the Disney Principle…Evolution, learning and the like impose similarities at the more superficial levels…The Disney Principle, by contrast, indicates similarities at all levels…So there are reasons to expect less than total heterogeneity at both the design and realization levels. Since evolution enforces similarity only at the design level, we should expect more variation at the levels of realization than at the design level. And this is why we expect multiple realization.

In other words, beneath the similarities captured by the taxonomic kinds of the special sciences, which can be attributed to the effects of the various homogenizing forces found in nature, there is another stratum of similarities to be expected among the realizers of such kinds, which can be explained by the fact that these homogenizing forces must operate under certain constraints imposed by the laws of nature, which limit the variety of ways in which they can bring about the same effect. Since the homogenizing forces that generate the similarities one finds among instances of the same higher level kind are largely indifferent to how such kinds are realized, and the laws of nature are not so strict as to allow only one possible avenue through which these higher level similarities can be produced, we may expect there

to be multiple ways in which a higher level kind can be realized. But since the laws of nature do at least impose some substantial restrictions on how something that exhibits the defining characteristics of a certain higher level kind can be made, the various ways in which such kinds can be realized are also likely to be far less numerous or heterogeneous than Kim and Fodor seem to suppose.

The preceding observations lead Block to distinguish among multiply realizable properties of the special sciences between those "that are the product of channeled selection, learning and design – in conjunction with the Disney Principle," (which he calls "'D properties'"), and "those that are due to *peculiarities of the realizations*," (which he calls "'realization properties'").[87] This distinction seems to me best interpreted as a distinction between features or properties *of* properties. Thus, for any multiply realizable property of psychology (e.g., pain), the "D properties" of that property will comprise those of its features that have played some role in its production and preservation through natural selection or some other homogenizing force, and which all its instances consequently exhibit due to the fact that the only realizers of that property that have likewise been preserved are those that suffice for those features under the laws of nature. The tendency of high levels of pain to cause behavior aimed at eliminating its source might hence be thought to qualify as a "D property" of pain, for the capacity to experience pain seems to have been selected for at least in part because of this tendency, and the only realizers of high levels of pain liable to have been preserved through

[87] Antony and Levine (1997, p.92) draw a parallel distinction between "realization specific" and (non-analytic) "realization independent" regularities.

selection are consequently those that (under the laws of nature) increase the likelihood that the pain they realize will lead to behavior aimed at eliminating its source.

The "realization properties" of pain, on the other hand, will include any features of that property that only some of its instances display due merely to certain idiosyncrasies in the manner in which those instances are realized. As an example of such a property, Block (1997, p.124) cites the phenomenon of "'aerodontalgia', [which] was discovered by U.S. Air Force dentists who noted that pilots in unpressurized planes of World War II (in which the sinus cavities expanded) reported pains that turned out to be related to previous dental work in which local anesthetic had not been used." As the tendency of past dental pains to be re-experienced when the sinus cavities are expanded is presumably "just a by-product" of the way pain is realized in humans (and perhaps more broadly in organisms belonging to the same or neighboring branches of our phylogenetic tree), the property of being re-experienced during expansion of the sinus cavities appears to be a "realization property" of pain, as there are likely other realizations of pain that do not exhibit this feature.

The upshot of all this is that while we should expect mental properties to be unprojectible with respect to their "realization properties," the same is not true of general statements expressing relations between mental properties and their "D properties." As Block (1997, p.125) puts it:

> Kim is right about realization properties. [Such properties] depend on the realization of psychological phenomena. The science of such psychological properties is not part of psychology. We wouldn't expect such properties to generalize to [minded] creatures that are not evolutionarily closely related to us. [On the other hand,] Kim is wrong about D properties. [Such properties are] common to creatures and

> perhaps machines that are not very similar in realization of [their respective mental states].

The reasoning here is fairly straightforward: Kim claims that mental properties are unprojectible because they are nomically equivalent with disjunctions of physical kinds that are themselves unprojectible because they lack the causal/nomological unity necessary for general empirical statements about them to be confirmable by their positive instances. Kim's argument hence relies on the assumption that the disjunctions of physical kinds with which multiply realizable properties are nomically equivalent are indeed so heterogeneous or "wildly disjunctive" that we have no reason to expect there to be any non-trivial, empirical features that all of their disjuncts share in common. This assumption, Block argues, is false, for since the laws of nature impose substantial constraints on the ways in which a multiply realizable property can be realized, there is bound to be a significant degree of causal/nomological unity among the various physical kinds that are actually capable of realizing a given mental property, making it reasonable to expect that each realizer will prove sufficient not only for the instantiation of the mental property it realizes, but also to endow the instances of the property it realizes with certain features that all other existing realizers of that property are likewise sufficient for. But if all the different realizers of a mental property suffice for these features, all the instances of that mental property must exhibit them, and general statements attributing such features to instances of mental properties hence *will* prove projectible after all. Kim is therefore wrong to think that because mental properties are multiply realizable and hence nomically equivalent with disjunctions of physical kinds, they must be unprojectible and therefore unfit to participate in laws.

Contrary to Kim, then, we should expect our evidence to confirm those general statements about mental properties that attribute to them their associated "D properties." Building upon Block's proposal, we might further suggest that for any empirically confirmed general statement attributing a certain property *F* to a mental property *M*, there are in fact three different reasons that could be given for *why* the evidence confirms it: (a) because *F* is essential to *M*, so that nothing could be an instance of the latter without exhibiting the former, (b) because under the actual laws of nature, the only things that are capable of realizing *M* are likewise sufficient to endow its instances with *F*, or (c) because *F* was responsible for *M*'s being preserved through natural selection, and hence only those realizers of *M* that suffice to endow its instances with *F* are likely to have been preserved as well.[88] In other words, a generalization attributing *F* to *M* may be projectible (a) because *M* has *F* by definition, (b) because all nomologically possible realizers of *M* necessarily confer *F* on the instances of *M* they realize, as a matter of natural law, or (c) because natural selection has "weeded out" all potential realizations of *M* that fail to confer *F* on the instances of *M* they realize, due to the fact that instances of *M* that have *F* are more fit than those that do not. In cases where *F* is a "D property" of *M*, the explanation for the projectibility of *M* relative to *F* will be of type (c). In other cases, one of the other two types of explanations may be more appropriate. Any lawlike regularity that mental properties enter into, however, seems explainable in the manner of (a), (b), or (c). And since each of the existing realizers of a mental property *M* is, for one of these

[88] This trichotomy is not endorsed by Block himself, but is instead offered as a plausible extension of his view. Note that if (a) is the reason why a general statement about a mental property is repeatedly confirmed, that statement will probably be analytic, and hence unlikely to qualify as a law.

three reasons, sufficient to endow the instances of *M* it realizes with whatever other properties we find *M* to lawfully correlate with, we should also expect the evidence we gather to confirm general statements correlating those same properties with the disjunction of *M*'s realizers.[89] Any nomic regularities that mental properties are found to enter into can thus be expected to lead us to corresponding regularities at the physical level as well.

One might wonder, though, whether Block's response to Kim doesn't in fact prove too much, for since the projectibility of general statements about mental properties is underwritten by the claim that the realizers of any mental property are likely to be homogeneous enough to ensure the projectibility of general statements linking their disjunction with those properties that are nomically correlated with the mental property they realize, Block's argument shows that mental properties are projectible only by way of establishing that the disjunctions of their realizers are projectible as well. But the reason the nomological irreducibility of mental properties was thought to follow from their multiple realizability was that there can be no "bridge laws" identifying mental properties with the disjunctions of their realizers, because the latter are not natural kinds, and are hence unsuited to figure into laws. However, if such disjunctions are in fact projectible, as Block suggests, then we seem to face the question why such disjunctions shouldn't also qualify as natural kinds with an equal right to participate in laws. If no

[89] In cases where (c) is the reason why a certain mental property *M* is projectible with respect to some other property *F*, while all the realizers of *M* that have been preserved through natural selection will be sufficient to confer *F* on the instances of *M* they realize, *M* may have other nomologically possible realizers that are *not*, and which have been eliminated through natural selection for precisely that reason. In such cases, the general statement correlating *F* with the disjunction of all the nomologically possible realizers of *M* will hence *not* be a law; only the statement correlating *F* (or some closely related property) with the disjunction of all the realizers of *M* that have been preserved through selection will be.

answer can be given, then a nomological reduction of mental properties to the disjunctions of their realizers seems possible after all.

Block's own solution to this problem is to relativize the notion of kindhood to that of projectibility, by treating properties or Boolean combinations of properties[90] as kinds only with respect to those properties that can be attributed to them in statements that are confirmable by their positive instances. On this proposal, properties and Boolean combinations of properties may, and typically will, qualify as kinds relative to some properties, and fail to qualify as kinds relative to others. Hence, Block (1997, pp.127-9) suggests that "if pain is nomically equivalent to a physico-chemical disjunction, then both pain and the disjunction will be kinds with respect to some properties [viz. "D properties], but to a lesser degree with respect to others [viz. "realization properties"]. *Kinds are relative and graded*." Block then claims that if we think of kinds in this way, any worries about the reducibility of mental kinds become groundless, for "[i]f talk of reduction presupposes a non-relative non-graded notion of a kind, then there is no matter of fact about reduction," since, on such a view, "kindhood comes in degrees."

I, however, don't see why one need presuppose a "non-relative non-graded notion of a kind," to talk in a "matter of fact" way about reduction, especially in cases where the putative reducing and reduced properties or Boolean combinations of properties are projectible, and hence, by Block's account, kinds, with respect to the same properties. Put

[90] Since, as Audi (2013, p.751) notes, Boolean operations are truth functions, and "properties are not truth-valued," properties cannot, strictly speaking, be disjoined, conjoined, or negated. Talk of disjunctions of properties, e.g., should therefore be taken to refer to types of states that a thing is in iff it has *F* or it has *G* or…"(for some *F*, *G*,…, such that *F*≠*G*, etc.)" Similarly, *mutatis mutandis*, for the other Boolean combinations.

simply, if two properties or Boolean combinations of properties qualify as kinds with respect to the same exact set of entities, then the existence of reductive bridge laws between the two would seem to be a real possibility, whether kindhood comes in degrees or not. Thus, if we accept Block's relativization of kindhood to projectibility, then so long as mental properties and the disjunctions of their realizers both qualify as kinds relative to the same properties, there seems to be no obstacle to reducing the former to the latter. Given Block's argument for the projectibility of mental properties, though, this correspondence between the sets of projectible statements in which a mental property and the disjunction of its realizers figure is precisely what we should expect; both should come out as projectible, and hence as kinds, relative only to those properties that are nomically correlated with the mental property concerned. So if we take Block's recommendation to relativize kinds to the properties they project to, it looks as though Block's argument for the projectibility of mental properties effectively reopens the door to a nomological reduction of mental properties to the disjunctions of their realizers.

There are, however, other options available. Since Block's argument for the projectibility of mental properties is separable from his proposal regarding the relativization of kindhood to projectibility, it is open to us to accept the former while rejecting the latter, thus availing ourselves of Block's response to Kim's objection to the multiple realization argument for the autonomy of psychology without thereby paving a new way for the reduction of psychology to neuroscience. There are, I think, good reasons for doing just this, for while projectibility is commonly taken to be a necessary condition for kindhood, Block's proposal involves treating it as also a sufficient, and indeed the *only* condition for being a natural kind; for him, to be a natural kind just is to

be projectible relative to some property. This seems much too permissive, as virtually any property or Boolean combination of properties is projectible relative to *some* property.[91] A much more natural view thus seems to be that there are other, additional constraints, besides projectibility, that a thing must satisfy in order to qualify as a real kind, and it seems reasonable to expect that some of these constraints will be such that Boolean combinations of properties will fail to satisfy them, even if they are nomically, metaphysically, or even logically equivalent with real kinds.[92] Such constraints might, e.g., be derivable from the standard Kripke-Putnam semantics for natural kind terms, as it is difficult to see how one could directly refer to disjunctions or other Boolean combinations of properties.[93]

Assuming that disjunctions and other Boolean combinations of properties can be excluded from kindhood on these or other reasonable grounds, then since only kinds can participate in laws, it follows that there can be no bridge laws identifying mental properties with the disjunctions of their realizers, and hence no nomological reduction of the former to the latter, even though, if Block is correct, both are projectible with respect

[91] As Block (pp.126-7) points out, even jade is projectible to some degree. Davidson (1995, p.275) makes the same point in reference to grue.

[92] The existence of such constraints would give us grounds for refusing to countenance "was observed today or earlier and is grue, or was not observed before tomorrow and is bleen" as a natural kind predicate, even though it is logically equivalent with the natural kind predicate "is green." This would also entail (*contra* Quine (1969, p.43) that kinds are intensional, and thus more akin to properties than sets.

[93] Due in large part to its importance for the multiple realization argument for the autonomy of psychology, a substantial literature has developed around the question of whether disjunctions of properties can themselves be properties or kinds. For a brief sampling, see Clapp (2001), Antony (2003), Walter (2006), and Audi (2013).

to the same set of properties.[94] With these conditions in place, we can make use of Block's account of the projectibility of mental properties without worrying that by doing so, we set them up for nomological reduction to their realizers. Accepted on these terms, Block's account serves as a useful complement to Fodor's response to Kim's objection, which offers no real explanation for the regularities that mental properties exhibit across distinct realizations, leaving this instead as a somewhat mysterious fact that happens to be supported by a large body of empirical evidence.[95] Taking Block's account on board enables us to fill this lacuna and thereby bolster Fodor's own response, which relies on the existence of such lawlike regularities to justify distinguishing mental properties from the disjunctions of their realizers.

In conclusion, then, it appears that Kim's objection to the multiple realization argument for the autonomy of psychology fails on two counts. First, Kim's contention that multiply realizable properties are unsuited to participate in laws because they are nomically equivalent with unprojectible disjunctions of kinds relies on the false assumption that all disjunctions of kinds are indeed unprojectible. As Block argues, the fact that multiple realizable properties are produced and preserved by certain homogenizing forces operating under the constraints imposed by the laws of nature gives us reason to expect that the various realizers of a multiply realizable property will all share certain causal/nomological features in common, and that the property they realize will hence be projectible with respect to those properties that the realizers themselves all

[94] One further consequence of this proposal is that while all laws are projectible, not all projectible generalizations qualify as laws.

[95] Fodor (1997, pp.161-2) openly concedes this.

regularly correlate with. Second, Kim's claim that multiply realizable properties are not natural kinds because they are nomically equivalent with disjunctions that fail to qualify as natural kinds relies on the false assumption that the discovery that a property is nomically equivalent with a disjunction of lower level properties always leads us to reject the notion that it is itself a natural kind suited to participate in laws. As Fodor argues, when empirical evidence points to the existence of certain regularities among a group of lower level properties that cannot be adequately accounted for in terms of the science that studies them, our inductive practices warrant the postulation of a higher level, nondisjunctive kind that is nomically equivalent with, but distinct from the disjunction of those properties, and which figures into laws that account for the regularities among them. In short, we must allow for the existence of multiply realizable kinds that are nomically equivalent with disjunctions of lower level properties because framing our laws in terms of such kinds enables us to capture certain generalizations that would otherwise be missed. It seems, therefore, that the suggestion that mental properties are multiply realizable and hence nomically equivalent with the disjunctions of their realizers in no way entails that such properties are unprojectible "pseudo-kinds," unsuited for use in scientific laws. The multiple realizability of mental properties thus supports, rather than endangers, the status of psychology as an autonomous science.

2.ii. Bickle's objection: Multiply realizable properties are not ipso facto nomologically irreducible

I now turn to an objection to the multiple realization argument for the nomological irreducibility of mental properties raised by Bickle (1998). The nub of Bickle's criticism is that the multiple realizability of a property poses an obstacle to its nomological reduction only under the constraints imposed on such reductions by the Nagelian model of intertheoretic reduction that Fodor's argument assumes. Since, however, these constraints are (by Fodor's own admission[96]) significantly stronger than those found in other models of intertheoretic reduction, and give rise, moreover, to a number of potential problems for Nagel's account, Bickle claims that we have reason to reject Nagel's model of intertheoretic reduction in favor of an alternative that allows for the nomological reduction of multiply realizable properties, and that Fodor's argument is consequently unsound.

The first problem that Bickle notes with Nagel's account[97] is its seeming inability to handle "textbook" cases of intertheoretic reduction wherein the reduced theory is empirically false. Such cases pose a problem for Nagelian models, under which a theory T is reducible to another theory T* iff T is deducible from T* (perhaps in conjunction with certain bridge laws), because a theory that is false cannot be deduced from a theory that is true. Hence, on Nagel's model, it is impossible for a theory that is false to be reduced to another theory unless the reducing theory is false as well. But there are many "textbook" cases of intertheoretic reduction (e.g. the reduction of Galilean to Newtonian

[96] See Fodor (1974, fn2): "The version of reductionism I shall be concerned with is a stronger one than many philosophers of science hold."

[97] These problems have also been noted by P.M. Churchland (1985, pp.9-10) and P.S. Churchland (1986, pp.280-1).

mechanics) involving the reduction of a false theory to a theory that is true, or was at least regarded as such when the reduction was achieved (Bickle, 1998, p.24; Feyerabend, 1962, pp.46-7).

One might, Bickle (1998, pp.24-5) notes, attempt to deal with this problem without completely abandoning the Nagelian model by suggesting either (a) that the reduction of false theories is accomplished by deducing them from the conjunction of the reducing theory and any needed bridge laws along with certain "possibly counterfactual limiting assumptions and boundary conditions," or (b) that the reduction of a false theory consists in the deduction of some *corrected version of it* from the reducing theory. Bickle argues, however, that these modifications to Nagel's model do not save it from another related problem, which is that "[s]ometimes a theory reducible to (some portion of) its successor turns out to be so radically false (in certain respects) that central elements of its ontology must be rejected as empirically uninstantiated." Nagel's account seems to have difficulty in handling such cases, for if the terms in the reduced theory that lack referents do not appear in the vocabulary of the reducing theory, then one will have to introduce certain bridge laws connecting them with terms in the reducing theory to enable the laws of the reduced theory to be deduced from those of the reducing theory. Since, however, bridge laws must, by most accounts, have the form of identity statements, or at least biconditionals, and nothing (except for nothing) is identical or biconditionally equivalent with something that doesn't exist, it will prove impossible to connect any of the terms of the reducing theory *via* bridge laws to those terms in the reduced theory that lack referents without thereby depriving the correlated terms in reducing theory of their referents as well. Therefore, as no theory that quantifies over empty terms is true, the

only theories, on Nagel's account, that one can reduce false theories containing empty terms to are theories that also contain empty terms, and are hence false as well. But there again seem to be clear "textbook" cases of reduction wherein a term in the reduced theory that does not have an extension is reduced to a term in the reducing theory that does; consider, e.g., the reduction of Newtonian to relativistic mass (Bickle, 1998, p.26; Feyerabend, 1962, p.80).

The third and, for our purposes, most important problem that Bickle claims Nagel's model of intertheoretic reduction faces concerns precisely that feature of the model that provides the basis for Fodor's argument for the autonomy of psychology; viz., that it does *not* allow for the nomological reduction of multiply realizable properties. According to Bickle, this feature of Nagel's account is by itself sufficient proof of its inadequacy, for the history of science, he claims, provides us with numerous examples of multiply realizable properties that have been successfully nomologically reduced. In support of this claim, Bickle points out that even temperature, the quintessential successfully reduced property, is multiply realizable both across distinct types of individuals and in the same individual at different times. Thus, while temperature in gases is reducible to mean molecular kinetic energy, the same is not true, e.g., of temperature in solids, plasma, or vacuums[98], and even in the same fixed volume of gas, the same mean molecular kinetic energy, and thus the same temperature, can be realized in many different ways by distinct maximally determinate microstates "in which the

[98] This point is adduced by Enç (1983, p.289), P.M. Churchland (1984, pp.41-2), and P.S. Churchland (1986, pp.356-8) in support of the local reduction strategy rejected in the previous chapter. (See also Sklar (1993, p.352).)

location and the momentum…of each constituent molecule [of the gas] are individually fixed") (Bickle, 1998, pp.121-2, 125). Far from posing an obstacle to intertheoretic reduction, it thus looks as though multiple realization is already present even in one of the most representative cases of intertheoretic reduction from the history of science; the reduction of thermodynamics to statistical mechanics. The fact that Nagel's model disallows the reduction of multiply realizable properties and the sciences that study them would hence appear to indicate not, as Fodor would have it, that the special sciences are autonomous, but rather that Nagel's model is flawed. At this point, Bickle's skepticism towards Fodor's argument from the multiple realizability of mental properties to their nomological irreducibility may begin to seem quite justified, for nothing much follows from the observation that the multiple realizability of a property poses an obstacle to its nomological reduction under a model of intertheoretic reduction that is itself problematic (especially if the model's inability to allow for the reduction of multiply realizable entities is one of the reasons why it *is* problematic).

In light of the difficulties facing the Nagelian model of intertheoretic reduction, Bickle advocates the adoption of an alternative model of intertheoretic reduction along the lines of that proposed by Clifford Hooker (1981a).[99] Hooker's basic idea is that rather than considering intertheoretic reductions as deductions of the reduced theory itself (or some modified version of it) from the reducing theory, we should view them instead as deductions of certain structures *within the reducing theory* that can then be seen to stand in a certain "analog relation" to the theory that is to be reduced, which relation entitles us

[99] See also P.M. Churchland (1985, pp.10-1) and P.S. Churchland (1986, pp.282-4).

to claim that the latter theory is reducible to the former. Hooker (1981a, p.49) describes this procedure as follows:

> Within T_B [the reducing theory] construct an analog, T_R*, of T_R [the reduced theory] under certain conditions C_R such that T_B and C_R entails T_R* and argue that the analog relation, A_R, between T_R and T_R* warrants claiming (some kind of) reduction relation, R, between T_R and T_B. Thus $((T_B \,\&\, C_R \rightarrow T_R^*) \cdot (T_R^* \; A_R \; T_R))$ warrants $(T_B \; R \; T_R)$.

Thus, while Hooker retains the Nagelian conception of intertheoretic reduction as involving some form of deduction from the reducing theory by making the relation between T_B and T_R* one of entailment, by also requiring that T_R* be constructed *within* the proprietary terminology of T_B, he effectively jettisons the feature of Nagel's account that gives rise to the problems noted above, which is its "treat[ment of] reduction not just as deduction but as deduction of a *structure specified within the vocabulary and framework of the reduced theory* – either [the reduced theory] itself or some corrected version [of it]" (Bickle, 1998, p.27).

Once this departure from the Nagelian model is made, the problems posed by reduced theories that are empirically false and/or contain empty terms no longer arise, for the only things that allow the falsity and reference failure that such theories involve to be carried over to the theories to which they are reduced are the bridge laws and deductive relations that the Nagelian account maintains must be in place between the two theories. In Hooker's model, however, there are no deductive relations or bridge laws between the reduced and reducing theories. The only relations of entailment are those that hold between the analog structure and the reducing theory within which it is constructed, and no bridge laws are needed to facilitate the deduction of the former from the latter, for since the former is part of the latter, it will not include any terms that the latter lacks.

Thus, on Hooker's account, any falsity or reference failure in the reduced theory is prevented from being transmitted to the reducing theory by the fact that there are no direct logical or nomological connections (e.g. entailments or bridge laws) between the two to enable such transmission. The only relation that the reduced theory enters into is the non-deductive, non-nomic analog relation, A_R, that obtains between it and the analog structure T_R*.

The key point for our purposes, however, is that by doing away with the need for bridge laws, Hooker's model not only avoids the difficulties that Nagel's has in accommodating reductions of false theories, it also eliminates the obstacle that multiple realization poses to reduction under Nagel's account. Multiply realizable properties again look to be nomologically irreducible under Nagel's model because there can, it seems, be no bridge laws linking them with their realizers (for laws only relate kinds, and the disjunction of the various realizers of a multiply realizable property is not a kind). However, if, as Hooker argues, reduced theories are not deduced from the theories to which they are reduced, so that even in cases where the reduced theory contains terms that the reducing theory does not, no bridge laws are needed to facilitate the reduction of the former to the latter, the suggestion that there can be no bridge laws linking the respective properties of psychology and neuroscience does not pose an obstacle to the reduction of psychology to neuroscience. Thus, in one fell swoop, Hooker's proposal seems to overcome the problems facing Nagelian accounts and invalidate the inference from multiple realizability to nomological irreducibility that such accounts have been thought to allow.

Rather than taking Hooker's account as he finds it, though, Bickle (1998, p.59) suggests that a better account of intertheoretic reduction can be obtained by incorporating Hooker's ideas into a "semantic" or "structuralist" conception of theories[100], which differs from the "syntactic" or "axiomatic" view derived from the Logical Positivist tradition in holding (a) that scientific theories are to be conceived not as deductive systems of sentences, "but instead in terms of their *models*[101]," and (b) that the structure of a theory consists not in logical relations between sentences, but rather in certain mathematic or set-theoretical relations among its models.

Developing Hooker's ideas in this direction enables Bickle to derive two important results. First, intertheoretic reductions can be reconstructed in such a way that it becomes possible to quantify the degree of correction that must be made to a reduced theory in the process of reducing it.[102] This is significant, because the degree of similarity between reduced and reducing theories, and the amount of correction that must consequently be made to the former in order to reduce it to the latter, has significant ontological consequences. In relatively "smooth" reductions, where only a few minor corrections need to be made to the reduced theory, the reduced and the reducing theories are similar enough that most if not all of the kinds of the former can be identified with

---

[100] Proponents of the structuralist view of theories include Suppe (1974, pp.221-30), Suppes (1960; 1967), and van Fraassen (1980, pp.53-6, 64-9; 1987).

[101] As used within structuralist accounts, the term "models" must be understood as referring to the various things that (potentially) *satisfy* or are *true interpretations of* a theory, *not* the theoretical representations of such things. (See Bickle (1998, pp.62, 82fn21) and Suppes (1960, p.289)). This use of the term "model" must hence be distinguished from the use I make of it when speaking, e.g., of Nagel or Hooker's models of intertheoretic reduction, or structuralist models of theory structure.

[102] The relevant quantitative measure is provided by the cardinality of certain relations (called "blurs") that shift the models of the reduced and reducing theories in such a way that the conditions for reduction are satisfied. See Bickle (1998, pp. 82-96).

kinds of the latter, and the ontology of the reduced theory is hence largely retained in that of the reducing theory. In particularly "bumpy" cases, however, where substantial changes must be made to the reduced theory[103], certain central portions of the ontology of the reduced theory will end up without any place in the reducing theory, and must therefore be eliminated, with the result that the terms in the reduced theory that were formerly taken to refer to these "entities" are thenceforth treated as empty. Given the weighty ontological repercussions that follow from the degree of similarity between reduced and reducing theories, the fact that Bickle's structuralist interpretation of Hooker's model gives us a way of measuring this factor with much greater precision than more traditional accounts of intertheoretic reduction would seem to be a major point in its favor.

Second, since according to the structuralist conception of theory structure, the relation of reduction between theories is not a deductive relation between sentences, but rather a set-theoretic relation between their respective models, which can be one-one, one-many, or many-one[104], Bickle's structuralist construal of Hooker's account of intertheoretic reduction makes explicit provision for cases of intertheoretic reduction wherein the reduced theory quantifies over properties that are multiply realizable. The

[103] In very "bumpy" cases, the old theory may be replaced by rather than reduced to the new one. The idea that there are cases of reduction that are neither "smooth" enough to warrant retention of the reduced theory's ontology nor "bumpy" enough to qualify as instances of theory replacement is actually a rather contentious assumption of Bickle's account, for one way of defending Nagel's model against the above criticisms is to hold that the model is not intended to allow for reductions of false theories or theories containing empty terms, for such theories can only be replaced, never reduced. (Hooker (1981a, p.45) is himself in favor of treating all non-retentive cases in this way.)

[104] See Bickle (1998, pp.65-82) for further details on the nature of this relation and the conditions it must meet; the important point for the present discussion is simply that it needn't be an injective function.

multiple realizability of temperature can thus be represented in Bickle's account of the reduction of thermodynamics to statistical mechanics by making the relation of reduction between the two theories a many-one set-theoretic relation between the models of statistical mechanics and those of thermodynamics, such that each model of thermodynamics involving a certain fixed volume of gas of a certain temperature gets mapped onto a number of distinct models of statistical mechanics, viz. all of those that involve any of the many microstates capable of realizing that temperature in such a volume of gas.

Given the various advantages that his model appears to have over Nagel's, one might naturally find it difficult to resist Bickle's call to abandon the traditional Nagelian model of intertheoretic reduction in favor of one such as his own that allows for the nomological reduction of multiply realizable properties. Indeed, the fact that temperature, the prototypical reduced property, is multiply realizable by itself seems sufficient to show that any model (such as Nagel's) that *disallows* the reduction of multiply realizable properties must *ipso facto* be deemed inadequate. Conceding this point requires, however, that we give up on the multiple realization argument for the autonomy of psychology, and accept that the multiple realizability of mental properties alone gives us no reason to think that such properties might not turn out, like temperature, to be nomologically reducible after all. Serious consideration would then have to be given to the possibility that by applying Bickle's model of intertheoretic reduction to the relation between psychology and neuroscience, we might find the former to be reducible to the latter in roughly the same way that thermodynamics is reducible to statistical mechanics.

At this point, however, one might object (as Bickle (1998, pp.125-6) anticipates) that the case of temperature actually fails to show that multiple realizability poses no obstacle to nomological reduction, for while temperature is indeed both multiply realizable and nomologically reducible, *it is not nomologically reducible to the types of states by which it is multiply realizable*. Thus, while temperature in gases is multiply realized by distinct microstates, it is not reduced to any one microstate or disjunction of microstates, but rather to a single "mathematical construct" out of them, viz. mean molecular kinetic energy. In this respect, however, temperature in gases *differs* from mental properties, for (so far as we now know) there is no single *neurobiological* property that any mental property is uniquely realized by and hence potentially reducible to that stands between a given mental property and its various neural realizers in the way that the statistical-mechanical property of mean molecular kinetic energy stands between a given temperature in gases and the various microstates that realize it.[105] Thus, whereas thermodynamics is reducible to statistical mechanics because there is a single statistical mechanical property that temperature in gases is identical with or uniquely realized by (even if temperature in gases is also *multiply* realizable by other types of statistical-mechanical states), the absence of any single type of neural realizer for each mental property precludes a parallel reduction of psychology to neuroscience.

In responding to this objection, Bickle seems to grant that temperature in gases, while multiply realizable, is not multiply realized by the property it is reduced to (viz.

[105] As neuroscience advances, we may at some point discover some such neurobiological property, but until we do so, the analogy between mental properties and temperature in gases remains an unsupported conjecture.

mean molecular kinetic energy). He argues, however, that the analogy between temperature and mental properties nonetheless still holds, because just as temperature in gases is uniquely realized by (or identical with) and nomologically reducible to a single type of statistical mechanical state, viz. mean molecular kinetic energy, which is in turn multiply realizable by a number of distinct microstates, mental states with representational content are uniquely realized by, and nomologically reducible to certain "neurofunctional" states, described in connectionist terms as "*partitions in activation-vector-spaces*," that intervene between them and their various neurobiological realizers (Bickle, 1998, p.136).[106]

Setting aside the many criticisms that have been raised against connectionist models of cognition[107], it seems to me that the crucial problem with this analogy between temperature in gases, microstates, and mean molecular kinetic energy, on the one hand,

---

[106] Activation-vector-spaces are geometric models of the "activation values" produced in the hidden units of a connectionist neural network that mediate between the network's input and output units *via* differently weighted causal connections. Within such models, each of these hidden units is assigned a different "axis (dimension)," whose values represent the level of activation "produced in [that] unit during processing" (Bickle, 1998, p.132). "[T]he entire set of activation values produced in every unit of the hidden layer during the processing of a given input" can then be represented by a point in the "activation-vector space" defined by these axes (Bickle, 1998, pp.132-3). As the weights of the various connections in the network are modified so that different types of inputs produce different levels of activation in the network's output units, the activation-vector space will become "partitioned" into different "subvolumes", each of which encompasses all of the various points in the vector space (i.e., all the different sets of activation values across the network's hidden units) that generate an output of a certain type. The connectionist hypothesis that Bickle endorses is thus that mental representations can be identified with such partitioned subvolumes in the activation vector space of a neural network, so that the property of being in a mental state with the representational content "apple", e.g., is identifiable with the set of different total levels of activation across the hidden units of a neural network (represented by a certain partitioned subvolume in the network's activation-vector space) that generate an output that enables selective behavior towards apples.

[107] Perhaps the most prominent of these is Fodor and Pylyshyn's (1988, p.3) argument that connectionists models do not seem to be able to adequately account for the "'systematicity' of mental representation: i.e.…[the fact] that the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents."

and intentional states, neurobiological states, and partitions in activation-vector spaces, on the other, is that while mean molecular kinetic energy is a physical property, (inasmuch as while it is indeed a mathematical/statistical construct, it is nonetheless one that is necessarily constructed out of a purely physical quantity (viz. molecular kinetic energy), wherefore nothing non-physical can have a mean molecular kinetic energy), partitions in activation-vector spaces *are not*. In Bickle's (1998, p.132, emphasis added) own words, "[a]n activation-vector space is nothing more than a *geometrical representation*." But since geometrical representations are indifferent to the non-quantitative features of the things they represent, neural or even physical states are not the only sorts of things that partitions in activation-vector spaces qualify as representations of, for such states are not the only kinds of structures that satisfy them. *Anything* with the right set of quantitative features will do. Consequently, while the primary intended empirical *application* of these geometrical representations may indeed be to neural states in human brains, such states cannot be *identical* with partitions in activation-vector spaces, for they are merely a small subset of the potentially limitless different kinds of states that can be represented in such a way. The same partition in an activation-vector space that is said to represent a certain collection of brain states might thus also represent an infinite number of different purely mathematical structures, or the response of different aspects of the U.S. economy to the Cuban Missile Crisis.[108]

[108] To this point, consider also that one of the most frequent ways in which activation-vector spaces have been used by connectionists is to provide models to guide the construction of representational systems *in computers*, where the states that realize the partitions in the activation-vector spaces are clearly not neural states (except perhaps in the same metaphorical sense in which connectionist models are themselves called "neural" nets).

Given, then, that partitions in activation-vector spaces can represent things that are neither neural nor physical in nature, there is no non-metaphorical sense in which these representations can themselves be described as physical or "neural." To do so would be to unduly conflate the representations themselves with the neural states they are typically (but not exclusively or necessarily) used to represent. It is hence misleading for Bickle (1998, p.137, emphasis added) to describe the level of description at which intentional psychology is, on his account, uniquely realized in humans as "*neuro*functional," for there is nothing intrinsically *neural* (or even *physical*) about the partitions in activation-vector spaces to which he claims mental properties with intentional content might be nomologically reduced.[109] To reduce mental properties to such partitions is thus *not* to reduce them to any neural or physical state, but at most to a certain type of geometrically characterized functional state, which differs from the sorts of states traditionally invoked by functionalists only in having the structure of a connectionist network, as opposed to that of a Turing machine. As a result, the position Bickle argues for ends up amounting to nothing more than a connectionist variant of the familiar token-physicalist position (to be challenged in the following chapter) that identifies mental properties with functional properties that are realizable by but nomologically irreducible to various distinct types of physical states. The only way Bickle's proposal differs from this standard model is in holding that the functional states that mental properties are to be identified with are best represented not as computations

[109] Certain remarks of P.S. Churchland (1986, p.382) are, I think, open to this same criticism.

over combinatorially structured symbols, but instead as networks of input, output, and intermediary "hidden" units linked by variously weighted causal connections.

The account Bickle offers thus leaves us with the following important disanalogy between the relation of thermodynamics to statistical mechanics and that of psychology to neuroscience. Whereas temperature in gases is multiply realizable at the statistical-mechanical level by distinct microstates, but also uniquely realized by or identical with a property that is explicitly constructed out of such statistical-mechanical kinds, the connectionist functional states that Bickle proposes we identify mental states with are multiply realizable by physical, neurobiological kinds without also being constructed out of, uniquely realized by, or identical with any such kinds. Due to this difference, Bickle's favored model of intertheoretic reduction cannot allow for a reduction of psychology to neuroscience as it does the reduction of thermodynamics to statistical mechanics, for one of the conditions that Bickle (1998, p.81) lays down on the reduction relation that on his account must hold between reduced theories and the theories to which they are reduced is that this relation must be an "*ontologically reductive link*"; i.e., it must be "constructed out of local links obtaining between all of the empirical base sets of the reduced theory [i.e., the kinds of things that the theory is ontologically committed to] and elements of the potential models of the reducing theory in a way that respects how the two theories each carve up the world."[110] Since the functionally individuated kinds of cognitive psychology

[110] This condition is necessitated by Schaffner's (1967, p.145) observation that the standard semantic/structuralist definition of the intertheoretic reduction relation as a set-theoretic isomorphism between the models of the reduced and reducing theories is "too weak to be adequate," because "[d]ifferent and nonreducible (at least to one another) physical theories can have the same formal structure – e.g., the theory of heat and hydrodynamics – and yet one would not wish to claim that any reduction could be constructed here." To deal with this problem, structuralists must add to their definition the further

are multiply realizable by neurobiological kinds without also being explicitly constructed out of them, it is difficult to see how there could be any such link between psychology and neuroscience, for given the multiple realizability of the kinds of the former by those of the latter, the taxonomies of the two sciences will end up cross-classifying one another, and the "empirical base sets" of psychology will consequently fail to correlate with neuroscientific kinds "in a way that respects how the two theories each carve up the world." (Contrast this with the reduction of thermodynamics to statistical mechanics, where the requisite "ontological reductive link" *can* be secured between the two theories, because the various microstates related to each specific temperature in a gas will all fall under a single statistical mechanical kind of having a certain mean molecular kinetic energy, thereby enabling temperatures in gases to be related to microstates "in a way that respects how the two theories each carve up the world.") In sum, since there seems to be no single physical or neurobiological kind that can be constructed out of the various neurobiological states that realize a given mental property in the way that a single mean molecular kinetic energy can be constructed out of the various microstates by which a given temperature in gases is multiply realized, it appears that mental properties cannot be nomologically reduced to any single physical or neurobiological kind in the way that temperature in gases is reducible to mean molecular kinetic energy.

In addition to these problems with Bickle's attempt to establish the reducibility of psychology to neurobiology by way of analogy with the reduction of thermodynamics to statistical mechanics, Ronald Endicott (1998) also points out that many of the advantages

---

requirement that there be "appropriate local links between the components of" the models of theories that are reductively related (Bickle. 1998, pp.74-82).

that Bickle claims for his own model of intertheoretic reduction over Nagel's classical account actually fail to hold up under scrutiny. Despite the fact that it does not depend, like Nagel's, on any deductive relations or bridge laws between reduced and reducing theories, Bickle's model still ends up encountering some of the same obstacles to the reduction of false theories and those that deal in multiply realizable kinds that Nagel's seems to face.

The first of these obstacles concerns the Hooker-inspired analog structure, T_R*, which under Bickle's model is deduced from the reducing theory in conjunction with certain counterfactual conditions that enable it to "mimic" the theory that is targeted for reduction. Recall that since this analog structure is introduced for the sole purpose of preventing any falsity or reference failure in the reduced theory from carrying over to reducing theory by eliminating the direct deductive and nomological connections between the two that are posited by Nagelian accounts, it is essential that T_R* be constructed using solely the vocabulary and conceptual resources of the reducing theory. Any input from the reduced theory in the construction of the analog structure is absolutely prohibited.

This restriction conflicts, however, with the fact (which is granted and even emphasized by Bickle (1998, pp.148-50) and others, e.g. Hooker (1981a, pp.48-51) and Patricia Churchland (1986, pp.284-5), who endorse similar models of intertheoretic reduction) that prior to their reduction, reduced theories often *coevolve* with the theories they ultimately reduce to. As a result of such coevolution, the theoretical vocabulary, conceptual resources, and explanatory models of a reducing theory have oftentimes been shaped by those of the theory that reduces to it (and *vice versa*). From this it follows,

however, that in such cases, the analog structure deduced from the reducing theory will be "infected" by the concepts and terminology of the reduced theory. As Endicott (1998, p.65) puts it:

> The new-wave constraint on theory construction stipulates that the basic T_B and *not* the original T_R must supply the conceptual resources for constructing the corrected image T_R*. Yet this seems flatly contradicted by the fact that, once co-evolution has run its natural course, T_R* has become a mutual product of T_B and T_R. How, then, is T_R* specified 'within the idiom of T_B' in any meaningful sense that *excludes* T_R?

If, then, the contribution of the reduced theory to the laws, concepts, and terminology of the analog structure cannot be discounted, it seems that the analog structure, and indeed the reducing theory itself may sometimes end up inheriting a portion of any falsity or reference failure that the laws, concepts, and terms of the reduced theory involve. Bickle's model hence appears to face the same problem as Nagel's in accounting for the reduction of theories that are false.

Endicott's second criticism of Bickle's model of intertheoretic reduction calls into question its alleged ability to dispense with Nagelian bridge laws and the ensuing difficulties they create for the nomological reduction of multiply realizable kinds. As Endicott points out, while Bickle's account makes no explicit appeal to any statements of identity or nomological covariance between the kinds of the reducing and reduced theories, such bridge laws are nonetheless a necessary concomitant of those "smooth" cases of reduction that fall on the ontologically retentive end of the reductive spectrum, for reductions in such cases are taken as warranting the cross-theoretical identification of the theories' respective kinds. Consequently, while one may not, on Bickle's account, need bridge laws to carry out a reduction, they are nonetheless part of the consequences

of any reductions that are smooth enough to justify retention of the reduced theory's ontology. But if bridge laws number among the consequences of smooth reductions, then any reasons we might have for thinking that there can be no such laws linking the kinds of two theories are, by *modus tollens*, thereby also reasons for thinking that neither theory can be smoothly reduced to the other. Given, then, that there *are* compelling reasons (which Bickle (1998, pp.4-5) himself accepts) for thinking that there can be no bridge laws between multiply realizable kinds and their various realizers, it follows that theories that deal in multiply realizable properties cannot, on Bickle's model, be smoothly reduced. Endicott (1998, p.69) summarizes the argument as follows:

> (i) If a case falls at the retentive end of the [reductive] continuum, then crosstheoretic property identities exist between reduced and reducing theories.
> (ii) If crosstheoretic property identities exist between reduced and reducing theories, then biconditional bridge laws exist between reduced and reducing theories.
> (iii) Therefore, if a case falls at the retentive end of the [reductive] continuum, then biconditional bridge laws exist between reduced and reducing theories…
> (iv) It is *not* the case that biconditional bridge laws exist between intentional psychology…and more basic physical theories.
> (v) Therefore, it is *not* the case that intentional psychology…will fall at the retentive end of the [reductive] continuum.

If, however, intentional psychology doesn't fall on the retentive end of the reductive continuum, then there are only two other places it can go: either it will fall on the "bumpy" end of the spectrum, among those theories whose ontologies are outright eliminated, or it will prove to be a truly autonomous, irreducible science that does not fall on the reductive spectrum at all. As a result, Bickle's model of intertheoretic reduction seems to leave us with the same choice regarding the status of psychology as the Nagelian model he claims to supersede: autonomy or elimination. In the absence,

therefore, of any convincing argument for the view that current psychology is radically false, the most reasonable conclusion to be drawn under Bickle's model would appear to be that psychology cannot be reduced to any physical science.

Summing up, the preceding analysis of Kim and Bickle's objections to the multiple realization argument for the nomological irreducibility of mental properties has found both their criticisms unsuccessful. Whereas Kim's relies on a false analogy between mental properties and jade, and the unwarranted assumption that the potential realizers of mental properties are too heterogeneous to be capable of realizing a single property that has the requisite causal/nomological unity to participate in laws, Bickle's relies on a false analogy between mental properties and temperature, and the mistaken assumption that the multiple realizability of mental properties poses an obstacle to their nomological reduction only under an outdated Nagelian model of intertheoretic reduction. As these objections nonetheless number among the most prominent and forceful that have been raised against the argument from the multiple realizability of mental properties to their nomological irreducibility, and incorporate, in various ways, many of the other criticisms that have been leveled against it, it seems reasonable to take their failure as representative, and conclude that the multiple realizability of mental properties indeed suffices for their nomological irreducibility. If this conclusion is correct, and the multiple realizability of mental properties does entail both that there is no type-identity between mental and physical properties, and that the former cannot be reduced to the latter by way of a more general reduction of psychology to neuroscience, the only options left for the physicalist, aside from abandoning his/her position, are to

either accept some form of token/non-reductive physicalism, or else embrace eliminative materialism. The rejection of these two remaining forms of physicalism will be the task of the following chapter.

CHAPTER 6

THE IMMATERIALITY OF MENTAL EVENTS

The previous two chapters have made the case that due to their multiple realizability, mental properties cannot be identified with or reduced to physical properties, and that type-physicalism (i.e., the view that each type of mental event is identical with some type of physical event) is therefore false. This conclusion, however, does not yet eliminate the possibility of a physicalist solution to the Exclusion Problem, for it is consistent with the alternative (and currently more widely held) token/non-reductive form of physicalism, which differs from type-physicalism in holding that mental properties are non-identical with and irreducible to physical properties, even though each token instance of any mental property is nevertheless identical with some physical event.[111] Should this position prove viable, then it may seem that even if mental properties cannot be identified with or reduced to physical properties, the most promising response to the Exclusion Problem is still to follow the physicalist in rejecting (4*) Mind-Body Dualism and maintaining that any mental cause of a physical effect must itself be a physical event. The present chapter seeks to challenge the adequacy of any such response to the Exclusion Problem by arguing that token/non-reductive physicalism is likely false as well.

[111] I remind the reader that I'm assuming a property exemplification view of events along the lines of that proposed by Kim (1973, p.222), according to which an event is an instantiation of a property by an object at a time.

The chapter is divided into four sections. Section 1 questions whether token/non-reductive physicalism is even a coherent position. Section 2 addresses Donald Davidson's (1970) argument for the view. Section 3 notes various problems involved in any attempt to identify instances of phenomenal properties with physical events. Section 4 points out some parallel problems facing the identification of instances of intentional properties with physical events. I then conclude the chapter by combining the results of these sections with those of the previous two chapters to derive a general argument for the view that a plausible solution to the Exclusion Problem should not require us to reject (4*).

1. Is token/non-reductive physicalism a coherent position?

Given the ascendance of physicalism in contemporary Anglophone philosophy, and the widespread acceptance of Fodor's arguments for the nomological irreducibility of mental properties, it should come as no surprise that token/non-reductive physicalism (T/NRPism) has for some time been the orthodox view in philosophy of mind. The position indeed has considerable appeal, as it claims to provide a simple monistic ontology that avoids postulation of mysterious immaterial substances or events, while also paying deference to the popular aversion to reductionism and the natural intuition that the mind is a special sort of entity that is in some sense different from the physical events by which it is realized. Any position that claims such striking advantages is, however, bound to raise some suspicion as to whether it is in fact capable of delivering on its promises, and these doubts are exacerbated in the present case by the apparent tension between the two ideas that T/NRPism purports to reconcile. Put simply, it is

difficult to see how the mind could be "nothing over and above" a certain collection of interrelated physical states and processes (as any physicalist is obliged to maintain), while being at the same time irreducible to the physical states it depends on, as T/NRPism avers. This difficulty has given rise to the charge that T/NRPism is actually incoherent. This objection can be presented in two different ways, which I'll now discuss in turn.

The most immediate difficulty for T/NRPism is that its claim that mental properties are irreducible to physical properties even though every token instance of any mental property is identical with an instance of some physical property is inconsistent with one of the main conceptions of events; viz. the property exemplification view developed by Jaegwon Kim (1973, p.222), according to which an event "consists in the instantiation of a property *P* by an object *x* at a time *t*." Since, on this view, an event just is the instantiation of a certain property by an object at a time, to say that two events are identical is to say that they are instantiations of one and the same property. Hence, on this view of events, the T/NRPists' claim that every mental property instance is identical with an instance of some physical property would require them to abandon their claim that mental properties are irreducible to physical properties. For if an event consisting in the instantiation of a mental property *M* is identical to an event consisting in the instantiation of a physical property *P*, and to say that two events are identical is to say that they are instantiations of the same property by the same object at the same time, it follows that *M*=*P*.[112]

[112] Bickle (1998, pp.12-3) makes this point nicely.

The obvious response to this objection would be for T/NRPists to reject Kim's theory of events in favor of some alternative view that allows for one and the same event to consist in the instantiation of two distinct properties (by the same object at the same time). The most suitable candidate for this role would likely be Davidson's (1969, p.179) "coarse-grained" theory of events, according to which events are particulars that are "identical if and only if they have exactly the same causes and effects." While it is somewhat unclear to what extent Davidson's criteria of event individuation actually conflict with those proposed by Kim,[113] Davidson (1969, pp.170-1) nevertheless distinguishes his conception of events from Kim's by insisting that on his view, the event e.g. of Brutus' stabbing Caesar is the same as the event of Brutus' killing Caesar, even though stabbing and killing are distinct properties or relations. This contrasts with Kim's account, according to which Brutus' stabbing Caesar and his killing him qualify as different (albeit closely related) events, since each consists in the exemplification of a different property or relation. By adopting a Davidsonian conception of events, T/NRPists might thus maintain that the event that makes it true to say that Caesar is in pain is one and the same as the event that makes it true to say that a certain type of neural activity *N* is currently taking place in Caesar's brain, even though the property of being in pain is distinct from and irreducible to that of being in *N*. The only problem with this maneuver is that a number of reasons can be given for favoring Kim's more "fine-grained" conception of events to the model proposed by Davidson, some of which appeal to Davidson's own causal criterion of event identity to argue that events such as Brutus'

[113] Kim (1976, p.164) suggests that their respective criteria may in fact be coextensive.

stabbing Caesar and Brutus' killing Caesar cannot be identical, precisely *because* they can cause and be caused by different things.[114] TN/RPists will hence have to come up with a compelling response to these arguments if they wish to ensure the coherence of their position.

The second challenge to the coherence of T/NRPism touches directly on the topic of mental causation and the Exclusion Problem. To understand this objection, which has been pressed most persistently by Kim (1989; 1992a; 1993c), we can start by noting that for any position that warrants the title of physicalism, the Exclusion Problem should pose no problem at all, for if the mind is physical (as any non-eliminativist physicalist must maintain), then mental causation of physical effects is perfectly consistent with (2) the Causal Self-Sufficiency of the Physical and (3) the Absence of Systematic Overdetermination. T/NRPism claims to satisfy this constraint on any physicalist position by maintaining that while mental and physical properties are distinct, mental events are nevertheless physical. On the assumption that any mental causes of physical effects are mental events, the physical nature of those events is thought by T/NRPists to be enough to ensure the consistency of mental causation with (2) and (3). The problem, however, is that while any mental event *e* is, according to T/NRPism, an instance both of some mental property *M* and of some physical property *P*, given (2) (which is a principle that any physicalist seems obliged to accept), any physical effects that *e* produces should be able to be fully accounted for by appeal to the fact that *e* is an instance of *P*. The fact that *e* is also an instance of *M* is consequently irrelevant (or at least superfluous) to any causal

[114] See Goldman (1970, pp. 1-10).

impact that *e* has on the physical world. In this way, the commitments of T/NRPism appear to lead straight to epiphenomenalism, for given those commitments (which include (2)), the mental properties instantiated in any mental event do not seem to make any additional contribution to its causal powers. It is instead only by virtue of its physical properties that any such event causes anything physical.

If this is correct then it looks as though T/NRPism does not avoid the Exclusion Problem after all. For despite the T/NRPist's identification of mental with physical events, the Problem simply resurfaces at the level of the distinct mental and physical properties that the T/NRPist claims these events instantiate. Since the physical properties instantiated in any mental event seem capable of doing all the causal work that the event performs, there appears to be nothing left for the mental properties instantiated in that event to do. It hence becomes difficult to avoid the conclusion that, according to T/NRPism, events *qua* mental don't do anything at all. Given, then, that any physicalist position should have no difficulty in addressing the Exclusion Problem without being forced to accept epiphenomenalism, the fact that T/NRPism faces these difficulties raises the question whether it can truly claim to be a form of physicalism.

Once again, the problem can be traced to the tension between the non-reductivist and physicalist components of T/NRPism. For in light of the difficulties just mentioned, it seems that the only way for the T/NRPist to ensure the causal efficacy of the mental properties instantiated in mental events is to either (a) identify them with the physical properties that those events are also said to instantiate, or (b) retain the distinction between mental and physical properties, but claim that some physical effects of mental events cannot be accounted for in terms of the physical properties they involve, which is

to reject (2). Option (a) is consistent with physicalism, but it requires abandoning the non-reductivist component of T/NRPism. Option (b), on the other hand, enables the T/NRPist to remain true to their non-reductivist commitments, but it requires them to deny (2), which is a principle that any physicalist must accept. It thus appears that in order to avoid epiphenomenalism, T/NRPists must give up on one of the two core tenets of their view. One can be a non-reductivist, or one can be a physicalist, but unless one is willing to accept epiphenomenalism[115], one cannot be both. Given this result, one naturally wonders whether T/NRPism was ever a stable position to begin with.

2. Davidson's (1970) Anomalous Monism

Granting, for the sake of argument at least, that T/NRPism is not inherently unstable or self-contradictory, what positive reasons might we have to accept it? Standard arguments for T/NRPism come in two different forms. One proceeds from certain general metaphysical premises that together entail the distinct nature of mental and physical properties and the identity of mental and physical events. The other proceeds from the assumptions that an exhaustive functional analysis can be given for any mental property, and that the functional description in terms of which a mental property is analyzed will be satisfied on each occasion of its instantiation by some purely physical event. The remaining three sections of the present chapter attempt to respond to these arguments by

[115] Some arguments against epiphenomenalism will be offered in the following chapter. For now, I will proceed on the assumption that the view is sufficiently counter-intuitive that it should be avoided if at all possible.

showing (a) that certain of the metaphysical premises that supposedly lead to T/NRPism are false, or at least disputable, and (b) that mental properties either do not admit of an exhaustive functional analysis, or else the functions that define them cannot be performed by any purely physical entity. The present section addresses what is undoubtedly the most famous of the first type of argument for T/NRPism; viz. Davidson's (1970) argument for Anomalous Monism.

Davidson's (1970, pp.80-1, 99-100) argument is impressive in its concision and simplicity. It consists of only three premises, and can be stated as follows:

(i) "Principle of Causal Interaction": "at least some mental events interact causally with physical events."

(ii) "Principle of the Nomological Character of Causality": "events related as cause and effect fall under strict deterministic laws."

(iii) "Anomalism of the Mental": "there are no strict deterministic laws on the basis of which mental events can be predicted and explained."

(iv) Therefore, "every mental event that is causally related to a physical event is a physical event."

In light of the preceding chapters, it should be clear that premise (iii), the Anomalism of the Mental, entails that mental properties are distinct from and irreducible to physical properties, for if there are no strict laws "on the basis of which mental events can be predicted and explained," then clearly there can be no strict type-type correlations between mental and physical events, for such correlations would enable the very sort of prediction and explanation that (iii) rules out. But as argued in the previous chapter, without such "bridge laws" to establish the requisite connections between types of mental

and physical events, mental properties cannot be reduced to or identified with physical properties. To accept premise (iii) of Davidson's argument is hence to accept the non-reductivist component of T/NRPism. The conclusion of the argument then simply asserts the remaining, physicalist component of T/NRPism, which is the identity of mental and physical events. Hence, if Davidson's argument is sound, then T/NRPism is true.

Assuming, as seems plausible, that the only events capable of being related as causes to physical effects by way of strict, deterministic laws are themselves physical, Davidson's argument is clearly valid. So if T/NRPism is false, at least one of the argument's premises must be false as well. For this reason, one of the more common criticisms of Davidson's argument is, I think, ultimately insufficient to refute it. This criticism is that, for reasons similar to those noted in the previous section, Davidson's argument appears to lead to epiphenomenalism, for if causation requires subsumption under strict laws (premise (ii)), and there are no strict laws relating mental and physical properties (premise (iii)), then it seems that mental events cannot cause physical effects by virtue of their mental properties, but instead only *qua* physical events.[116] While this does not entail that premise (i) of the argument is false (since mental events can still "interact causally with physical events" by virtue of being identical with physical events that do so), it does mean that mental *properties*, at least, are causally inert. Given, however, the validity of Davidson's argument, those who accept it may reply that such epiphenomenalism is simply a surprising consequence that is forced upon us by the truth of Davidson's premises. Hence, while Davidson (1993) did attempt to respond to the

[116] This criticism has been raised, e.g., by Honderich (1982, pp.62-4), Kim (1984, p.267), and Sosa (1984, pp.277-8).

charge of epiphenomenalism by claiming that the supervenience of mental properties on physical properties is enough to secure their relevance to the causation of physical effects, even if his response proves inadequate (which many think it does[117]), the die-hard Anomalous Monist might instead simply grant the charge and claim that the causal inefficacy of mental properties is just something we must learn to accept. The drawback of the present objection is thus that while the epiphenomenalist consequences of Davidson's argument give us reason to suspect that something in it is amiss, they fail to specify precisely where the error lies. Rather than noting the implausibility of some of its consequences, a more conclusive objection to the argument should provide a direct attack on one of its premises.

Given the *prima facie* plausibility of premise (i), it should, I take it, be fairly uncontroversial to suggest that it is premises (ii) and (iii) of Davidson's argument that are most vulnerable to criticism. While I think both of these premises are false, my objections to (iii) are not such as to raise any serious trouble for the argument. Here it should suffice to recall the claim advanced in the preceding chapter that while the multiple realizability of any mental property M by certain distinct physical properties P_1, P_2,… means that there can be no "bridge laws" of the form $(\forall x)(Mx \equiv (P_1x \vee P_2x \vee \ldots))$, (since laws are relations between natural kinds, and there are no disjunctive kinds), each individual realizer of M may nevertheless be nomologically sufficient for it in such a way that $(\forall x)(P_1x \supset Mx)$, $(\forall x)(P_2x \supset Mx)$, etc. all qualify as laws of nature. As there is no reason to think that such laws could not be strictly deterministic, their existence would entail that

[117] See Kim (1989; 1993a), McLaughlin (1993), and Sosa (1993).

(ii) is false. This, however, does not threaten the basic tenor of Davidson's argument, for his conclusion still follows if (i) is interpreted as affirming the causation of physical effects by mental causes, and (iii) is replaced by the following, weaker premise:

> (iii*) there are no strict deterministic laws on the basis of which physical events can be predicted and explained in terms of mental events.

This premise is not falsified by the existence of strict laws of the form $(\forall x)(Px \supset Mx)$, for such laws would only be useful in explaining and predicting the occurrence of mental events in terms of physical events, not *vice versa*. More importantly, seeing as any laws relating mental causes to their physical effects will undoubtedly be *ceteris paribus*, (iii*) also seems likely to be true. Given, then, that Davidson's argument still goes through on this weaker premise, the potential falsity of (iii) does not pose a major problem for his reasoning.

The case is different, however, with regard to premise (ii). For clearly, if causes and effects needn't be related by strict, deterministic laws, then the fact that only physical events seem capable of standing in such strict nomological relations to physical consequences would give us no reason to think that mental events must be physical in order to produce physical effects. Thus, if (ii) is false, then Davidson's argument is wrecked beyond repair.

The conception of causation contained in (ii) is an instance of a more general Regularist theory of causation that goes back to Hume. Such theories hold that for one thing to cause another is for both to be tokens of types whose correlation is entailed by some nomic regularity, such that (given certain background conditions) things of the former type are invariably followed by things of the latter type. Humean Regularist

accounts of causation have, however, long been known to face a number of serious objections, all of which apply equally to (ii). The first of these concerns the seeming inability of such accounts to provide a satisfactory explanation for the ease and readiness with which people recognize causes and effects. When a person recognizes the striking of a match as having caused it to light, their judgment certainly does not seem to be predicated on the assumption that these events are tokens of types that are nomically correlated in such a way that (*ceteris paribus*) tokens of the former are invariably followed by tokens of the latter.[118] The young age at which children become capable of identifying causes and effects tells strongly against the idea that causal judgments rest on such theoretical presuppositions. The fact that judgements about causation are genetically prior to speculations about nomic regularities would seem, moreover, to suggest that we come by our knowledge of the latter only by way of picking up on similarities between situations that we have *already* identified as instances of causation. If this is so, then rather than causation being dependent upon law-like regularities, laws may instead be mere idealizations of patterns that are observed among situations that we recognize *as* instances of causation *prior* to our recognition of any regularities among them.

Secondly, as David Fair (1979, p.225) points out, if Regularism were true, then "we could only come to know that *A* causes *B* by recognizing *A* and *B* as an instance of some general regularity...[But] [w]e recognize *unique* occurrences as causally connected even if they *defy* regularities in our experience." To demonstrate this point, Fair asks us to imagine a person who had never seen glass struck by something hard enough to shatter

[118] See Searle (1983, pp.118-21).

it, and had many times seen something they mistook for glass (e.g. clear plastic) fail to break upon being violently struck. The regularities in the experience of such a person might lead them to believe that glass is unbreakable. Yet the first time they should happen to see a piece of real glass shatter upon being struck, say, by a baseball, it seems likely that they would have no trouble at all in identifying the baseball's striking the glass as the cause of its shattering. If this is so, then it would appear that our concept of causation, as well as our ability to correctly apply it, does not depend in any way on our grasp of nomic regularities.

In the Regularist's defense, one might note that the objections raised thus far bear solely on our knowledge of causation, or our ability to recognize and identify causes and effects. But the Regularist is (one might argue) not interested in such epistemological issues. Their aim is instead to describe what causation actually is. In response, it could be said that surely any acceptable account of what causation really is should not be such as to entail that the vast majority of our causal judgments are unjustified, and it is hard to see how Regularism can avoid this result if the objections just raised are valid. But no matter. Even setting these objections aside, Regularism still faces a number of counterexamples that are entirely non-epistemic.

The first type of counterexample shows that Regularism is too permissive, since there are things that satisfy the conditions of the Regularist analysis of causation, but which nevertheless do not seem to be related as cause and effect. Given, e.g., that drops in atmospheric pressure are nomically correlated both with drops in the mercury level of nearby barometers and the occurrence of storms, there is also a regular correlation between changes in barometer readings and the subsequent occurrence of storms. A strict

Regularist account would hence seem to force us to treat changes in barometer readings as causes of storms (and *vice versa*, unless a further stipulation is made requiring causes to be temporally prior to their effects). But this is false. Changes in barometer readings don't cause storms (nor *vice versa*). Rather, both are effects of a common cause; viz., drops in atmospheric pressure.[119]

A second type of counterexample shows that Regularism is also too strict, as there are certain seemingly straightforward instances of causation that Regularism fails to treat as such. This problem is particularly noticeable for Regularists, such as Davidson (1970, p.81), who require "events related as cause and effect [to] fall under strict deterministic laws," for the seemingly non-deterministic nature of quantum level events does not appear to prevent them from producing effects.[120] Consider the following case, also taken from Fair (1979, p.223): "Suppose I trigger a bomb by placing radium near a Geiger counter connected to the bomb. Am I any less the cause of the explosion because one link in the causal chain, the radium decay, cannot be described deterministically?" The fact that Davidson's premise (ii) demands a positive answer to this question is by itself reason enough to view it with suspicion.

In light of the various objections and counterexamples that have now been raised to the Humean Regularist theory of causation, it seems reasonable to conclude that premise (ii) of Davidson's argument is probably false. This conclusion seems all the more warranted, considering the particular stringency of the form of Regularism that (ii)

---

[119] See also Reid (1788/2010, IV.9) and Lewis (1973a, pp.556-7).

[120] A compelling defense of indeterministic causation can be found in Mellor (1995, ch.5).

endorses. Without premise (ii), though, the entire argument for Anomalous Monism collapses. If there is to be a sound argument for T/NRPism, it will hence have to come from some other source.

3. Qualia

With the collapse of Davidson's argument for Anomalous Monism, the most promising attempt to derive T/NRPism from general metaphysical principles has proven unsound.[121] This suggests that a different approach is needed. As noted above, the primary alternative method of arguing for T/NRPism is to first maintain that mental properties can all be exhaustively analyzed in strictly functional terms[122], and then claim that as a matter of empirical fact, while the individuative function of any mental property might be performed by distinct types of physical events, each time a mental property is instantiated, the event that performs its individuative function is always purely physical

[121] Another general argument for T/NRPism might be offered on the grounds that even if the multiple realizability of mental properties precludes their type-identification with physical properties, instances of mental properties should still be token-identified with physical events because the resulting view is ontologically simpler and thus preferable to the view that mental properties and events are dependent upon, but distinct from physical properties and events. (See Smart (1959, pp.142-3) and Block and Stalnaker (1999, pp.23-5).) Against this proposal, the remainder of the present chapter argues that mental events exhibit certain qualitative and intentional features that cannot be plausibly attributed to physical events, and thus that the former must be distinguished from the latter even if this does require an inflation of our ontology, as there is no other position that offers an equally plausible explanation of the available data. (See also Kim (2011, pp.100-10).)

[122] The content of this claim of course depends largely on how one specifies the inputs and outputs of the functions in terms of which mental properties are to be analyzed. One option is to have the inputs and outputs be types of mental states, physical stimuli, and behavioral responses. Another option is to require that the inputs and outputs be specific types of neurobiological events. As Block (1978) points out, the latter, more fine-grained approach is open to the charge of "chauvinism," for it entails that beings with bodies significantly different from our own would, as a matter of definition, be incapable of possessing minds. Throughout the following discussion, I therefore assume that the functions that the T/NRPist aims to equate mental properties with are specified in more general, behavioral terms.

in nature. Since the instances of any functionally individuated property are identical to the events that perform its function, it follows that mental events (mental property instances) are identical with physical events. Yet given that the same mental property can be realized by distinct types of physical events, it likewise follows that, due to the nomological irreducibility of multiply realizable properties, mental properties are distinct from and nomologically irreducible to physical properties. As these two conclusions are equivalent to the two central tenets of T/NRPism, the forgoing argument hence entails that T/NRPism is true.

While the functional nature of mental properties is typically taken by those who advance this type of argument to be a necessary truth (which is arrived at either through *a priori* analysis of folk psychological concepts, or through the empirical investigations of a more sophisticated scientific psychology), the identity of mental and physical events is usually viewed as both contingent and *a posteriori*, as it is thought that the functions that individuate mental properties could be carried out by non-physical entities in other possible worlds. The same holds likewise for the nomological irreducibility of mental properties and the consequent falsity of type-physicalism, as proponents of this argument also typically hold that mental properties could have turned out to be type-identical with physical properties, if the laws of nature had been such that each mental property could only be realized by events of a single physical type.

Framed in terms of the discussion of realization in Chapter 3, the conclusion of this alternative argument for T/NRPism can be seen to entail that mental properties are all physically realized in the manner of Type 1, wherein the realized entity is individuated in strictly functional terms, and its realizers ensure its occurrence or instantiation by

performing the very function that individuates it. Henceforth, I will refer to this thesis as the view that mental properties are *functionally reducible* to physical properties, where a type of property *A* will be said to be functionally reducible to a different type of property *B* iff all properties of type *A* can be fully analyzed in functional terms, and the functions that individuate them can all be performed by instances of properties of type *B*. By this definition, the foregoing argument for T/NRPism purports to establish that mental properties are functionally, but not nomologically reducible to physical properties. Introducing this distinction between functional and nomological reduction is meant to help underscore the fact that although the proprietary laws and kinds of psychology cannot, according to the above argument, be nomologically reduced to or identified with those of physics, there is still some sense in which the mental ends up getting reduced to the physical, inasmuch as the extensions of mental properties are (if the argument is sound) populated solely by physical events.

The remaining two sections of the present chapter will offer some reasons for denying that mental properties can be functionally reduced. The present section defends this claim with respect to phenomenal properties, or *qualia*.

Phenomenal properties, such as the smell of a skunk, the feeling of pain, or the sensation of red, have long been viewed as the primary sticking point for any attempted functionalist reduction of the mind. The basic reason for this is that there appears to be "something it is like" to experience such properties that cannot be equated with the causes or effects of their instantiation, or their relations to other types of mental states. The experience of pain, e.g., clearly feels a certain way, but its having this distinctive character does not seem to consist merely in the fact that it is typically caused by bodily

damage, that it often produces feelings of distress or anxiety, that its normal causes can fail to produce it if one is temporarily unconscious or currently engrossed in some demanding task, that it usually results in behavior aimed at removing its cause, etc. To provide a functional analysis of a given property, however, just is to define it in such strictly relational terms. (More precisely, we might say that a property *P* admits of functional analysis iff it can be shown to be equivalent to the property of producing outputs/effects O_1, O_2, O_3,… in response to inputs/causes I_1, I_2, I_3,… under conditions C_1, C_2, C_3,…, where each combination of a certain input/cause I_i with a certain set of conditions C_j corresponds to a certain distribution of probabilities over *P*'s various possible outputs/effects, such that for each potential output/effect, there is a certain probability that *P* will produce it in response to that input under those conditions. To say that *P* is functionally analyzable is hence to say that the nature of *P* is fully determined by the relations it bears to the properties involved in O_1, O_2, O_3,…, I_1, I_2, I_3,…, and C_1, C_2, C_3,…) Thus, if phenomenal properties are individuated even just partly in terms of the intrinsic, non-relational quality of "what it is like" to experience them, then such properties cannot be functionally reduced, as there will be something essential to them that escapes functional analysis.

Reflecting upon the character of my own experience, it seems to me almost certain that (a) the properties I refer to by way of such expressions as "the smell of a skunk" or "the sound of a horn" do in fact exhibit certain distinguishing features that are neither identical with, nor logically entailed by any sort of functional role that these properties might occupy, and (b) that these features are the primary marks whereby I identify such properties, and are moreover the characteristics that seem most essential to

them. By this I mean that the very nature or identity of the property that I refer to as, e.g., "the smell of a skunk" appears to be constituted primarily by that feature that is revealed to me in my conscious experience of it, so that the idea that that property could be instantiated without exhibiting that feature strikes me as somehow contradictory. The same does not seem to be true, however, of the functional role that that property occupies, for while the smell of a skunk does indeed have certain characteristic causes, effects, and relations to other mental states (e.g., it is typically caused by the presence, in one's nasal passages, of molecules of the sort emitted from skunk scent glands, and it often produces feelings of nausea or disgust, as well as wrinkling of the nose and other expressions of displeasure), I nonetheless do not encounter the same sort of difficulty in attempting to conceive of that property's being instantiated without standing in such relations, or of something's standing in such relations without being an instance of that property, as I do in attempting to conceive of an instance of that property that is not accompanied by the distinctive type of experience whereby I distinguish it from other properties, or of something's possessing that experiential character without being an instance of that property.

If the foregoing observations apply equally to the experience of others, then it would seem to be an empirical fact about our experience of phenomenal properties that such properties strongly *appear*, at least, to be incapable of being analyzed in purely functional terms. On the other hand, it must be admitted that there also seems to be nothing internal to our experience of such properties that could enable us to determine whether or not this appearance is accurate. For this reason, many of the familiar thought experiments that purport to establish the functional irreducibility of consciousness (e.g.

Jackson's (1982) Knowledge Argument, Ned Block's (1978) Absent and Inverted Qualia Arguments, and David Chalmers' (1996) Conceivability Argument) are I think better interpreted as illustrations of the empirical fact just mentioned, than as demonstrations of the accuracy of the appearance it describes. For the reason why it seems to (most of) us that Mary acquires new (propositional) knowledge upon leaving her black and white room, that a homunculi-headed robot would lack consciousness, and that zombies and inverted qualia are possible (or at least readily conceivable), is that phenomenal properties exhibit certain features that appear to be distinct from the functions that their instances perform. Our reactions to these thought experiments hence cannot be cited as grounds for assuming that this appearance is veridical, for the appearance is what accounts for these reactions, and it would be just as capable of producing them whether it is accurate or not.

Given, moreover, that only some of the essential features of phenomenal properties need be distinct from the performance of any function in order for such properties to qualify as functionally irreducible, those who deny that phenomenal properties are functionally reducible can allow that there may be certain functions that such properties cannot be instantiated without performing, or whose performance is metaphysically sufficient for the instantiation of such properties. For (a) one might hold that phenomenal properties are partly, but not wholly individuated in functional terms, in which case each phenomenal property would be associated with a certain function whose performance was metaphysically necessary, but not sufficient for its instantiation. Or (b) one might hold that while such properties are not even partly individuated in functional terms, each phenomenal property is nonetheless associated with a certain function whose

performance is both metaphysically necessary and sufficient for its instantiation. The latter option is made available by the fact, demonstrated by Kit Fine (1994), that something's being necessarily true of a thing is not sufficient for its also being essential to it. It may hence be necessarily true of a certain phenomenal property that it cannot be instantiated without its instances performing a certain function, even though that function is not even partly constitutive of the essence of that property.

All this goes to show that while the seemingly non-functional nature of phenomenal properties helps to explain the conceivability of situations (e.g. zombie worlds and qualia inversions) wherein phenomenal properties and the functions they are typically associated with come apart, such situations needn't be nomologically, or even metaphysically possible in order for that appearance to be accurate. On the other hand, any evidence for the possibility of such situations would likewise count as evidence in favor of the view that qualia cannot be functionally reduced, for though the latter thesis does not entail the former, it is nevertheless entailed by it. While the mere conceivability of such situations cannot, for the reasons noted above, be given much weight in deciding this issue, more compelling empirical evidence for the independent variability of qualia and their characteristic functions might perhaps be drawn from studies of prefrontal lobotomy and cingulotomy patients who claim to have experiences of pain without displaying any of the aversive behavior typically associated with such experiences[123], or from considerations noted by C.L. Hardin (1988, pp.140, 142), which suggest that many of the more serious obstacles to the possibility of functionally undetectable spectrum

[123] See Aydede (2000, pp.546-8).

inversions in trichromats would be avoided if, instead of switching blue with yellow and/or red with green, red and green were both switched, respectively, with yellow and blue.[124]

Although such evidence may help bolster the claim that phenomenal properties are functionally irreducible to physical properties, its role in the case against T/NRPism is I think ultimately secondary, for the most compelling reason to deny that phenomenal properties can be functionally reduced still seems to consist in the observation that our experience of such properties presents them as having certain features that appear to be distinct from any function they perform. While this appearance may turn out to be misleading, and thus falls short of *demonstrating* that phenomenal properties are functionally irreducible, it nonetheless seems vivid enough that we would be reasonable to accept it as accurate unless substantial reasons can be given for thinking otherwise. In short, I take the apparent non-functional nature of phenomenal properties to provide a defeasible, but very strong *pro tanto* reason to believe that phenomenal properties are functionally irreducible. This reason also seems to me to be both more compelling than any other reason that has been offered in favor of adopting such a belief, and strong enough to place the onus on those who insist that phenomenal properties *can* be functionally reduced to provide some principled means of explaining the apparent non-functional nature of such properties away. The remainder of the present section responds to the two most prominent strategies that have been employed to this end.

[124] Hardin (1988, pp.140-2) does, however, note a number of phenomenal and psychophysical asymmetries between the red-green and blue-yellow opponent channels that would remain to be dealt with. See also Hilbert and Kalderon (2000).

3.i. The Phenomenal Concept strategy

The first of these strategies[125] (commonly referred to as the Phenomenal Concept strategy (PCS)) maintains that the seemingly non-functional nature of phenomenal properties (and the seemingly non-physical nature of phenomenal events) can be fully explained in terms of certain differences between our phenomenal and physical/functional concepts. These differences are said to account for the appearance of contingency that surrounds the relationship that the phenomenal character of an experience bears to both the functional role it occupies and the physical events that realize it. With the disparity between the phenomenal and the physical/functional thus explained in purely epistemic terms (i.e., in terms of differences between concepts), the suggestion is then made that the properties and events that our phenomenal and physical/functional concepts pick out are nonetheless metaphysically the same; the result being that the apparent non-identity of phenomenal and physical/functional properties and events is attributed solely to the different ways we have of conceptualizing the physical/functional events taking place in our own brains.[126]

---

[125] Proponents of this strategy include Loar (1997), Hill and McLaughlin (1999), Tye (1999; 2003), Block (2007), and Papineau (2007). Tye (2009, ch.3) has, however, since changed his view.

[126] The Phenomenal Concept strategy bears certain similarities to the Ability Hypothesis, defended by Lewis (1990) and Nemirow (1980, p.475), which also seeks to provide a purely epistemic explanation for the apparent differences between phenomenal and physical/functional states, by maintaining that to know "what it is like" to have a certain experience is not to have access to any non-physical "phenomenal information," but rather to have "an ability to place oneself, at will, in a state representative of the experience." The objection raised to PCS below should indicate some of the difficulties facing any such approach. (More direct criticisms of the Ability Hypothesis can also be found in Gertler (1999, pp.322-6) and Raymont (1999).)

While proponents of PCS hold differing opinions as to the exact nature of phenomenal concepts, the ability of such concepts to account for the apparent differences between phenomenal and physical/functional properties and events is usually thought to hinge on their possession of two key features. First, such concepts are understood to pick out their referents directly, "without the use of any descriptive, reference-fixing intermediaries" (Tye, 1999, p.713). In this they are said to differ from concepts (e.g. our folk concept of *water* as *the transparent, odorless, tasteless stuff that falls from the sky, fills lakes and rivers, comes from taps, etc.*) that pick out their referents by way of some "contingent mode of presentation" (Loar, 1997, p. 600). Second, the direct manner in which phenomenal concepts refer is taken by proponents of PCS to render them distinct from and unanalyzable in terms of physical/functional concepts.[127] Those who take this view thus disagree with "analytic" functionalists, e.g. David Lewis (1966) and David Armstrong (1968), who hold that the equivalence of mental properties with certain functional properties can be established through *a priori* conceptual analysis. According to proponents of PCS, the special features of phenomenal properties are instead thought to ensure that "no amount of a priori reflection on phenomenal concepts alone will reveal phenomenal-physical or phenomenal-functional connections, even of a contingent type" (Tye, 1999, p.715). Yet since the apparent distinctness of phenomenal and physical/functional properties can be attributed to the lack of any *a priori* connections between our phenomenal and physical/functional concepts, and the lack of such

[127] This claim is sometimes bolstered by the observation that phenomenal and physical/functional concepts "have quite different conceptual roles" (Loar, 1997, p.602). See also Hill and McLaughlin (1999, p.448)).

connections can supposedly be fully explained in terms of the features of phenomenal concepts just noted, the apparent distinctness of phenomenal and physical/functional properties, and our consequent ability to conceive of the one without the other, cannot be taken to warrant the metaphysical conclusion that such properties are in fact distinct. To do so would be to unjustifiably infer "from the lack of resemblance in our phenomenal and physical-functional conceptions a lack of sameness in the properties to which they refer" (Loar, 1997, p.605).

One problem that seems to cast serious doubt on the ability of PCS to provide a satisfactory explanation for the apparent non-identity of phenomenal and physical/functional properties and events concerns its seeming *in*ability to account for the rationality of those who deny that such properties and events are in fact identical.[128] While those who make this denial may of course be wrong, it seems fair to say their denial is not wholly irrational. The explanation that the PC strategist offers for this fact is that the physical/functional nature of pain is *a posteriori*. Those who lack the empirical information that reveals the identity of the property of being in pain with a certain functional role, or the identity of a particular experience of pain with a certain physical event that occupies that role, can thus rationally reject these identities, just as someone who lacks the empirical information that supports the identification of water with H_2O can rationally deny that water is H_2O.

The question of how we are to explain the *a posteriori* status of certain identity claims (e.g. Hesperus = Phosphorus, or Scott = the author of *Waverly*) is itself a

[128] A forceful statement of this objection can be found in White (2007).

philosophical puzzle that goes back at least to Frege. A standard solution, adapted from Saul Kripke (1980, pp.150-5), is to maintain that in order for an identity statement to be capable of being *a posteriori*, at least one of the terms flanking the identity sign must pick out its referent by way of some contingent[129] mode of presentation that is conceptually distinct from the mode of presentation (if any) through which the other term of the identity statement refers. The *a posteriori* status of the statement "Water is H_2O" can thus be accounted for by the fact that "the reference of ["Water"] was determined by an accidental property of [its] referent [(H_2O)]," viz. that of appearing to us as the bearer of certain superficial qualities (e.g., transparency, lack of odor and taste, potability, etc.) (Kripke, 1980, p.152). Since the observable features of H_2O could conceivably be very different from those it actually has (one might, e.g., easily imagine it appearing to us instead as a black, sticky, pungent goo), the possession of this property by H_2O is at least *epistemically* contingent. Given, then, that the features that make up the mode of presentation through which our concept of *water* refers to H_2O are epistemically contingent features of the latter, in that there are no *a priori* connections between these features and H_2O itself, it is possible for someone to rationally deny that water is H_2O, despite the fact that "Water is H_2O" is necessarily true. For one who lacks the empirical information showing that the bearer of the superficial features associated with water *is* H_2O would have no reason to believe that this is so.

[129] The explanation of the *a posteriori* status of the identity statements at issue seems to require only that the relation between the referent and its mode of presentation be *epistemically* contingent. It need not be *metaphysically* contingent as well.

The trouble for PCS is that those who endorse it cannot avail themselves of this explanation for the *a posteriori* status of identity statements of the form ⌜ Phenomenal property *M* = Functional property *F* ⌝ or ⌜ Phenomenal event *m* = Physical event *p* ⌝, because according to PCS, phenomenal concepts pick out their referents *directly*, without the mediation of any contingent mode of presentation, and the same likewise seems true of our concepts of the physical/functional properties and events that phenomenal properties and events are supposedly identical with. For even assuming that functional concepts (e.g. that of producing ameliorative behavior in response to bodily damage) and concepts of types of neural events (e.g. the release of dopamine from axon terminals in the midbrain) have certain modes of presentation that are distinct from the functions or types of neural events they refer to, the ability of such concepts to pick out their referents does not seem to depend on the descriptive features that their associated modes of presentation involve in the way that the referent of one's concept of *water* depends upon the descriptive features involved in its mode of presentation. If this is correct, though, then the requirement that at least one of the terms of an *a posteriori* identity statement must pick out its referent by way of some (epistemically) contingent property of the latter will fail to be satisfied in the case of identity statements of the form ⌜ Phenomenal property *M* = Functional property *F* ⌝ or ⌜ Phenomenal event *m* = Physical event *p* ⌝. In order to make good on their claim to account for the apparent distinctness of phenomenal and physical/functional properties and events, and the consequent rationality of those who deny their identity, proponents of PCS must hence reject the requirement just

mentioned[130] and offer some alternative theory of *a posteriori* identity statements that allows the terms of such statements to be conceptually distinct and to both pick out their referents directly. The problem is that it unclear how any such theory could work. For if I have two distinct concepts that have no *a priori* connection to one another, and both these concepts pick out their referents directly, without relying on any superficial qualities that are distinct from the referents themselves, then I see no way to avoid the conclusion that my concepts are of two distinct things (though they may be of two distinct properties or parts of the same thing). It would seem, therefore, that the PC strategist's attempt to provide a strictly epistemic explanation for the apparent non-identity of phenomenal and physical/functional properties and events ultimately proves unsuccessful.

3.ii. Representationalism

The second major strategy used by those seeking a functionalist reduction of the mind to address the problems raised by phenomenal properties goes by the name of Representationalism, or Intentionalism. While distinct from PCS, Representationalism is nonetheless consistent with it, and the two strategies are sometimes employed in conjunction with one another.[131] The core thesis of Representationalism is that

[130] Loar (1997, p.600) does indeed explicitly reject this requirement. White (2007, pp.22-5) responds with a compelling defense of a weaker version of it that proponents of PCS are still obliged to deny.

[131] Tye (1995), e.g., has at times advocated both Representationalism and PCS. Other prominent Representationalists include Dretske (1995), Harman (1990), and Lycan (1996, ch.4). Spinoza (*Ethics* IIax3, IIp11proof) might also be cited as an early adherent of the view. Anyone who allows for the existence of phenomenal states and accepts Brentano's (1874) thesis that intentionality is "the mark of the mental" will qualify as a Representationalist as well.

phenomenal states are intentional states whose content represents certain physical/functional properties of external objects, or states of one's own body. On this view, the "redness" involved in a visual sensation of red might hence be treated as part of the content of an intentional state that represents a certain external surface as having a certain spectral reflectance distribution, and the sensation of pain might likewise be identified with the content of an intentional state representing damage at a certain location in one's body. Assimilating phenomenal states to intentional states in this manner helps the T/NRPist in two ways: First, those who do distinguish between phenomenal and intentional states typically view the latter as being much more susceptible to functional analysis than the former. Thus, if phenomenal states *are* intentional states, then those strategies that have proven useful in the attempt to functionally reduce intentionality can be applied to phenomenal properties as well, thereby opening up the possibility of a unified functionalist reduction of mind. Second, the fact (if it is a fact) that phenomenal states are intentional would enable T/NRPists to explain the apparent distinctness of phenomenal and physical/functional properties and events as due merely to certain special features of the representational content of phenomenal states (e.g., as due to the fact that the content of phenomenal states is non-conceptual, whereas that of propositional thoughts about physical/functional properties and events is conceptually structured). Given these advantages, it is easy to see why T/NRPists might be inclined to accept a Representationalist theory of phenomenal consciousness.

While some initial support for Representationalism might be drawn from the fact that phenomenal states are often treated as having veridicality conditions, (which would

suggest that such states must hence have some kind of intentional content that can be more or less accurate of whatever it is they represent), the principal argument for the view is the so-called argument from "transparency." This argument adverts to the fact that when we attempt to focus our attention upon the character of our conscious experience, our attention often seems to look straight through our experience to whatever it is that our experience seems to be an experience of, and any phenomenal features that we come across while engaged in this reflective activity consequently seem to present themselves as features of the objects of our experience, rather than of our experience itself. A compelling statement of this point can be found in the following, oft-quoted passage from Gilbert Harman (1990, p.39):

> When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience. Nor does she experience any features of anything as intrinsic features of her experience. And that is true of you too…When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree 'from here.'

On the basis of such observations, the argument goes on to claim that since further introspection fails to reveal any intrinsic properties of our experience that do not seem more properly attributable to the objects of our experience, the character of our experience seems to be fully constituted by the properties that it represents other things as having. The only properties that our experiences themselves have are the relational properties of representing other things as being a certain way.

While the most immediate response to the argument from transparency would be to simply deny that the only introspectible features of our conscious experiences are those that such experiences represent other things as having, and insist that we do in fact have introspective access to a certain phenomenal element in our experiences that "**outruns**" whatever representational content they might have, I think it best not to try and argue the issue on such grounds, for the Representationalist can match any appeal to phenomenology as showing that "representational content is [*not*] all there is to phenomenal character" with their own phenomenological evidence indicating that conscious states are transparent (Block, 1996, p.20). In short, given that both sides in the current debate can appeal to phenomenological evidence that weighs in their favor, any attempt to settle the debate on phenomenological grounds alone seems destined to end in stalemate. For this reason, many of the putative counterexamples to Representationalism that appeal to various thought experiments (e.g. Block's (1990) Inverted Earth, or the cases cited by Peacocke (1983, pp.12-7)) to suggest that it is possible for two distinct states to share the same representational content but differ phenomenologically, or be phenomenologically indiscernible while having different representational content, strike me as inconclusive, since those who do not already take their experience to show that there is more to any phenomenal state than its representational content will be unlikely to view such thought experiments as demonstrating this to be so. There are, however, two other points that together seem to me to significantly undermine the idea that Representationalism can provide T/NRPists with the functionalist reduction of phenomenal consciousness they seek.

First, it should be remembered that the Representationalist's assimilation of phenomenal to intentional states will only render the former amenable to functionalist reduction if intentionality is itself functionally reducible. While many are (as noted above) quite optimistic about the prospects for such a reduction of intentionality, its success is not a foregone conclusion, and the following section will indeed raise some objections to various "naturalization projects" of this sort. Thus, even if Representationalism turns out to be true, it is by no means certain that it will enable T/NRPists to account for phenomenal consciousness in a manner consistent with their position.

The second point, drawn from George Molnar (2003, pp.74-80), calls into question one of the key implications of Representationalism; viz. that phenomenal states possess non-natural meaning. As introduced by Paul Grice (1957), the distinction between natural and non-natural meaning is meant to distinguish the sense of "meaning" in which, e.g., the rings on a tree trunk mean that the tree it belonged to grew to be *x* years old before being cut down, or certain spots on the skin mean measles, from the sense in which, e.g., the chiming of the clock means that it is noon, or a person's assertion that "*p*" means *p*. Among the criteria that Grice uses to distinguish between these two senses of meaning, one of the most crucial is that in cases of natural meaning, *x means that p* entails *p*, whereas in cases of non-natural meaning, no such entailment holds. This criterion is by itself sufficient to show that the relation between the representational content and the object of an intentional state must be of the non-natural sort, for such states can be directed towards things that do not exist. Thus, if phenomenal states are, as the Representationalist claims, intentional, then the relation between the

content of such states and the things they represent must also be non-natural. But it isn't. Therefore, Representationalism is false.

Consider pain (Molnar, 2003, pp.77-80). As previously mentioned, the standard Representationalist account of pain treats it as the content of an intentional state representing damage to a certain part of the body. If this is correct, then the relation between an experience of pain and the bodily damage and location it represents must be non-natural. Now whatever it means for one thing to represent another, I take it that all representations must at the very least be distinct from the things they represent. For this reason, though, there seems to be no sense in which pains can be said to mean or represent the bodily locations where they are felt as being, for the location where a pain is felt is partly constitutive of, and perhaps even essential to the very pain that is experienced. As Molnar (2003, p.77) puts it, "a headache does not *represent* my head as hurting, it *is* my head hurting." Pains thus cannot represent the bodily locations at which they are felt, for there is not enough "space" between the two for the former to be described as representing the latter.

What about the relation between pain and bodily damage? The problem here is that this relation does not seem to satisfy the condition for non-natural meaning noted above; viz. that if *x* non-naturally means *p*, then it does not follow that *p*. For while pain does in all cases seem to signify some form of bodily damage, it also seems that whenever pain is felt, the damage it represents is none other than the damage that actually caused it. And since a pain cannot be caused by damage that does not exist, it follows that if a given experience of pain means that a certain part of the body is damaged, then that part of the body has indeed been damaged. This means that the relation between an

experience of pain and the bodily damage it represents cannot be non-natural, and pain therefore cannot be an intentional state (Molnar, 2003, p.79).

"Phantom pains" (pains experienced in phantom limbs) may seem to present a counterexample to this argument, since in such cases one might take the location of the represented damage to be in a body part that no longer exists (or is at least no longer part of the subject's body). I suggest, however, that the correct interpretation of such cases is to view phantom pains as instead representing the actual body damage (e.g. some disorder in the subject's brain) that truly causes them. In support of this, one might easily imagine an analogous situation in which a doctor corrects an earlier misdiagnosis of the cause of a subject's pain (wherein the cause was perhaps wrongly identified with damage at the location where the pain was felt) by saying: "It turns out that your pain does not mean damage to body part *x* after all. In fact, it means damage to body part *y* (which is what's really causing it)." If this interpretation is correct, then it seems that, unlike intentional states, pain at best has only natural meaning.

Certain Representationalist treatments of experiences of color may, however, seem to escape this kind of objection, for as noted above, the standard Representationalist account of phenomenal colors treats them as contents of intentional states that represent the spectral reflectance distributions (SRDs) of the surfaces of external objects, with each color corresponding to a different SRD. Since, as discussed in Chapter 4, phenomenally indistinguishable color experiences can be produced by surfaces with different SRDs, phenomenal colors would on such an account be capable of misrepresenting, and would thus have non-natural meaning. The problem with this suggestion, though, is that color experiences would then end up misrepresenting far too often, for in addition to the fact

that surfaces with different SRDs can produce matching color experiences, surfaces with the same SRD can also produce different color experiences when viewed under different illuminants, or when viewed under the same illuminant by individuals with slightly different cone sensitivities. Given, moreover, that different SRDs can generate matching experiences of color, the selection of a single SRD as the *representatum* of a given type of color experience also seems arbitrary, for there will always be a number of other SRDs that are equally capable of producing that same type of experience. At any rate, since there would be little motivation for treating phenomenal colors as representational if they would thereby end up *mis*representing as frequently as they would under the account just described, the Representationalist seems compelled to find some other physical/functional property to serve as the *representatum* of phenomenal color. The most promising alternatives appear to be either the disjunction of all the SRDs that can give rise to a given color experience, or the relational property of emitting, transmitting, or reflecting light whose spectral distribution is such as to push the perceiver's opponent processing channels into a state that corresponds to the experience of a certain color. The formal proposal seems rather *ad hoc*, especially when one considers that the disjunctions associated with different phenomenal colors would overlap. The second proposal has problems in accounting for those colors (e.g. brown) that are seen only in contrast with others, for the same response in the opponent processing channels that gives rise to the experience of such a color when other, contrasting colors are present in the visual field will instead generate an experience of one of the spectral hues when all contrasting colors are removed.

With these proposals rejected, the next best option for the Representationalist would seem to be to suggest that phenomenal colors just represent whatever brain state is currently producing them. But in this case, color experiences cannot misrepresent, and therefore do not have the non-natural meaning characteristic of intentionality. In sum, as with pain, there does not appear to be any satisfactory Representationalist account of phenomenal colors under which they have anything more than natural meaning. Since there is no apparent reason to think that experiences of pain or color differ from other phenomenal states in this regard, we may conclude provisionally that phenomenal states in general are distinct from intentional states, and that Representationalism is therefore false. All this is, however, consistent with the idea (which will be explored further below) that while phenomenal states are themselves non-representational, the phenomenal character of an experience may nevertheless contribute to fixing the representational content of the perceptual and other intentional states that are formed in relation to it.

If the forgoing criticisms are sound, then neither PCS nor Representationalism can provide T/NRPists with a satisfactory explanation for the apparent distinctness of phenomenal and physical/functional properties and events that might clear the way for a functionalist reduction of phenomenal consciousness. As these two strategies are the most promising and well-developed of the proposals that T/NRPists might employ to this end, it seems reasonable to conclude that no alternative strategy currently on offer is likely to fare any better. As things now stand, then, it looks as though the simplest explanation for the apparent distinctness of phenomenal and physical/functional properties and events is also the best one; viz. that phenomenal properties and events *appear* to be non-functional

and non-physical because they in fact *are* so. Until a better explanation is put forward, we thus arrive at property dualism for phenomenal properties by way of inference to the best explanation.

4. Intentionality

Having argued for the functional irreducibility of phenomenal states, the remainder of the present chapter will be spent arguing that intentional states, e.g. believing that *p*, desiring that *p*, perceiving *q*, or thinking about *r*, are functionally irreducible as well. As these two types of states are together typically viewed as exhausting the class of mental states, the success of this and the previous section will imply that the mind in general cannot be functionally reduced, either in whole or in part, and thus that the functionalist route to T/NRPism is bound to fail.

As previously mentioned, the claim that intentional states are functionally irreducible is much more contentious than the claim that phenomenal states are so. This is because whereas the phenomenal character of an experience strongly appears to be distinct from any function that experience performs, intentional attitudes seem to be fully analyzable in strictly functional terms. The attitude of believing that *p* might thus, e.g., be plausibly defined as the state that is typically caused by perceiving that *p*, or being told that *p*, or believing both that *r* and that if *r*, then *p*, etc., and which itself typically causes affirmative responses to the question whether *p*, and behavior of type *x* whenever one desires that *y* and believes that if *p*, then acting in manner *x* is the best way to bring it about that *y*, etc. The fact that the various intentional attitudes can be analyzed in such

terms is, however, still insufficient to establish that intentional states are fully functional in nature, for while such states are partly individuated in terms of the attitudes they involve (e.g. belief, desire, perception, thought, etc.), the complete individuation of any intentional state also requires reference to its representational *content* (that which is believed, desired, perceived, thought, etc.). This is because in addition to the differences between, e.g., desire and belief *simpliciter*, individual beliefs and desires also differ among themselves depending on what they are "about" (as, e.g., the belief that it's raining differs from the belief that the Pope is infallible). Thus, while the attitudes of believing, desiring, thinking, etc. may indeed admit of a complete functional analysis, the same must also be true of the representational content of intentional states in order for such states as a whole to be functionally reducible. And if such states are, moreover, to be functionally reducible to *physical* states, as T/NRPism requires, the functions in terms of which their individuative attitudes and contents are analyzed must also be such as can be performed by purely physical events. The aim of this section, then, will be to defend the idea that the representational nature of intentional states cannot be functionally reduced in a manner consistent with T/NRPism, because the representational contents of such states cannot be analyzed in terms of any functions that physical events are capable of performing.

While physical states do seem perfectly capable of satisfying the functional descriptions that pick out the various intentional attitudes, it is, I take it, much more difficult to see how such states might serve as the representational content of an intentional state, as it is *prima facie* unclear how any purely physical entity could be said to (non-naturally) represent something in such a way as to be true, false, or more or less

accurate of that thing it is supposedly "about," in the way that the representational contents of intentional states inevitably are. In light of this fact, the burden of proof seems to be on the physicalist to provide some naturalistic explanation[132] for how physical states might come to possess the kind of representational content that any intentional state must have. If no such explanation can be given, our conclusion must be that the representational features of intentional states (which are indeed what *make* them intentional) cannot be reduced to or explained in terms of any purely physical states or processes.

Of the various naturalistic theories of representational content that have been developed thus far, the two that to me seem the most promising are the Causal theory proposed by Fodor (1987), and Teleological theories of the sort endorsed by Fred Dretske (1988) and Ruth Millikan (1984; 1989).[133] While these theories do much to further our understanding of the causal and biological conditions that give rise to intentionality, both face serious problems that are I think significant enough to suggest that the representational contents of intentional states cannot be functionally reduced in the manner they propose.

4.i. Causal Theories of Representational Content

[132] I.e., one that makes no use of any semantic, intentional, or representational notions, but instead draws solely on the explanatory resources of the natural sciences.

[133] One of the other major theories of representational content, Conceptual/Inferential Role Semantics, is criticized effectively by Fodor (1987, pp.76-83) and Fodor and Lepore (1991), as well as being open to some of the objections raised to Causal theories of representational content below.

Causal theories of representational content (CTR) attempt to facilitate the reduction of intentional to physical states by suggesting that for one thing to represent another in the way that the representational content of an intentional state represents its object is for it to be causally related to the thing it represents in the way that certain physical states of intentional beings are related to various features of their environment. In its most basic form, CTR holds that *x* represents *y* as *F* iff *y* causes *x* and *x* is of some type whose tokenings are reliably caused by instances of *F* (Fodor, 1987, p.99). Thus stated, however, CTR is clearly inadequate, for if representations were to represent whatever causes them, then it would be impossible for any representation to *mis*represent (Fodor, 1987, pp.101-2). To see why, imagine that whenever a certain type of object *F* (an apple, say) interacts with my sensory organs, it causes the tokening in me of a certain type of physical state *x*. According to the most basic formulation of CTR, this may seem to support the claim that *x* represents *F*. But as is the case with all representations (or at least with all those that possess non-natural meaning, and are thus capable of serving as the content of intentional states), *x* is not a perfectly reliable indicator of the presence of *F*s, for it is, or at least should be possible for it to sometimes (or perhaps even *always*) represent my environment *in*accurately. Such occasions would seem, on the present model, to consist in the tokening of *x* in response to some object that is *not* of type *F*, but rather of some other type *G* (e.g. an orange seen in dim lighting). Under such conditions, *x* might be thought to *mis*represent the *G* that caused it as an *F*. Yet if representations represents whatever causes them, then since tokenings of *x* can be caused by both *F* and *G*, *x* in fact never represents just *F*, but rather the disjunction of *F* and *G*, as well as any other type of thing that happens to be causally sufficient for *x*. And if this is so, then

tokens of *x* caused by *G*s do not misrepresent those *G*s as *F*. They instead accurately represent them as $F \vee G$ ($\vee$ *any other type of thing that is causally sufficient for x*). As the same reasoning can be applied equally to anything that any potential representation might be thought to *mis*represent, the basic formulation of CTR thus entails that representations can never *mis*represent anything, which in the case of intentional states, at least, is plainly false.

In an attempt to resolve this problem (which is commonly referred to as the "Disjunction Problem"), Fodor (1987) proposes some modifications to the basic formulation of CTR that enable one to distinguish, among the causes of a given representation, between those that do and those that do not contribute to its representational content. If successful, these modifications would enable advocates of CTR to handle examples of the sort just discussed by maintaining that *x* represents *F*, not *G* or $F \vee G$, because *F* is the only cause of *x* that meets the conditions for determining its content, and therefore any token of *x* that is caused by a *G* (or indeed any non-*F*) thereby *mis*represents its cause as an *F*. Fodor's proposal, then, is that we can distinguish between those causes that do and those that do not help fix the content of a given representation *x* on the basis of the (synchronic) asymmetric dependence of the causal relations between *x* and those of its causes that do not contribute to its content on the causal relations that hold between *x* and those of its causes that do. The basic idea is that if *x* is caused by, but does not represent *G*, this is because the ability of *G*s to cause tokenings of *x* depends on the fact that tokenings of *x* are also caused by some other type of object *F*, where the fact that *F*s cause tokenings of *x* is not likewise dependent upon the fact that *G*s do so as well. In modal terms, if *x* is caused by both *F* and *G* but represents only *F*, this is because in

the closest possible world in which *F*s do not cause tokenings of *x*, *G*s don't either, whereas in the closest possible world in which *G*s do not cause tokenings of *x*, *F*s still do (Fodor, 1987, pp.106-11).

While various reasons have been noted for questioning the adequacy of Fodor's proposed solution to the Disjunction Problem, two objections in particular seem to me to highlight certain basic deficiencies in any potential form of CTR, Fodor's Asymmetric Dependence Theory included.[134] The first of these stems from the fact that since mental representations supervene on (and are, according to the T/NRPist, token-identical with) brain states, they can be produced by what Fred Adams and Ken Aizawa (1994, pp.216-7) call "pathological causes," e.g. "[a] severe blow to the head, hallucinogenic drugs, a brain tumor, a high fever, or a current passed through a series of well-placed microelectrodes in the brain." The problem this raises for CTR is that since tokenings of any mental representation can be produced by such "pathological" causes as well as by the thing it supposedly represents (e.g. *apple* or *horse*), even assuming that the Disjunction Problem can be resolved so as to allow that mental representations do not represent the disjunction of all their potential causes, it is unclear what principled grounds advocates of CTR can give for saying that it is its alleged *representatum*, rather than some one of its potential "pathological" causes, that a given mental representation in fact represents. Fodor's Asymmetric Dependence condition is of no help here, for the laws that render the various "pathological" causes of a given representation sufficient for its

[134] Fodor (1987, p.108 fn5) gestures towards a potential response to one of the more obvious objections to CTR, viz. that intentional states can be about things that do not exist, and which thus cannot cause what represents them.

tokening do not appear to (synchronically) depend in any way on the laws that connect it to its alleged *representatum*, nor *vice versa*. What seems needed is instead an appeal to the fact that it is its alleged *representatum* (and not its "pathological" causes) that a representation is *supposed* to represent, and which it does represent when it and/or the internal systems that regulate its tokenings are functioning properly. The only way of developing such a response within a naturalistic framework would, however, be to advert to the biological function(s) of the representation and/or the internal systems that are responsible for producing it. To take this route, though, is to allow that in determining the content of a given representation, it is not enough to consider merely its contemporary causal relations, but that one must take its evolutionary history into account as well; which is to say that CTR is inadequate unless supplemented by a Teleological account of the sort to be discussed further below.

The second objection to CTR is that the representational contents of intentional states are individuated at a much finer level of grain than CTR seems able to account for. As noted by Gary Gates (1996), the gist of this criticism can be brought out by an analogy with Quine's (1960, ch.2) argument for the indeterminacy of translation. For just as the necessary coincidence of rabbits, rabbit stages, undetatched rabbit parts, etc., and the equivalence in the effects they produce on one's sensory apparatus makes it impossible, by Quine's lights, to tell which of these candidate referents the term "gavagai" refers to (even if uttered only in the presence of rabbits), the necessary coincidence and causal equivalence of these various candidate *representata* likewise make it impossible for any purely Causal theory of representational content to explain how something could represent just one without also representing all the others. The

evident fact that we *can* have distinct and independent thoughts about rabbits, rabbit stages, undetached rabbit parts, and many other necessary coincident things that are indistinguishable in terms of their effects thus seems to show that CTR slices representational contents too thickly.

In response to this objection, proponents of CTR might attempt to argue that thoughts about, e.g., rabbits and rabbit stages can be distinguished from one another on the grounds that the concept of rabbit stages contains the concept of rabbit as a part, and that one could hence not have thoughts about rabbit stages unless one was already capable of thinking about rabbits, even though one could have rabbit thoughts without being able to think about rabbit stages (in which case the ability of rabbit stages to cause representations of rabbit stages would be asymmetrically dependent on the ability of rabbits to cause rabbit representations) (Fodor, 1987, pp.86-7). Against this proposal, however, Gates (1996, p.335) is I think right to point out that "the fact that [a] property is picked out by a complex expression (e.g., of English) – even if it could not be picked out by any simple expression (also of English) – could not preclude its being the cause of simple concept tokens." In other words, there is no apparent reason why the concept of rabbit stage might not, for some people, be a simple concept that is just as or even more basic than their concept of rabbit (as would, e.g., be the case if one conceived of rabbits as coherent series of rabbit stages). If such is the case, though, then the proposed response is inadequate.

Another potential response to the objection just raised is offered by Fodor (1994, ch.3), who suggests that one might distinguish, e.g., between rabbit thoughts and rabbit stage thoughts on the grounds that one who thinks that *x* is a rabbit will *ipso facto* be

disposed to make different inferences therefrom than one who thinks that *x* is a rabbit stage. While this maneuver requires supplementing CTR with a measure of Conceptual/Inferential Role semantics, the addition called for is relatively modest, and can, if Fodor is correct, be confined to the logico-syntactic components of one's system of mental representations in such a way as to avoid wholesale semantic holism. Proponents of CTR can hence adopt Fodor's proposal as a minor modification, rather than a radical revision of their basic view. Even this injection of Conceptual/Inferential Role Semantics is, however, still not enough to inoculate CTR against the problems raised by necessarily coincident, causally equivalent *representata*. For as Gates (1996, p.343) shows, each time one finds a difference in the inferences that are licensed by or that one would be disposed to draw from thoughts about some "gavagaish" properties *X* and *Y*, further reflection will inevitably discover some other "gavagaish" property *Z* whose conceptual/inferential role is indistinguishable from *X* or *Y* at precisely that point where *X* and *Y* differ. Thus, while one might attempt to distinguish mental representations of rabbits from mental representations of rabbit stages by appealing to the fact that one who thinks that *x* is a rabbit would be more likely to infer from this that *x* existed prior to the present moment than one who thinks that *x* is a rabbit stage, one could not distinguish in this way between representations of rabbits and representations of undetached rabbit parts, or representations of rabbit stages and representations of momentary rabbit events, since someone who believes, e.g., that *x* is a rabbit would be just as likely to infer from this that *x* existed prior to the present moment as one who believes that *x* is a bunch of undetached rabbit parts (and likewise, *mutatis mutandis*, for those who believe that *x* is a rabbit stage and those who believe that *x* is a momentary rabbit event). Generalizing from

this example, it seems that whenever some inferential grounds are found for distinguishing between the content of thoughts about the necessarily coincident, causally indistinguishable properties *X* and *Y*, there will always be some further property *Z* (which is likewise necessarily coincident with and causally indistinguishable from *X* and *Y*) such that those grounds are inadequate to distinguish between our thoughts about *Z* and those about *X* or those about *Y*. As Gates (1996, p.343) puts it:

> [T]his method of constructing new gavagaish predicates will generalize to each purported solution of this sort…Gavagaish properties are legion. Trying to eliminate individually the possible alternatives is like trying to diminish an infinite pile of stones by removing one pebble at a time. And each time we succeed in identifying an acceptance and inference pattern which distinguished a property from one of its gavagaish alternatives, we reveal more pebbles in the pile.

In sum, then, even when supplemented with an appeal to conceptual/inferential roles, there appears to be no way for CTR to account for distinct mental representations of distinct but necessarily coincident and causally indistinguishable entities. Intentional content is, in short, too fine grained for CTR to be true.

4.ii. Teleological Theories of Representational Content

In light of the forgoing considerations, those seeking a naturalistic reduction of representational content may attempt to remedy the inadequacies of CTR by proposing that what it is for a physical state *x* to represent a certain property or object *F* is not, or not merely for it to stand in a certain causal relation to *F*, but rather for it to have the *biological function* of indicating *F* (where *x* indicates *F* iff *x* regularly covaries with or "carries information about" *F*). While those who adopt such Teleological theories of

content (TTR) continue to hold that representation rests on causation, the relevant causal relation is, on such views, not that (if any) which holds between a token representation and the thing it represents, but rather that which held between the ancestral forebears of the representation and earlier instances of the represented property or type. More specifically, under TTR, the fact that a given state *x* represents *F* is said to consist in the fact (a) that *x* is a token of some physical type *G* that indicates *F*, (b) that the capacity to produce tokens of *G* is a trait that has been preserved under some selective process (e.g. learning[135] or natural selection), (c) that this capacity has been selected for at least partly *because G* indicates *F*, and (d) that *x* was itself produced by such a capacity.[136] Although these conditions do not require that *x* itself be caused by *F* in order to represent it, the satisfaction of conditions (b) and (c) is naturally read as requiring past instances of *G*, at least, to have been caused by *F*s, for it is unclear how the fact that *G*s indicate *F* could have led to the selection of the capacity to produce tokens of *G* unless *G*s were at some point reliably caused by *F*s. However, just as the past causation of *G*s by *F*s needn't carry over to present tokens of *G* that represent *F*, so too a current capacity to produce tokens of *G* need no longer provide the evolutionary/biological advantage that led to its prior selection in order for the *G*s it produces to retain their status and function as representations of *F*. All that's needed is for that capacity to have been inherited or

[135] For reasons of simplicity, the following discussion proceeds largely on the assumption that the functions that TTR makes use of to account for representational content are generated by natural selection. The criticisms of TTR raised below nonetheless apply equally to versions of TTR, e.g. Dretske's (1988), which hold that learning is the *only* selective process capable of generating the functions that distinguish representations from mere indicators.

[136] Millikan (1989) might prefer to say that *x* represents the condition which it must correspond to in order for its "consumers" (i.e., those systems that make use of *x* to guide their operations) to function properly.

preserved due to the fact that it *did* provide such an advantage at some point in the individual or phylogenetic history of the being that now possesses it.

By analyzing representational content in terms of biological functions, TTR incorporates an element of normativity that is lacking in CTR, for whereas causal relations are, in themselves, strictly non-normative, things with biological functions can be normatively evaluated as functioning better or worse depending on whether they achieve or fail to achieve the ends for which they were selected. This fact enables TTR to account for the normativity of mental representations by equating the veridicality of a given representation with the degree of success that it and and/or the mechanisms that regulate its tokening achieve in performing their proper biological functions. According to TTR, it is hence *because* mental representations have the biological function of indicating some property or object *F* that they are subject to normative evaluation as being more or less accurate of their objects. As hinted above, this feature of their position makes it relatively easy for proponents of TTR to handle those cases involving "pathological" causation of mental representations that gave CTR such trouble. For whereas CTR seems unable (even when augmented by Fodor's Asymmetric Dependence condition) to provide any principled grounds for claiming that a perception of a goat caused by a blow to the head represents a goat and not the blow to the head, adherents of TTR can treat this as a simple consequence of the fact (a) that the state that serves as the representational content of the perception has the function of indicating goats, not blows to the head (for it is the covariation of such states with goats that has led to preservation of the capacity to produce them in the perceiver), and (b) that its having this function does not depend on its present cause. The faulty nature of the resulting perception can

likewise be attributed by TTR to the failure of the goat representing state to perform its proper goat indicating function, as there are presumably no goats in the perceiver's vicinity for the state to indicate the presence of, and even if there were, the state's tokening would not, on this occasion, have been caused by a goat, as it must in order for it to function in the right way (i.e., in the way that led to its selection as a goat indicator).

While TTR thus fares much better than CTR in dealing with "pathologically" caused representations (or indeed with any kind of case that likewise seems to require reference to what a given representation is *supposed* to represent), it is nevertheless vulnerable to the second objection to CTR raised above, concerning the fine-grained nature of representational content. This should come as no surprise, for although TTR does not require representations to represent their current causes, it still holds the content of any representation to be fixed by a purely causal process involving the causation of certain of its previous tokenings by some object or property *F* and the consequent selection of the capacity to produce such tokens due to their usefulness as indicators of *F*. As argued above, though, no set of purely causal relations is sufficient to differentiate between distinct mental representations of necessarily coincident, causally equivalent entities, e.g. rabbits and undetached rabbit parts. Hence, just as the fact that a given representation *x* is caused by a given object or property *F* is by itself inadequate to determine whether *x* represents *F* or some other necessarily coincident, causally equivalent property *G*, so too the fact that previous tokenings of *x* have regularly covaried with or been caused by *F*s in such a way as to lead to the preservation of the capacity to produce such *x*s is likewise inadequate to determine whether *x* represents *F* or *G*. For given that *x* covaries equally with *F* and *G* (and $F \vee G$), there is no way to tell which of

these candidate *representata x* was selected as an indicator of. As Gates puts it (1996, p.336 fn12): "Whatever has been 'selected for' its prowess as a horse-detecting mechanism has also been selected for its uncanny ability to detect undetached horse parts." In sum, since TTR follows CTR in explaining what a representation represents in terms of some causal relation between (previous tokens of) the representation and (previous instances or manifestations of) its *representatum*, it is ultimately no more able than CTR to account for the existence of distinct mental representations of necessarily coincident, causally equivalent entities. Both theories are thus incapable of individuating mental representations at a level of grain fine enough to match our own representational capacities.

A second problem for TTR arises from an apparent mismatch between the norms that pertain to biological functions and those that furnish the conditions of satisfaction for mental representations. As noted above, TTR is committed to the view that these norms are in fact identical, for according to TTR, the accuracy or inaccuracy of mental representations consists wholly in the successful or unsuccessful performance of certain biological functions, which implies that the norms that specify the conditions under which a given representation is accurate must be identical with the norms that specify the conditions under which some biological function qualifies as having been successfully performed. An obstacle to this proposed identification of representational with biological norms is, however, raised by Tyler Burge (2010, pp.292-308), who notes that whereas the successful performance of any biological function must always be such as to have made some prior contribution to the fitness of the organisms (or the ancestors of the organisms) whose parts or capacities possess it, it does not appear similarly constitutive of the

accuracy of any representation that its correct or incorrect mapping of the environment increase or decrease the fitness of the beings (or the ancestors of the beings) in which it is tokened. The resulting "mismatch" between representational and biological norms makes it easy to imagine scenarios wherein inaccurate representations end up *contributing* to, or accurate ones *detracting* from, the fitness of the organisms in which they are tokened (Burge, 2010, p.301). Consider, e.g., the case, presented by Burge (2010, p.302), of an animal whose "avoidance mechanism functioned to increase strength and agility – in avoiding [some] predator – even in cases in which the animal engaged in avoidance behavior, because of a misrepresentation as of a predator, when no predator was present." Even assuming that this animal's predator representations are consistently inaccurate, their being so is *not* detrimental to the successful functioning of its avoidance response mechanisms, for the overactive tokening of these representations actually ends up making the animal better able to avoid predators, thereby *improving* the ability of its avoidance response mechanisms to achieve their proper end and *increasing* the overall fitness of the animal as a whole. The possible existence of such a creature hence demonstrates that "[f]ailure of accuracy need not be failure to realize *any* biological function" (Burge, 2010, p.302).

While the accuracy/inaccuracy of mental representations must thus be distinguished from the successful/unsuccessful performance of any biological function, this does not controvert the fact that mental representations nevertheless may and no doubt *do* make some contribution to the fitness of the organisms in which they are tokened by causing behavioral responses that increase the organism's chances of survival and reproduction. Indeed, if the ability to produce mental representations did *not* make

some such contribution to fitness, then the preservation and proliferation of that capacity would appear entirely miraculous. It is therefore reasonable to assume that mental representations have in fact been selected for their ability to cause certain evolutionarily advantageous responses, which they consequently now have the biological function of producing. Their possession of such functions remains, however, a separate matter from their status as representations, for as illustrated by the case discussed above, a representation can succeed in performing the function for which it was selected (viz. that of producing evolutionarily advantageous responses) while failing to accurately represent the environment. In short, since it is only by virtue of their non-semantic, causal features that mental representations make any real contribution to fitness, it is their causal, rather than their semantic features that they have been selected for (Burge, 2010, p.302). Hence any biological function that a mental representation has must be concerned solely with the fitness of the responses it causes, rather than with the representational accuracy of its tokens. And thus it is that the success of a representation in performing its biological function is largely tangential to its veridicality.

If what Burge says is true, then TTR is fundamentally false. What it is for a state to represent something cannot be merely for it to have a certain kind of biological function if the selection process that generates such functions is indifferent to representational accuracy. But as remarked above, the only differences that natural selection is sensitive to are differences in the contributions that various traits make to an organism's fitness, and there is no necessary correlation between such differences and the differences between accurate and inaccurate representations. Put simply: "Evolution does not care about veridicality. It does not select for veridicality *per se*" (Burge, 2010, p.303).

To which we might add, that veridicality likewise does not care about evolutionary fitness, or the successful performance of biological functions. In which case, *contra* TTR, the representations that serve as the content of intentional states cannot just be states that have the biological function of indicating their *representata*. For representations, when successful, are accurate, and nothing can be accurate simply by virtue of successfully performing some biological function.

4.iii. Phenomenal Intentionality

In light of the forgoing criticisms, it appears that the two leading naturalistic theories of representational content, CTR and TTR, both fail in their attempts to analyze the representational content of intentional states in terms that are strictly confined to the vocabulary and explanatory resources of the natural sciences. While the objections raised to these theories have thus far been largely negative, there is another, positive proposal that poses further problems for any functionalist reduction of intentionality by making such reduction contingent upon the prior reduction of phenomenal states. The core of this proposal, which has been advanced by Terence Horgan and John Tienson (2002) and Brian Loar (2003) under the title of Phenomenal Intentionality (PI), is that the representational content of (at least some) intentional states is (at least partly) determined by an associated phenomenology. If this proposal is correct, then what a given belief, thought, or desire is about may ultimately depend upon the phenomenal character of the experience of the individual that is currently in that intentional state. Thus stated, PI hence serves as a complement or foil to the Representationalist position discussed

above.[137] Whereas the latter seeks to ground the phenomenal character of experiences in their representational content, according to PI, it is rather the representational content of certain intentional states that must be viewed as dependent upon the phenomenal states with which they are associated. And whereas Representationalism is often employed by physicalists to make phenomenal states more amenable to reduction by assimilating them to intentional states (which are widely held to present fewer difficulties in this regard than qualia), PI, in contrast, makes the task for physicalists much harder, for if PI is true, then no reduction of intentional to physical states can be regarded as complete unless the phenomenal states that help fix their representational content have been reduced to physical states as well.

While attempts to provide a fully worked out theory of PI are at this point still in their initial stages, there are nevertheless at least two considerations that seem to suggest that the core thesis of PI is basically correct. The first of these consists in the introspective observation that our experiences (particularly our perceptual experiences) seem to exhibit a certain "directedness," whereby they purport to refer to beyond themselves to certain objects, events, and properties out in the world. The following thought experiment, proposed by Loar (2003, p.239), helps bring this feature of our experience to light:

> Suppose some indistinguishable lemons are one after the other brought to my visual attention. The lighting, the position of my eyes, and so on, are held constant. I am asked to think something about each lemon in turn, say 'that's yellow'. Afterwards I am told that some of the apparent lemons were hallucinations…I am asked whether, despite this, my successive

[137] While Loar (2003, pp.238-9) and Horgan and Tienson (2002, p.520) both explicitly reject Representationalism, they do nonetheless hold that there are no "'purely' qualitative, that is, in themselves non-intentional" aspects of experience (Loar, 2003, p.238).

> visual demonstrative thoughts all visually presented their objects in the same way. Surely a natural reply is yes…

The similarity among the thoughts that warrants this reply is, as Loar notes (2003, p.239), "an *intentional* feature. For [the] demonstrative concepts [the thoughts employ] (both the ones that succeed in referring and the ones that do not) all purport to pick out some object visually." What is it, then, that accounts for the fact that these successive thoughts all share the same intentional feature? It cannot be any external object or property out in the world that the thoughts all refer to, for in many cases the putative referent of the thought does not exist, so some of the thoughts are not referentially related to anything at all. The answer must therefore instead lie in some "nonrelational phenomenal feature" present on the occasion of each thought, for aside from the intentional feature that the various thoughts share in common, it is the internal, phenomenal situation that alone remains the same in each case (Loar, 2003, p.239). In sum, as we "apparently can tell that hallucinatory experiences have a 'purporting to refer' property that is also present when [phenomenally indiscernible] visual experiences pick out real objects in the normal way," it seems to follow that this intentional directness that we are able to identify in our experiences must somehow be rooted in their internal, phenomenal features, for these features appear to be the only thing (apart, again, from the directedness just mentioned) that phenomenally indistinguishable veridical and hallucinatory experiences have in common, and hence the only thing that could explain their intentional similarity (Loar, 2003, p.240).

A second, more speculative source of support for PI can be found by adopting the perspective of a brain in a vat whose experience is phenomenally indiscernible from

one's own (Horgan and Tienson, 2002, pp.524-7; Loar, 2003, pp.246-7, 250-1). While the evaluation of such wild hypotheticals relies heavily on intuitions, which can at best offer only defeasible, *prima facie* evidence for any claim, it nevertheless bears noting that intuitions in this case seem to weigh heavily in favor of the view that there is something that the representational contents of my own intentional states and those of my B.I.V. twin share in common. For quite generally, the world as represented by my own perceptions would appear to be no different from the world as represented by the (mis)perceptions of my B.I.V. twin. If it appears to me, e.g., that I am currently sitting courtside with Chewbacca at a Knicks game eating alligator sauce out of a fishbowl, the same must surely be true of my B.I.V. twin. The only difference between the contents of our intentional states seems to be that my twin's perceptually based representations are radically inaccurate, whereas mine are by and large correct. The fact remains, however, that the world as represented to us by our respective intentional states is the same. But if this is so, and the content of my intentional states is really no different from that of my B.I.V. twin's, then there seems to be no natural alternative to the conclusion that intentional content is at least partly dependent upon phenomenology, and that PI is therefore true.

The main objection to this line of reasoning is that it seems to stand in open conflict with the widely held Externalist theory of reference, meaning, and mental content persuasively argued for by Kripke (1980), Putnam (1975), and Burge (1979), according to which the referents of many important terms and concepts (e.g., proper names and singular concepts, natural kind terms/concepts, and "socially deferential" terms/concepts) are determined by certain causal or other contingent relations between the person

employing the term or concept and his/her natural environment and linguistic community. The content of statements or thoughts containing such terms or concepts is thus said by Externalists to be "wide," meaning that it does not depend merely on the internal state of the speaker or thinker, but also on his/her external surroundings. This sets up a conflict with PI, for (a) according to PI, there is a certain kind of intentional content (viz. that which is shared by me and my B.I.V. twin) "that constitutively depends on phenomenology alone," and (b) it seems relatively clear that phenomenal states, at least, "depend only on narrow factors," i.e., factors that are internal to the individual, such as the current state of their nervous system (Horgan and Tienson, 2002, p.527). From this it follows, however, that many, if not all intentional states have a form of content that is *not* wide, but strictly *narrow*. As Horgan and Tienson (2002, p.527) put it:

> [T]he theses of phenomenal intentionality and the narrowness of phenomenology jointly entail that there is a kind of *narrow* intentional content…pervasive in human life, such that any two creatures who are phenomenal duplicates must also have exactly similar intentional states vis-à-vis this kind of narrow content.

PI thus seems to be committed to an Internalist view of mental content, according to which the representational content of many intentional states is fixed independently of any factors external to the individual that is in them.

The problem for PI, then, is that Externalism is supported by a number of familiar thought experiments that tell (or are at least widely taken to tell) rather conclusively in its favor. If acceptance of PI should require rejecting the standard reading of such thought experiments, then the prospects for PI would appear rather bleak. Fortunately for its advocates, however, there is no need for PI to be saddled with any such requirement, for while PI does require the rejection of an unrestricted Externalism about mental content

(i.e., the position that *all* content is wide, and *none* of it narrow), PI is fully consistent with the view that certain intentional states may possess both wide and narrow content, which is a position that is likewise consistent with standard readings of Externalist thought experiments. A simple way of developing such a view, which is suggested by both Horgan and Tienson (2002, pp.527-9) and Loar (2003, pp.253-4), is to distinguish between reference and purported reference, or more specifically, between what (if anything) the content of an intentional state actually refers to, and what it represents by virtue of the directness that is grounded in its associated phenomenology. The wide content of an intentional state can then be identified with its actual referent (assuming it has one), this being determined in the typical Externalist manner by certain relations between the subject of the state and his/her physical and social environment, whereas its narrow content is identified with what the state merely *purports* to refer to, this being determined *not* by any relations to anything outside the state's subject, but instead solely by the phenomenal character of the subject's concurrent experiences. Since the experiences that each of us associates with what we respectively conceptualize as water would be phenomenally indiscernible, my own thoughts about water would thus possess the same *narrow* content as those of my B.I.V. twin and a counterpart of me living on Putnam's (1975) Twin Earth.[138] The *wide* contents of our respective thoughts, however, would differ, for whereas my water thoughts refer to H_2O, those of my Twin Earth counterpart refer to XYZ, and those of my B.I.V. twin either don't refer at all, or else

[138] *Contra* Millikan (1984, p.93) and other advocates of TTR, the same would also be true of my own thoughts about water and those of a Davidsonian (1987) "Swampman" duplicate of me.

they refer to some pattern of electrical activity in the circuitry of the supercomputer to which he is hooked up.[139]

As there is nothing obviously incoherent or wildly implausible about such a position, it appears that PI can be made consistent with the Externalist intuitions brought out by the thought experiments of Kripke, Putnam, and Burge, while at the same time preserving the Internalist intuition that the contents of the intentional states of phenomenal duplicates must share something in common. Externalist thought experiments therefore cannot be used to reject PI out of hand. But if the position just outlined is correct, and there is in fact a form of intentional content that constitutively depends on narrow, phenomenological factors alone, then, as Horgan and Tienson (2002, p.521) point out, "theories that ground all intentionality in connections to the external world, such as causal and teleological theories of intentionality, are deeply mistaken." For it clearly cannot be a necessary condition for one thing's representing another that the former be causally related to the latter in some way (as CTR maintains), or that the former have been selected for its usefulness as an indicator of the latter (as *per* TTR), if there exist certain representational states whose content is determined solely by the phenomenal experience of the subject who is in them. PI thus threatens to undermine any attempt to functionally reduce intentionality not only by making the success of any such project dependent upon a prior reduction of phenomenal states, but also by demonstrating

[139] As noted by Horgan and Tienson (2002, p.528 fn26), this distinction between two types of content shares certain similarities with the 2-dimensional semantics developed by Chalmers (1996) and Jackson (1998), which might indeed be useful in elucidating the satisfaction conditions for wide and narrow content respectively (the former being determined by the secondary or C-intension of a given thought or sentence, and the latter with its primary or A-intension). This does not mean, however, that advocates of PI must also accept Chalmers' (1996) use of 2-D semantics to support an inference from conceivability to metaphysical possibility.

the inadequacy of any account of representational content (e.g. CTR and TTR) that assigns the determination of such content solely to factors external to subjective experience. The plausibility of a physicalist theory of intentionality being thus inversely related to the plausibility of PI, it is difficult to see how any such theory can be made compelling unless the arguments for PI can first be shown to be unsound.

Our evaluation of the two leading attempts to functionally reduce intentional properties to physical properties has uncovered a variety of difficulties that suggest that the prospects for such a reduction are rather dim. While the various intentional attitudes do seem susceptible to functional analysis, the representational content that is equally essential to the individuation of intentional properties (and arguably constitutive of their intentionality) cannot, it seems, be fully analyzed in purely causal or biological terms. The attempt to establish T/NRPism for intentional states by way of functionalist reduction hence faces the problem that the contents of such states cannot be equated with any functions that physical events seem capable of performing. And if the contents of intentional states cannot be functionally reduced in this manner, then neither of course can intentional states themselves. Putting this together with the results of our discussion of phenomenal properties in section 3 above, it thus seems that neither of the two types of properties that are distinctive of mentality can be functionally reduced to physical properties. The functionalist route to T/NRPism has therefore proven no more successful than the attempt to argue for T/NRPism on the general, metaphysical grounds proposed by Davidson (1970). Adding to this the worries raised in section 1 as to the very

coherence of T/NRPism, it would seem best at this point for us to set the position aside and explore other options.

Where does this leave us, then, with respect to (4*) Mind-Body Dualism? Having argued, in the previous two chapters, that mental properties cannot be type-identified with physical properties due to their multiple realizability, and rejected, in the present chapter, the weaker view that all instances of mental properties are token-identical with physical events, the only option remaining for those who insist on rejecting (4*) is to embrace eliminativism. For with both type- and token-physicalism off the table, the only position left that is consistent with the claim that there are no non-physical mental properties and events is the view that there are no mental properties and events at all. While such eliminativism does have its proponents[140], one might reasonably think that if acceptance of physicalism requires the loss of one's mind, then the cost of physicalism is simply too high. Of course no one can foresee what shocking discoveries future neuroscience has in store, and the possibility remains that our understanding of the brain will someday reach a point where explanations of animal behavior in terms of beliefs, desires, sensations, and feelings are superseded by explanations stated in terms of various sorts of neural activity, but to maintain that this result is inevitable or even probable is to place a large bet on long odds. At any rate, as things now stand, belief in the existence of mental properties is I take it more than adequately justified by our introspective access to *qualia*, and the success, reliability, and continued fruitfulness of psychological explanations of animal behavior. More could surely be said in support of realism about the mental, but rather

[140] See, e.g., P.M. Churchland (1981), Stich (1983), P.S. Churchland (1986), and Dennett (1988).

than embarking on a lengthy argument in favor of a position that few will see much reason to question, let alone deny, it would perhaps be better to follow the advice contained in Hume's observation that "Next to the ridicule of denying an evident truth, is that of taking much pains to defend it" (*Treatise* I.3.xvi., as quoted by Molnar (2003, p.99)).

Assuming, then, that the basic claims of the past three chapters are correct, and mental events can neither be eliminated nor token- or type-identified with physical events, we seem to have no choice but to allow that mental properties and their instances are distinct from, irreducible to, and unexplainable in terms of the physical events on which they depend. If this is indeed the case, then the Exclusion Problem cannot be satisfactorily resolved by simply rejecting (4*) Mind-Body Dualism. Accepting this conclusion consequently places us under an obligation to provide some other explanation for how the apparent causal efficacy of our mental states (1) can be reconciled with the apparent truth of (2) the Causal Self-Sufficiency of the Physical and (3) the Absence of Systematic Overdetermination. As the physicalist alternatives to (4*) all seem impracticable, the seeming conflict between (1), (2), (3), and (4*) must be resolved in some other way, either by rejecting (1), (2), and/or (3), or else by showing that the conflict is illusory. The remaining three chapters explore each of these options with the aim of showing that the problems involved in the rejection of (1) and/or (4*) are far more substantial than those that are thought to arise from the rejection of (2), (3), and/or the alleged inconsistency between (1), (2), (3), and (4*), and that the Exclusion Problem is therefore much more easily and plausibly resolved by denying (2) the Causal Self-Sufficiency of the Physical, (3) the Absence of Systematic Overdetermination, or the

incompatibility of (1), (2), (3), and (4*), than by denying either (4*) Mind-Body Dualism or (1) the Causal Efficacy of the Mental.

CHAPTER 7

MENTAL CAUSATION WITHOUT OVERDETERMINATION

In light of the challenges that the previous three chapters have raised for any physicalist theory of mind, it seems reasonable to consider mind-body dualism as an alternative that is at least worth exploring. Assuming, however, that the multiple realizability of mental properties, the resistance of qualia to functional analysis, and the difficulties facing any functionalist reduction of intentionality all lend credibility to the view that mental events are not identical with, reducible to, or fully explainable in terms of physical events, it remains to show that the former can, on such a view, still be said to cause the latter without thereby generating a metaphysical picture that is either incoherent or wildly implausible. In short, we have yet to establish that dualists can reasonably reject epiphenomenalism. Here, again, the central difficulty (raised by the Exclusion Problem) is that any non-epiphenomenalist form of dualism seems flatly inconsistent with the conjunction of two theses that both seem independently plausible, viz.:

(2) *Causal Self-Sufficiency of the Physical:* Every physical effect has a sufficient physical cause; and

(3) *Absence of Systematic Overdetermination:* Causal overdetermination is rare.

By taking the arguments adduced in the preceding chapters as grounds for accepting dualism, we thus incur an obligation to either provide some reason for thinking that the conjunction of (2) and (3) is false or compatible with the view that the mind is both non-physical and causally efficacious, or else accept epiphenomenalism as well. The burden

of the remaining three chapters will be to show that this obligation can be discharged without taking the epiphenomenalist way out.

These last three chapters are organized as follows: The present chapter points out the difficulty of providing a precise formulation of (3) that is immune to counterexamples, notes some problems that rigorous adherence to (3) creates when combined with the view that all causal power resides at the level of fundamental physics, and develops a strategy that dualists might employ to argue that if causal overdetermination is rare, then effects with distinct and independently sufficient physical and mental causes needn't be overdetermined. The following chapter notes some problems involved in finding a formulation of (2) that is strong enough to pose a problem for interactionist dualism without begging the question against it, evaluates the support that the thesis derives from conservation laws, and considers some potential objections and counterexamples to the principle that all physical effects have sufficient physical causes. The final chapter then builds on the results of the preceding chapters to provide two accounts of mental causation that can avoid the Exclusion Problem while remaining consistent with dualism.

1. Epiphenomenalism

Before commencing our discussion of (3), it is perhaps worth adding a few words on the subject of epiphenomenalism, for one might naturally wonder why dualists should go to such lengths to account for mental causation when the problems they face in doing so can be much more easily resolved by simply denying that there is any such thing.

Given the multitude of social practices that rest on our collective pre-theoretical belief in the mind's causal efficacy, and the centrality this belief has to our normal conception of ourselves as human agents, such a move is likely the most counterintuitive of the possible answers to the Exclusion Problem. That said, a number of contemporary dualists (e.g., Chalmers (1996), and (formerly) Jackson (1982)) have nonetheless been drawn towards epiphenomenalism as a way of avoiding conflict with (2) and (3). Support for this maneuver might be drawn from certain scientific studies (including, most famously, the experiments of Benjamin Libet (1985)), which seem to suggest that the neural activity that initiates voluntary behavior actually occurs approximately 350 ms *prior* to any conscious decision to act. However, even if these studies (which many, including Libet himself, have noted can be interpreted in ways that do *not* entail epiphenomenalism[141]) can help quiet the protests that epiphenomenalism raises from untutored folk intuitions, the epiphenomenalist still faces another problem: For the vast majority of cases, our knowledge of concrete events seems to require that there be some causal relation, however indirect, between the object of our knowledge and a certain true belief we have about that object. Hence, if mental events are causally inert, it becomes a mystery how we know anything about them (including the alleged fact that they are causally inert!).

A further problem for epiphenomenalism is raised by the fact that since selective pressure is exerted only on traits that actually *do* something, if mental states are causally inert, it becomes difficult to explain how or why they evolved and have since been preserved under natural selection. The fact that mental states have evolved not only in our

[141] See, e.g., the remarks of Breitmeyer, Eccles, Jung, Latto, Näätänen, and Stamm in the Open Peer Commentaries on Libet's paper, as well as McCall (2013).

species, but in a significant portion of the animal kingdom, would seem to suggest that such states make some contribution to fitness by enabling minded individuals to do certain things that mindless beings cannot. This is because the simplest explanation for the apparent prevalence of mentality across different species of animals is that mental properties bestow certain causal powers on their bearers that (in certain species, at least) made those individuals that instantiated such properties more likely, on the whole, to survive and reproduce than those that didn't.

While these problems are perhaps not insurmountable[142], taken in conjunction with the strong resistance that epiphenomenalism faces from common sense, they nonetheless seem substantial enough to shift the burden of proof onto the epiphenomenalist and compel him/her to provide some positive reason for thinking that the mind is indeed causally inert. The standard motivation for accepting epiphenomenalism is, however, just the worry that the causation of physical effects by non-physical, mental events is somehow objectionable, either because such causation is superfluous (since every physical effect already has a sufficient physical cause), or because such causation is simply incomprehensible. If the remaining three chapters of this dissertation are successful, though, these concerns will be shown to be groundless, and we will consequently be left with no compelling reason to deny that the mind is causally efficacious. The remaining three chapters are thus just as much an argument

[142] In response to the first difficulty, the epiphenomenalist might appeal to certain reliable, non-causal covariations between our mental states and the beliefs we have about them, or postulate some non-causal, *sui generis* relation of "acquaintance" to serve as the requisite epistemic link between us and our minds. In response to the second, the epiphenomenalist could suggest that mental properties are "spandrels" that themselves do not contribute to fitness, but are nonetheless "preserved" under selection because they happen to supervene on non-mental states that do.

against epiphenomenalism as they are an argument for the viability of interactionist dualism, for given the various problems that epiphenomenalists must deal with, once interactionist dualism has been shown to be viable, epiphenomenalism loses much of its appeal.

2. Causal overdetermination

Of the four propositions that the Exclusion Problem presents as incompatible (viz., (1) the Causal Efficacy of the Mental, (2) the Causal Self-Sufficiency of the Physical, (3) the Absence of Systematic Overdetermination, and (4*) Mind-Body Dualism), it appears, then, that there are good reasons not to reject either (4*) Mind-Body Dualism or (1) the Causal Efficacy of the Mental. Of the two propositions that remain, many have fixed on (3) the Absence of Systematic Overdetermination as the source of the trouble. Those who favor this diagnosis of the Exclusion Problem might set about resolving it either by rejecting (3) as false, or else by denying that (3) is in fact inconsistent with the conjunction of (1), (2), and (4*). The solution to the Problem advanced in the remainder of the present chapter combines these two approaches by arguing that if overdetermination is defined in such a way that effects with distinct, independently sufficient mental and physical causes *ipso facto* qualify as overdetermined, then (3) is likely false, and dualists can hence freely reject it and retain (1) and (4*) without having to call (2) into question. On the other hand, if overdetermination is defined in such a way that (3) is most likely true, then distinct, independently sufficient mental and physical causes needn't overdetermine their joint effects, and one can hence

endorse (1), (2), (3), and (4*) without contradiction. If successful, this strategy will hence show that the conception of overdetermination under which (3) seems most plausible is likewise one according to which (3) is compatible with the conjunction of (1), (2), and (4*), thus providing us with a way of overcoming the Exclusion Problem without rejecting dualism or embracing epiphenomenalism.

2.i. Is causal overdetermination rare?

The idea that it is rare for an effect to have two or more distinct, sufficient causes is often motivated by appeal to certain central examples of overdetermination that have a markedly unusual, or accidental character. Take, e.g., the death of a person shot by a firing squad. Assuming that more than one of the shots fired by the members of the firing squad would have been enough, by itself, to cause the victim's immediate death, the latter event is clearly overdetermined. And while the overdetermination in this case is certainly no accident (since the victim had presumably been sentenced to die in precisely this way), such occurrences are nonetheless rather unusual, as people are generally not killed by more than one thing at a time. Consider, now, what it would be like if such overdetermination happened quite regularly, so that quite often when an effect occurred, a number of distinct causal processes sufficient to produce that effect could be found that were perfectly timed and coordinated so as to terminate at the very moment and location of the effect's occurrence. Taking firing squad cases as our guide, such a situation would seem like a kind of cosmic coincidence. And since postulation of coincidences is generally to be avoided in one's theorizing, attention to paradigmatic examples of

overdetermination that occur only by freak accident or in highly unusual circumstances might thus easily be taken to show that overdetermination happens only rarely.

The problem with this sort of argument for (3) is that, according to the standard definition of causal overdetermination as the production of an effect by two or more distinct, independently sufficient causes, there are a large swath of cases that seem to qualify as instances of overdetermination, but whose occurrence is neither rare, unusual, nor coincidental in any way. Consider, e.g., an anvil dropped on an egg. Is the egg's breaking causally overdetermined? By the standard definition, it would seem so.[143] For if the anvil was dropped on the egg, so too were its top and bottom halves, and if the former's being dropped on the egg caused the egg to break, it seems that the same must be said of the latter as well. Since either of these parts of the anvil would, however, have been by itself sufficient to cause the egg to break if dropped on the egg alone, it seems that the egg's breaking has at least three distinct, independently sufficient causes (viz., the anvil's being dropped on it, the anvil's top half being dropped on it, and the anvil's bottom half being dropped on it). While it is of course not often that an anvil gets dropped on an egg, occurrences of the same general type are certainly not rare or unusual. They can be found whenever a composite produces an effect that certain of its parts could have caused on their own. Nor is there anything coincidental or odd about the overdetermination that such cases involve, for considering how the causal powers of a

[143] Consider also a causal chain wherein an event c_1 is sufficient to cause another event c_2 that is in turn sufficient to cause a certain effect e. Here the effect e would seem to qualify, under the standard definition of overdetermination, as overdetermined by the distal and proximal causes c_1 and c_2. Such cases are also neither rare, unusual, nor coincidental in any way, and thus seem to present another counterexample to (3).

whole depend on those of its parts, it should come as no surprise that many of the effects caused by a whole could have been produced by certain of its parts by themselves.[144] Reflection on such instances of overdetermination thus suggests that overattention to firing squad type cases may have the misleading effect of making overdetermination appear much rarer and more unusual than it actually is. If this is correct, and (3) is in fact false, then the Exclusion Problem can be resolved without abandoning either (1), (2), or (4*), for the fact that the conjunction of these three propositions entails that every physical effect of a mental cause is overdetermined would not seem to pose a problem if overdetermination is indeed as widespread and routine as anvil-dropping type counterexamples to (3) indicate.

To ward off such counterexamples, adherents of (3) must either find some way to deny that the effects in such cases indeed qualify as overdetermined under the standard definition of overdetermination, or else modify their definition of overdetermination so as to apply only to cases of the firing squad variety, while excluding cases of the anvil-dropping sort. The best chance at implementing the former strategy would seem to be to deny that the allegedly overdetermining causes in the proposed counterexamples to (3) are indeed distinct. One might thus refuse to countenance the egg's breaking as being overdetermined by the impact of the anvil and the impact of the anvil's top or bottom half on the grounds that the latter are not truly distinct from the former, for the very reason that they help constitute or compose it. The problem with this proposal, however, is that it is unclear what it could mean for two things to be distinct if it is not simply for them to be

[144] See Sider (2003, pp.722-3).

non-identical, and understood in this way, there are few better assurances of the distinctness of two things than a difference in their respective persistence conditions and/or causal powers. Given then that the persistence conditions and causal powers of any part of the anvil differ from those of the anvil as a whole, it seems that the anvil must be distinct from its parts. Any effect caused by the anvil that certain of its parts could have also caused on their own would therefore seem to qualify under the standard definition as overdetermined.

Another, more extreme way of denying that the apparent counterexamples to (3) qualify as instances of overdetermination under the standard definition is simply to deny that all or all but one of the allegedly overdetermining causes in such cases actually exist or have the causal efficacy ascribed to them. In response to the anvil-dropping case, one might hence argue that the egg's breaking isn't causally overdetermined by its being struck by the anvil and the anvil's bottom half, either because (a) there are no such things as anvils, or anvil parts, or eggs for that matter, since the only things that exist are elementary particles, the fundamental physical properties they bear, and the fundamental physical forces that govern their behavior, or because (b) while macro-physical objects like eggs, anvils, and anvil parts exist, the only things that have any causal efficacy are the elementary particles of which such entities are composed. Instead, then, of saying that the egg's breaking is caused both by its being struck by the anvil and by its being struck by certain of the anvil's parts, we ought rather to say that the egg's breaking (or the collection of quantum events that we conceptualize as the egg's breaking) is caused solely by a collection of quantum events that compose or are conceptualized as the dropping of the anvil and its macro-physical parts.

Although some have endorsed similar views in print[145], the eliminativist/ epiphenomenalist consequences of (a) and (b) raise such difficulties for any realist interpretation of the special sciences that it is hard to think of an argument for (a) or (b) compelling enough to justify the acceptance of either.[146] For if all causal efficacy is confined to the level of fundamental physics as (a) and (b) would have it, not only must all mental states (barring panpsychism) be treated as epiphenomena; all macro-physical entities and micro-physical entities at the molecular and atomic levels must be stripped of their causal powers as well. This poses a major problem for the interpretation of theories in every scientific discipline besides basic physics, for interpreted realistically, any theory that attributes causal powers to entities that do not appear at the fundamental physical level would have to be regarded by adherents of (a) or (b) as strictly false. The vast majority of our current best theories at the atomic, molecular, biological, psychological, and social levels would consequently have to either be tossed out, or retained as merely instrumentally useful. Note also that having denied the existence and/or causal efficacy of all scientific kinds besides those of basic physics, proponents of (a) and (b) cannot preserve the legitimacy and success of the causal explanations employed in the non-

[145] See, e.g., Unger (1979), van Inwagen (1990), and Merricks (2001).

[146] Moreover, even disregarding the apparent incompatibility of such views with scientific realism, the arguments that Unger, van Inwagen, and Merricks offer in support of their favored forms of eliminativism remain open to criticism on other grounds. The Sorites-style arguments employed by Unger (1979) can be undermined by adopting certain theories of vagueness. In response to van Inwagen's (1990) contention that there is no satisfactory answer to the question of what conditions must be satisfied in order for different things to compose an additional object (unless that object is a living being), one might suggest that while standard answers to this question may have problems, there is no apparent reason why science couldn't eventually provide a more satisfactory answer in terms, e.g., of chemical bonding. The Exclusion-style argument used by Merricks (2001) rests on the contentious assumptions that (a) macro- and micro-physical objects overdetermine their joint effects, and (b) that all forms of causal overdetermination are objectionable. The second of these assumptions is criticized effectively by Sider (2003).

fundamental sciences by simply reinterpreting them as "program explanations," in the manner of Jackson and Pettit (1990a; 1990b), for as was shown in Chapter 2, such explanations are inadequate substitutes for explanations backed by real causal relations between their *explanantia* and *explananda*. In short, any view that locates all being and/or causal power at the fundamental physical level seems committed to rejecting the potential truth of virtually all contemporary scientific theories and the explanations they support. Such a result is, it seems to me, sufficiently extreme to render any view suspect that has it as a consequence.

The idea that all existence and/or causal efficacy resides at the fundamental physical level also faces the further difficulty that the existence of a fundamental physical level is itself an open question. As Block (2003) points out, it may be that each time we seem to hit upon a fixed set of fundamental particles and forces (such as those that make up the current Standard Model), further investigation will reveal these particles and forces to be further divisible into or dependent upon some more basic set of particles and forces, which will in turn be later broken down into some even more basic entities, *ad infinitum*.[147] The history of science itself would in fact seem to lend some inductive support to this possibility, as micro-physical entities that were once thought to be fundamental have often later been found not to be so. At any rate, the idea that nature has no bottom is surely at least as coherent and plausible a hypothesis as the conjecture that

[147] See also Sider (1993). While Merrick's (2001, pp.115-6) brand of eliminativism is, as he notes, "consistent with matter's infinite divisibility," his principles enable him eliminate macro-entities in such conditions only if one also assumes either (a) that microscopic wholes have non-redundant causal powers, whereas macroscopic wholes "cause only what their parts cause," or (b) that systematic macroscopic overdetermination is objectionable in a way that "systematic microscopic overdetermination" is not. Both these assumption seem rather *ad hoc*.

only fundamental physical entities exist or have causal powers. Should this hypothesis turn out to be true, though, then the latter conjecture would entail either that nothing exists or that the things that do exist don't cause anything. Both of these conclusions are patently absurd. (a) and (b) can thus be seen to run up against the same problem of causal drainage that was raised in criticism of Jackson and Pettit's (1990a; 1990b) attempted solution to the Exclusion Problem in Chapter 2.[148] Considering, then, the significant difficulties involved in the restriction of existence and/or causal efficacy to the level of fundamental physics, it seems reasonable to conclude that such a maneuver fails to provide a satisfying defense of the claim that causal overdetermination, as standardly defined, is rare.

The issues (3) raises when interpreted according to the standard definition of overdetermination as the production of an effect by two or more distinct, independently sufficient causes also bear on a certain response to the Exclusion Problem that one often encounters in the literature.[149] This response points out that if every physical effect has a sufficient *fundamental* physical cause, as (2) (the Causal Self-Sufficiency of the Physical) seems to intend, then the Exclusion Problem raises difficulties not only for mental

---

[148] Kim (1998, pp.77-87, 112-20; 2003, pp.167-76) attempts to block causal drainage by appealing to the notion of "micro-based properties"; i.e., macro-properties of an object that "[tell] us what sorts of micro-constituents the object is made up of and the structural relations that configure these constituents into a stable object with substantival unity." He suggests that the powers of macro-physical objects are in no danger of draining away, because many of the macro-physical properties we treat as causally efficacious are "construable as" micro-based, and micro-based properties are simply *identical* with (and therefore have the same causal powers as) the mereological configurations of micro-entities that they characterize. Block (2003, pp.145-50), however, notes that the identification of micro-based properties with mereological configurations of lower-level entities runs up against the problem that some macro-level properties may be "micro-based in *alternative ways*." (See also Walter (2008, pp.689-92), Noordhof (1999), Gillett and Rives (2001), and Bontly (2002, pp.82-90).)

[149] See, e.g., Fodor (1989, pp.60-3), Baker (1993, pp.86-90), and Bontly (2002).

causation, but for all non-fundamental physical causation as well. For if every physical effect produced by a non-fundamental physical event already has a sufficient cause at the fundamental physical level, then such effects will all qualify, according to the standard definition, as overdetermined. If, however, such overdetermination is as rare as (3) avers, then the apparent frequency with which causation occurs at the atomic, molecular, and macro-physical levels poses a significant problem, which can be resolved only by either identifying all physical events with (aggregates of) events at the fundamental physical level (thereby adopting the view that the only physical events that exist are those that occur at the fundamental level), or else treating all non-fundamental physical entities as causally inert. When combined with the idea that events at the fundamental physical level are by themselves sufficient to account for everything that happens in the physical world, adherence to (3) (where (3) is interpreted according to the standard definition of overdetermination) thus leads to a position similar, if not identical to one of the two positions ((a) or (b)) whose infeasibility was just shown. If the Exclusion Problem indeed generalizes in this way to *all* non-fundamental causation, both mental and physical, one might begin to suspect that if there's anything that the Exclusion Problem demonstrates, it is not that dualism faces special problems in accounting for mental causation, but rather that there is something wrong with the assumptions whereby the Exclusion Problem suggests that it does. For these same assumptions lead quite naturally to the rejection of all causation aside from that which occurs at the fundamental physical level, and one might reasonably think that any set of premises that problematizes such a large swath of our causal talk must itself be problematic.

2.ii. What must causal overdetermination be in order for it to be rare?

If (3) is to be made defensible, it seems then that a more restrictive definition of overdetermination will have to be introduced that will enable us to label firing squad type cases as instances of overdetermination without having to classify effects produced by both fundamental and non-fundamental causes, or by a whole and certain of its parts, as overdetermined as well. With such a definition in place, the apparent counterexamples to (3) would be summarily dealt with, for since they would no longer qualify as instances of overdetermination, the frequency and regularity of their occurrence would be perfectly compatible with overdetermination's being rare. Any worries about the generalization of the Exclusion Problem to non-fundamental physical causation would likewise be allayed, for if effects do not count as overdetermined simply by virtue of having both fundamental and non-fundamental causes, then the apparent frequency of macro-physical causation and the causal dependence of all physical events on events at the fundamental physical level would not stand in conflict with the idea that overdetermination is rare.

How, then, might the standard definition of overdetermination be modified so as to achieve these desired results? A potential answer can be found in Karen Bennett's (2003, p.476) suggestion that in order for an effect *e* to qualify as causally overdetermined by two of its causes c_1 and c_2, the following counterfactual statements must be nonvacuously true:

> (O1) if c_1 had happened without c_2, *e* would still have happened: $(c_1 \& \sim c_2) \Box\!\!\rightarrow e$, and
>
> (O2) if c_2 had happened without c_1, *e* would still have happened: $(c_2 \& \sim c_1) \Box\!\!\rightarrow e$.

Although they are offered only as necessary conditions for overdetermination, these two counterfactuals nonetheless entail, but are not entailed by the conditions contained in the standard definition of overdetermination as the production of an effect by two or more distinct, independently sufficient causes, and as such express a more restricted notion of overdetermination than does the latter. For given that (O1) and (O2) must be non-vacuously true, their antecedents cannot be impossible, meaning that it must be possible for each overdetermining cause of a given effect to occur in the absence of all the others, which entails that overdetermining causes must be distinct. And given the requirement represented in (O1) and (O2) that each overdetermining cause of a given effect must be capable of producing it in the absence of all the other overdetermining causes of that effect, taking (O1) and (O2) as necessary for overdetermination likewise entails that overdetermining causes must be independently sufficient for the effects they overdetermine. In contrast, while the standard definition of overdetermination requires any overdetermining causes of a given effect to be distinct from (i.e., non-identical with) one another, it does *not* require that it be possible for each to occur in the others' absence, as the non-vacuous truth of (O1) and (O2) does. Taking (O1) and (O2) as necessary for overdetermination hence adds to the standard definition the further requirement that overdetermining causes be capable of occurring in each other's absence.

Does this further condition restrict the notion of overdetermination enough, though, to rule out the seeming counterexamples to (3) discussed above? The answer depends on how finely we individuate the events that serve as overdetermining causes, and on the modal strength that we attribute to the laws of nature. Take, e.g., the causation of an effect *e* by both a certain non-fundamental physical event c_1 and the collection of

fundamental physical events c_2 on which c_1 depends. Such cases presented a problem under the standard definition of overdetermination, for they would appear to qualify, under that definition, as instances of overdetermination, in which case the regularity with which they occur would stand in conflict with (3). Since the conditions (O1) and (O2) only add to the standard definition the further requirement that overdetermining causes be capable of occurring independently of one another, if (O1) and (O2) are to avoid classifying *e* as overdetermined by c_1 and c_2, and thus succeed where the standard definition failed, it must somehow be shown either that c_1 cannot occur without c_2, or that c_2 cannot occur without c_1.[150] (Note that since *both* (O1) and (O2) must be non-vacuously true in order for *e* to qualify as overdetermined, it is not necessary to show *both* that c_1 cannot occur without c_2, *and* that c_2 cannot occur without c_1.)

As hinted at above, there are, I think, two ways one might do this: The first would be to say that it is part of what makes c_1 and/or c_2 the events that they are that the former depends on precisely those fundamental physical events that are c_2 and/or that the latter gives rise to precisely that non-fundamental physical event that is c_1. This would ensure that c_1 and c_2 could not both occur independently of one another by building their co-occurrence into the very individuation conditions for one or both of them. A second way of achieving this result would be to maintain that the laws that render c_2 sufficient for c_1 are metaphysically necessary. If such is the case, then while c_1 might be capable of

[150] Bennett (2003) notes another way in which *e* might fail to be overdetermined by c_1 and c_2 under (O1) and (O2), which is if c_2 only suffices for c_1 given certain background conditions, where these background conditions are also such that c_2 is unable or unlikely to be able to cause *e* without them. In such situations, (O1) will come out false, since the closest possible world wherein c_2 occurs without c_1 will be one in which the background conditions that enable c_2 to cause *e* do not obtain. For the purposes of the following discussion, I will assimilate such cases to those wherein c_2 cannot occur without c_1 by incorporating the relevant background conditions into c_2 itself.

occurring without c_2 (if, e.g., it is multiply realizable), it would be impossible for c_2 to occur without c_1, thus rendering (O2) merely vacuously true, and thereby disqualifying c_1 and c_2 as overdetermining causes of *e*. Against the first proposal, however, one might object that it is somewhat counter-intuitive that events like c_1 and/or c_2 should be so modally fragile that a single minute change in the fundamental physical events that give rise to c_1 should be enough to replace c_1 with some different event that just so happens to be exactly like c_1 in all but this one infinitesimal respect, and/or that one could not alter the physical laws that make it so that c_2 gives rise to c_1 without thereby also replacing c_2 with some distinct collection of fundamental physical events that is indistinguishable from c_2 in every respect save for the fact that it is not sufficient for c_1. In opposition to the second proposal, our intuitions might again protest that the clear conceivability of radical changes in the actual laws of nature tells against the metaphysical necessity of the laws correlating c_2 with c_1.

The intuitions weighing against these two proposals for ensuring that c_2 and c_1 cannot each occur without the other might be counterbalanced by invoking a *dispositional essentialist* view of properties, according to which "[a]t least some sparse, fundamental properties have dispositional essences," meaning that they bestow the same dispositions on their instances in all possible worlds (Bird, 2007, p.45).[151] If such a

[151] Prominent advocates of dispositional essentialism include Swoyer (1982), Molnar (2003), Mumford (2004), and Bird (2007). While dispositional essentialism is a controversial view, many of the arguments for it are, to my mind, quite convincing. The strongest seems to me to be that the basic properties that current physics ascribes to subatomic particles (e.g. charge, mass, and spin) appear inherently dispositional, and assuming that such particles are simple, the dispositions ascribed to them cannot be grounded in any categorical properties at some more basic level, but must instead be treated as fundamental properties in their own right. (See Ellis and Lierse (1994, pp.32, 42-3) and Mumford (2006).) Bird (2005, pp.447-53) also offers some compelling negative arguments against the rival, "quidditist" view of properties, which has the odd, and arguably unacceptable consequence that there is a possible world

conception of properties is correct, then just as the property of *being water* is necessarily identical with the property of *being H_20*, the property of *being negatively charged* may be necessarily related to the property of *being positively charged* in such a way that any object instantiating the former property is thereby disposed to attract and be attracted by objects instantiating the latter property (and *vice versa*). Adopting a dispositional essentialist view of properties lends immediate support to the second proposal (i.e. that c_2 and c_1 cannot each occur without the other because the laws correlating the two events are metaphysically necessary), for as many have noted, dispositional essentialism coincides quite naturally with the view that at least some laws of nature are metaphysically necessary.[152] This is because if one thinks that certain properties bestow the same dispositions on their instances in all possible worlds, then any laws that are made true by the fact that instances of such properties are disposed to behave in certain ways will be true in all possible worlds as well (although they will only be *non-vacuously* true in those worlds wherein such instances exist). Thus, if it is essential to the properties of *being negatively charged* and *being positively charged* that any object instantiating the one is thereby disposed to attract and be attracted by any object instantiating the other, the law stating that oppositely charged objects attract one another will be true in all possible worlds. Adopting dispositional essentialism hence enables one to ensure that c_2 is incapable of occurring without c_1 (and that *e* is hence not overdetermined by c_2 and c_1

identical to the actual world in every respect save for the fact that "charge has all the causal or nomic roles associated with gravitational mass" and *vice versa*. (See also Black (2000) and Wilson (2010).)

[152] The argument from dispositional essentialism to the metaphysical necessity of natural laws is discussed at length by Bird (2007, ch.3). See, however, Corry (2011) and Mumford (2004).

under (O1) and (O2)) by maintaining that since it is essential to the fundamental physical properties involved in c_2 that their instances be disposed to produce an occurrence of c_1 whenever they co-occur in the manner of c_2, the laws that correlate c_2 with c_1 are metaphysically necessary.

A moment's reflection shows that dispositional essentialism can be employed in defense of the first proposal (i.e. that c_2 and c_1 cannot each occur without the other because their co-occurrence is individuative of c_1 and/or c_2) as well. For if properties have dispositional essences, then the events which are the instances of such properties must be individuated at least partly in terms of the dispositions they necessarily possess as instances of the properties that they are instances of. Consequently, if the fundamental physical properties that c_2 is the instantiation of are essentially such as to bestow on their instances the disposition to give rise to c_1 whenever they are co-instantiated as they are in c_2, then it is essential to (individuative of) c_2 that it give rise to c_1, and therefore no event that fails to co-occur with c_1 could be the same event as c_2. The dispositional essentialist view of properties can thus be used to support, and in a way combine, both of the methods outlined above for ensuring that non-fundamental physical events and the fundamental physical events they depend on do not count as overdetermining their common effects under conditions (O1) and (O2). Advocates of (3) would therefore do well to accept dispositional essentialism, since doing so enables them to dismiss an important class of potential counterexamples to (3) as failing to qualify as legitimate instances of overdetermination.

Having dealt with those cases involving causation by both fundamental and non-fundamental physical events, the remaining putative counterexamples to (3) can be

handled quite easily once (O1) and (O2) are taken as necessary for overdetermination. Cases of the anvil-dropping sort involving the causation of an effect by both a whole and certain of its parts can, e.g., be quickly ruled out as instances of overdetermination, for since an anvil cannot be dropped on an egg without its top and bottom halves also being dropped on the egg, these two seemingly overdetermining causes of the egg's breaking cannot each occur without the other. When applied to such cases, one of the antecedents of (O1) and (O2) would hence be impossible, thereby rendering either (O1) or (O2) merely vacuously true; the result being that the common effects of wholes and their parts would not qualify as overdetermined.[153] As Bennett (2003, pp.478-9) points out, the adoption of (O1) and (O2) as necessary conditions for overdetermination also helps dismantle another class of potential counterexamples to (3), (noted in footnote 143) involving the seeming overdetermination of an effect by its proximal and distal causes. For given that if the distal cause of an effect had managed to occur without producing the effect's proximal cause, the effect itself would not have occurred (i.e. if c_1 causes c_2, which in turn causes e, then if c_2 did not occur, neither would have e, even if c_1 still did), either (O1) or (O2) would, when applied to such cases, turn out false, thereby ensuring that proximal and distal causes do not overdetermine their common effects under (O1) and (O2). The worry that these two types of cases (viz. those involving causation by both wholes and their parts, and those involving distal and proximal causes) might constitute

[153] Some further conditions may be needed, however, to prevent effects from being overdetermined by distinct proper parts of a whole (e.g. the top and bottom halves of the anvil), each of which is capable of existing and producing a given effect (e.g. the egg's breaking) on its own.

counterexamples to (3) can thus be easily allayed by adopting conditions (O1) and (O2) as necessary for overdetermination.

While restrictive enough to rule out cases that posed problems for (3) when interpreted according to the standard definition of overdetermination, (O1) and (O2) are not so restrictive as to prevent paradigm cases of the firing squad variety from qualifying as instances of overdetermination as well. This is because each shot that is independently sufficient to cause the death of a firing squad victim could also have occurred on its own, without any of the other shots being fired; e.g., all the other guns could have jammed. And while acceptance of dispositional essentialism might make it reasonable to think that fundamental physical events are incapable of occurring without the non-fundamental events they actually give rise to, it seems highly unlikely that it is also essential to the properties involved in each shot taken by the members of a firing squad that they could not have been co-instantiated in the event of that shot's being fired without all the other shots occurring as well. Since, then, any of the shots that were sufficient for the victim's death could have occurred without the others, any two such shots will non-vacuously satisfy (O1) and (O2), thus making them potentially overdetermining causes of their common effect (the victim's death).

2.iii. If causal overdetermination is rare, then are the effects of distinct, independently sufficient mental and physical causes overdetermined?

Supplemented with a dispositional essentialist view of properties, (O1) and (O2) thus provide a conception of causal overdetermination that retains paradigm cases within

the extension of the term while excluding cases whose regular occurrence would otherwise pose a threat to (3). In sum, we seem to have hit upon a notion of overdetermination that renders (3) defensible. Where does this leave us, though, with respect to mental causation and the Exclusion Problem? Now that we have found a way to vindicate (3), aren't we now back where we started in trying to make sense of the evident incompatibility of (1) and (4*) with (2) and (3)? Not entirely. For having seen what steps must be taken in order to make (3) defensible, it can now be argued that these same steps also lead to the conclusion that distinct, independently sufficient mental and physical causes needn't[154] overdetermine their common physical effects. The results of the preceding discussion thus provide the basis for a potential solution to the Exclusion Problem by putting us in a position to show that the conception of overdetermination under which (3) seems most plausible is likewise one according to which (3) is compatible with the conjunction of (1), (2), and (4*).

To see how this works, first recall that in order to handle the various counterexamples to (3) discussed above, overdetermination must be defined so as to apply only to cases wherein an effect is produced by causes that could each have occurred without the other. To establish that mental causation does not give rise to widespread overdetermination (at least in the sense of overdetermination under which it is plausible to suggest that overdetermination is rare), one therefore need only show that

[154] Dualists of course needn't (and shouldn't) maintain that mental and physical causes *never* overdetermine their joint effects, for this does happen, if rarely. Consider, e.g., the case of a person who decides to raise their arm and does so at the very same instant that a strong wind blows it into a raised position. (Note how the fortuitous, accidental character of the example seems to mark it as similar in kind to firing squad cases.)

most mental causes could not have occurred without the physical causes with which they share a common effect (or *vice versa*). Since this physical cause will typically be the physical event that the mental cause itself depends on (as, e.g., the deliberate movement of one's arm might be thought to be caused by both one's conscious decision to move it and by the neural event that realizes one's decision), this is consequently tantamount to showing that mental events that produce physical effects couldn't have occurred without the physical events they themselves depend on (or *vice versa*). Secondly, recall that in order to deal with problematic cases involving causation by both non-fundamental physical events and the fundamental physical events they depend on, it proved necessary (or at least advisable) for advocates of (3) to accept a dispositional essentialist view of properties as a way of ensuring that such causes could not both occur independently of one another, and therefore could not overdetermine their common effects under (O1) and (O2). This means that dualists are entitled to make use of the same strategy to argue that mental causes and their physical realizers cannot each occur without the other, and that the joint effects of such mental and physical events hence also do not qualify as overdetermined under (O1) and (O2).

With these points in mind, the dualist might reason as follows[155]: Seeing as the defense of (3) has already led us to accept, or at least seriously entertain the idea that at

[155] The following argument is inspired by an approach to the Exclusion Problem taken by Bennett (2003), Kallestrup (2006, pp.471-3), and Mellor (1995, pp.101-5). The suggestion that the psychophysical laws are metaphysically necessary and the appeal to dispositional essentialism as a way of accounting for the necessity of these laws do not, however, appear in their respective formulations of this approach. The only precedent for these ideas that I am aware of is Wilson (2005, p.438; 2011, p.142). Bennett (2008, esp. pp.296-9) would likely see these proposals as incompatible with dualism, as she takes dualists to be committed to the view that the mental does not supervene with metaphysical necessity upon the physical. For this reason, she argues that her "compatibilist" solution to the Exclusion Problem is only available to physicalists. I, however, do not see why dualists should have to deny that mind-body supervenience is

least some physical properties have dispositional essences, there seems no reason the exclude from the set of dispositions that may be essential to such properties the dispositions that their instances have to realize instances of specific mental properties. In other words, if we are willing to consider the idea that certain fundamental physical properties are essentially such as to give rise to certain types of non-fundamental physical events when co-instantiated in certain ways, why shouldn't we consider the possibility that such properties may also be essentially such that their instances are disposed to realize certain types of mental states when certain conditions are met as well?[156] By extending dispositional essentialism in this way to the dispositions that the bearers of certain physical properties have to realize instances of certain mental properties, the psychophysical laws which render the physical realizers of mental properties sufficient for the mental properties they realize are thereby made *metaphysically necessary*.[157] For if it is essential to the physical properties that are instantiated in the various physical conditions P_1, P_2,…P_n that are sufficient for a given mental property *M* that their instances be disposed to generate instances of *M* when related in the manner of P_1 or P_2 or…P_n, then the law stating that an instance of *M* occurs whenever P_1 or P_2 or…P_n obtains will be true in all possible worlds. Just as it offered proponents of (3) a way of

---

metaphysically necessary, and Bennett herself concedes that she does "not really [have] an argument" for thinking that they must. More on this below.

[156] Wilson (2005, pp.445-6) offers an argument in favor of this possibility from a form of holism about natural laws.

[157] Assuming that mental properties are multiply realizable, these laws might be conceived as many-to-one functions from physical descriptions of world-states to psychological descriptions of world-states. In order to ensure that the mental properties instantiated in a given world-state cannot be identified with, reduced to, or fully explained in terms of the physical properties instantiated in that world-state, the dualist must also insist that these psychophysical laws are distinctly non-physical, in that they are not included in, determined by, or deducible from the totality of purely physical facts.

attributing metaphysical necessity to the laws correlating non-fundamental physical events and the fundamental physical events they depend so as to ensure that such events do not overdetermining their joint effects under (O1) and (O2), dispositional essentialism can therefore also be used by dualists to attribute metaphysical necessity to the psychophysical laws linking mental states to their physical realizers so as to ensure that their joint of effects are also not overdetermined under (O1) and (O2), thereby avoiding any potential conflict between (3), and (1), (2), and (4*).

While endorsement of the view that the physical realizers of mental properties are essentially disposed to realize the properties they do under the conditions they actually do so thus offers dualists a way of addressing the Exclusion Problem without having to reject (1), (2), (3), or (4*), as Bennett (2003, p.491) points out, adopting such a solution to the Problem does require one to reject the metaphysical possibility of "zombies" (i.e., beings that are physically and functionally indistinguishable from normal human beings, but lack consciousness). This is because to establish, in the manner outlined above, that mental causes and their physical realizers do not overdetermine their common physical effects under (O1) and (O2), one must maintain that it is metaphysically impossible for the realizer of a given mental state to occur without it. But a zombie *just is* a hypothetical being that can be in the exact same physical state as a conscious human without having any (conscious) mental states at all, so to hold that a realizer of a mental state cannot occur without it is to deny that such beings are possible. Many contemporary dualists are likely to balk at this result, seeing as one of the more well-known arguments for their view, viz. David Chalmers' (1996) Conceivability Argument, rests on the supposition that since zombies are conceivable, such beings must also be metaphysically possible,

and some form of dualism must consequently be true. While otherwise quite friendly to dualism, as it allows one to retain both (1) and (4*) without having to contest either (2) or (3), the present solution to the Exclusion Problem hence also seems to require dualists who adopt it to give up on what many take to be the strongest reason for accepting dualism in the first place.

Further reflection, however, suggests that relinquishing the Conceivability Argument may not be so damaging to dualism after all. The job is made significantly easier by the fact that, in order to make use of the above solution to the Exclusion Problem, dualists needn't (as some physicalists do[158]) go so far as to deny that zombies are conceivable; all they need do is deny that such beings are metaphysically possible. They can hence confine their criticism of the Conceivability Argument to the inference from the conceivability of zombies to the conclusion that such creatures are metaphysically possible. This inference can be reasonably questioned by dualists and physicalists alike, for regardless of one's stance on the mind-body problem, one might naturally view with skepticism the suggestion that conceivability is anything more than a defeasible guide to what is possible. Following Stephen Yablo (1993 pp.33-6), one might, e.g., think that a scenario that is in fact impossible may nonetheless appear conceivable to those who lack knowledge of certain facts that demonstrate the impossibility of that scenario, so that certain scenarios may be conceivable, and hence seem possible, without their actually being so. In support of this point, Yablo (1993, pp.30-2) notes that there are certain propositions (e.g. the denial of Goldbach's conjecture) that are "undecidable"

[158] See, e.g., Dennett (1995).

(meaning that they are neither conceivable nor inconceivable), but which are either necessarily true or necessarily false. In such cases, either the proposition or its negation is necessarily false, yet neither is inconceivable, thereby belying the notion that our modal intuitions always track modal truths. Independent of the issue of whether or not the mind is physical, there is thus at least *prima facie* reason to reject the idea that our modal intuitions are reliable enough to provide us with adequate justification for believing in the im/possibility of whatever seems to us to be clearly in/conceivable. Dualists are therefore perfectly entitled to deny that the conceivability of zombies is sufficient proof of their possible existence, and can hence reject the Conceivability Argument in favor of the present solution to the Exclusion Problem without in any way compromising their position.

One might still wonder, though, whether the sort of position one would have to adopt in order to make use of the solution to the Exclusion Problem proposed above really warrants the title of dualism. The longstanding association of dualism with belief in the possibility of disembodied minds or (more recently) mindless bodies that are physically and functionally indistinguishable from those of a normal, conscious human might be taken to suggest that any view that does not allow for such possibilities must really be a form of physicalism in disguise.[159] Against this suggestion, however, it should

[159] This may be why Kroedel (2013, p.4), who proposes a similar solution to the Exclusion Problem, stops short of attributing full-blown metaphysical necessity to the psychophysical laws, and instead suggests that dualists should hold merely that the psychophysical laws have a privileged modal status *vis-à-vis* the physical laws such that "worlds where the psychophysical laws are violated are further from actuality than any worlds where only the ordinary laws are violated," for there are points (e.g. p.16) where he seems to suggest that ascribing metaphysical necessity to the psychophysical laws would be incompatible with dualism. (Bennett (2008) is much more explicit in her endorsement of this claim.) I, however, don't see why this should be the case. Assuming that it is indeed compatible with their position, holding that the psychophysical laws are metaphysically *necessary* also seems like the better option for

first be noted that there is nothing in the commitments one must adopt in order to employ the proposed solution to the Exclusion Problem that requires one to deny the metaphysical possibility of disembodied minds. To ensure that the joint effects of a mental event and its realizer are not overdetermined under (O1) and (O2), it is sufficient to maintain that the latter could not occur without the former. This leaves open the possibility that the converse is not also the case. Second, even if one rejects the possibility of disembodied minds, one could still maintain both that the psychophysical laws are metaphysically necessary and that the relation between mental and physical events is contingent by holding that while the psychophysical laws are necessary, the *physical* laws are *contingent*. Since the psychophysical laws are functions from distributions of physical properties and sets of physical laws to distributions of mental properties, introducing this asymmetry between the respective modal force of the physical and psychophysical laws would allow one to maintain that while the psychophysical laws are the same in all possible worlds, the same physical events could nonetheless give rise to different mental events in worlds governed by different physical laws.[160]

Finally, and perhaps most importantly, even if these two strategies for rendering the necessity of psychophysical laws consistent with the contingency of the mind-body

---

dualists, as such a move can draw independent motivation from dispositional essentialism, whereas Kroedel's proposal is open to the objection that it is entirely *ad hoc* (an objection to which Kroedel's (2013, p.15) responses are not fully satisfying).

[160] This would, however, put significant stress on the dispositional essentialist explanation for the metaphysical necessity of the psychophysical laws, since to account for the contrasting contingency of the physical laws, one would have to maintain that certain physical properties are essentially disposed to realize certain mental properties without having any essential dispositional relations to other physical properties, which seems odd.

relation prove impracticable, this still does not make the above solution to the Exclusion Problem inaccessible to dualists, for the claim that the correlations between mental and physical events are merely contingent is, I think, not something that dualists need endorse anyway. As defined in Chapter 1, dualism is merely the view that mental properties (i.e., properties that things exemplify insofar as they are endowed with intentionality and/or consciousness) and their instances are entirely distinct from, irreducible to, and incapable of being fully explained in terms of physical properties and their instances. According to this definition of their position, the most that dualists seem committed to regarding the modal status of the mind-body relation is that it cannot be logically necessary, as that would seem to imply the existence of some conceptual link between mental and physical properties that could potentially allow for the exhaustive explanation of the former in terms of the latter. This leaves them perfectly free, however, to maintain that the nomological correlations between the two sorts of properties are *a posteriori* necessities, which are just as metaphysically "brute" as the fundamental laws that govern purely physical events, and which, like the latter, can only be discovered through experience. While some may deem any definition of dualism that gives dualists this option inadequate, it seems to me, rather, that dualism has too often been saddled with commitments that are not essential to its core thesis; viz. that the mind exists, but cannot be identified with or fully explained in terms of anything purely physical. As the arguments advanced in Chapters 3-6 from the multiple realizability, intentionality, and phenomenal features of mental states illustrate, the latter thesis can be defended without making any assumptions as to whether the relation between mental states and their physical realizers is or is not metaphysically contingent. Despite its traditional association

with their position, the claim that correlations between mental and physical states are merely contingent thus strikes me as a thesis that dualists are entitled to reject. And considering the merits of the solution to the Exclusion Problem that rejecting this thesis gives them access to, they would, perhaps, be wise to do so.

The results of our discussion of (3) (the Absence of Systematic Overdetermination) can now be summed up as follows: First, it was shown that on the standard definition of causal overdetermination as the production of an effect by two or more distinct, independently sufficient causes, (3) is subject to a number of counterexamples involving kinds of cases that appear to satisfy the standard definition, but whose occurrence is frequent and widespread. Second, it was shown that by replacing the standard definition of overdetermination with conditions (O1) and (O2), most if not all of these counterexamples to (3) can be dealt with, especially if one also adopts a dispositional essentialist view of properties to ensure that the laws under which fundamental physical events suffice for the non-fundamental physical events they give rise to are the same in all possible worlds. Third, it was shown that by extending dispositional essentialism to the dispositions that certain physical properties bestow on their instances to realize certain mental properties when certain conditions are met, roughly the same reasoning used to rule out counterexamples to (3) involving the potential overdetermination of physical effects by fundamental and non-fundamental physical causes can also be used to show that physical effects with distinct, independently sufficient mental and physical causes are not overdetermined either. Fourth, it was argued that a position that makes use of the above reasoning to establish the compatibility of (3)

with (1), (2), and (4*) can be justly characterized as a form of dualism. Together, these four points lead us to the conclusion that under its most plausible interpretation, (3) implies a conception of overdetermination that need not apply to cases wherein a physical effect is produced by distinct and independently sufficient mental and physical causes, so that contrary to appearances, the propositions (1), (2), (3), and (4*) are not inconsistent. Dualists can therefore insist on (1) the Causal Efficacy of the Mental without having to deny either (2) the Causal Self-Sufficiency of the Physical or (3) the Absence of Systematic Overdetermination.

CHAPTER 8

THE CAUSAL SELF-SUFFICIENCY OF THE PHYSICAL

Having argued in the previous chapter that (3) the Absence of Systematic Overdetermination is either dubious or else compatible with the conjunction of (1) the Causal Efficacy of the Mental, (2) the Causal Self-Sufficiency of the Physical, and (4*) Mind-Body Dualism, the present chapter asks whether an alternative solution to the Exclusion Problem compatible with (1) and (4*) can be achieved through a critical examination of (2). The pursuit of an additional solution along these lines will likely strike some as necessary, inasmuch as the previous chapter may appear to leave one of the central worries expressed by the Exclusion Problem unaddressed; viz. that if the physical realm is causally self-sufficient, then any non-physical causation of physical effects seems utterly redundant. The arguments of the previous chapter do not directly address this issue, for even granting that distinct, independently sufficient mental and physical causes can produce the same effects without giving rise to any problematic form of overdetermination, the fact remains that if every physical effect already has a sufficient physical cause, the postulation of additional mental causes for such effects still appears superfluous. The present chapter responds to this concern by arguing that even if the solution proposed in the previous chapter is indeed inadequate, the Exclusion Problem can still be resolved without abandoning (1), (3), or (4*), simply by subjecting (2) to more careful scrutiny.

The structure of the chapter is as follows: Section 1 notes some difficulties involved in providing a non-question begging formulation of (2) that rules out

interactionist dualism. Section 2 then evaluates the support that (2) is often thought to derive from conservation laws of physics, and responds to various conservation-based arguments against interactionist dualism. Lastly, section 3 considers two ways in which dualists might go about rejecting (2).

1. Can (2) rule out non-physical causation of physical effects without begging the question against interactionist dualism?

Our discussion of (2) begins with a few words concerning another principle from which it must be sharply distinguished. As formulated by Jaegwon Kim (1998, p.40), this alternative principle, which I will refer to as the Causal Closure of the Physical (CCP), states that "If you pick any physical event and trace out its causal ancestry or posterity, that will never take you outside the physical domain. That is, no causal chain will ever cross the boundary between the physical and the nonphysical." In contrast to (2), which (as formulated in Chapter 1) states merely that every physical effect has a sufficient physical cause, CCP amounts to the assertion that no physical effect has a non-physical cause. As such, CCP is much too strong to be used in formulating the Exclusion Problem, since any formulation of the Problem citing CCP as a premise would beg the question against interactionist dualism by stating precisely what the Problem is supposed to show; viz., that if mental events are non-physical, then they cannot cause physical effects. Interactionist dualists are therefore only obliged to respond to the Problem if it is stated using (2) (or some other suitable principle) instead of CCP, and any statement of the Problem that relies on the use of CCP can be justly rejected by them as question begging.

With this point in mind, we can now raise the question that will be the focus of the present section, which is whether any formulation of (2) can be strong enough to rule out the possibility of non-physical causal intervention in the physical realm when conjoined with (3) while also being weak enough to avoid collapsing into the question-begging CCP. Certain arguments advanced by E.J. Lowe (2000) suggest that such a formulation of (2) may be hard to come by. If this is so, then just as they might argue (as proposed in the previous chapter) that (3) is either false or compatible with the conjunction of (1), (2), and (4*), dualists might likewise maintain that since the only formulations of (2) that are strong enough to be incompatible with the conjunction of (1), (3), and (4*) will end up entailing CCP, (2) is either question begging or else consistent with the conjunction of (1), (3), and (4*). If successful, such a strategy would again enable us to resolve the Exclusion Problem without having to deny the immateriality or causal efficacy of the mind.

The arguments I will draw from Lowe in support of this strategy show that for each of a series of increasingly strong formulations of (2), a potential case of causation can be produced that (a) satisfies that formulation, (b) contains no overdetermination, and (c) involves the causation of a physical effect by a non-physical, mental cause. As any formulation of (2) that is stronger than the strongest member of this series would seem to seriously risk entailing CCP, these potential cases will together pose a challenge to the idea that there is any non-question begging formulation of (2) that is strictly incompatible with the conjunction of (1), (3), and (4*). A further point that should not go unnoticed is that, in contrast with the type of case that was of primary interest in the previous chapter, the mental causes involved in the cases discussed below are distinctly non-redundant, in

that the fact that their effects would not occur without them isn't due merely to those effects having other, physical causes that are necessarily correlated with the alleged mental causes in such a way that if the latter failed to occur, so would the former. As such, the absence of overdetermination in these cases does not depend on any of the previous chapter's more controversial appeals to dispositional essentialism or necessary psychophysical laws. It instead rests solely on the fact that if one were to remove the mental causes from these cases, none of the remaining physical events would be by itself sufficient to bring about the relevant effect.

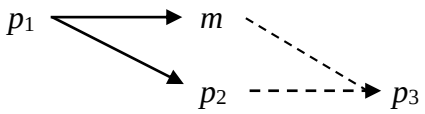
Of the various possible formulations of (2) that come most readily to mind, the weakest in fact seems to be the one offered in Chapter 1; viz. that every physical effect has a sufficient physical cause. As Lowe (2000, pp.575-6) points out, given that "…is a sufficient cause of…" is a transitive relation, it is rather easy to produce a case that contains no overdetermination and entails (1) and (4*) while also satisfying this principle. Consider, e.g., a case wherein a physical cause p_1 is the sufficient cause of a certain non-physical mental event *m*, which is in turn sufficient to cause another physical event p_2.

$$p_1 \longrightarrow m \longrightarrow p_2$$

Since "…is a sufficient cause of…" is transitive, p_2 has a sufficient physical cause (p_1), but is nonetheless also caused, without being overdetermined by, a non-physical, mental event (*m*). It seems, therefore, that (2) will have to be strengthened if its acceptance is to force us into rejecting (1), (3), or (4*).

A slightly stronger formulation of (2) that rules out the sort of case just considered is the following: Every physical effect has a sufficient physical cause, and at every time at which it has a cause, it has a physical cause. Since, in the previous case, p_2 has no

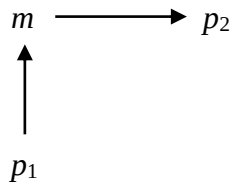
physical cause at the time of *m*'s occurrence, that case would be disallowed by the formulation of (2) just proposed. Here again, however, it is possible to produce an overdetermination-free case of mental causation that satisfies this principle, and thereby demonstrates its consistency with the conjunction of (1), (3), and (4*). Consider a situation wherein a physical event p_1 causes both a mental event *m* and another physical event p_2. Now let there be some further physical event p_3 that is caused by *m* and p_2, but only in conjunction with one another; i.e., while both *m* and p_2 are necessary for p_3, neither is sufficient on its own to produce it.[161]



Such a situation would satisfy the strengthened formulation of (2) now under consideration, for p_3 has a sufficient physical cause (viz., p_1), and, unlike the previous case, any time at which p_3 has a cause, it has at least one physical cause. Yet p_3 also has a non-physical, mental cause *m* whose non-redundant nature is evident from the fact that its removal from the situation described would result in the failure of p_3 to occur. Lastly, since neither *m* nor p_2 is by itself sufficient to produce p_3, p_3 is also not overdetermined, so the frequent occurrence of this kind of case would be fully consistent with (3). Since it allows for such cases, the strengthened formulation of (2) is hence still compatible with (1) (3), and (4*). It looks, then, as though an even stronger formulation is needed.

[161] The following two figures are simplified and adapted from Lowe (2000, p.577, 580). As with all the other figures used in this section, the events placed further to the left occur prior to those further to the right, and events on the same vertical occur simultaneously. Arrows represent causal relations; solid lines indicate sufficient causes, whereas dashed lines indicate that the event at the arrow's tail is not by itself sufficient to cause the event at the arrow's head.

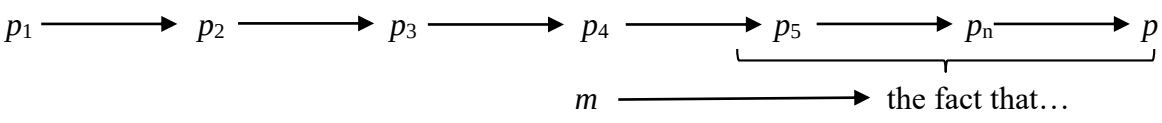
The following seems to be the next logical step: "At every time at which any physical event has a cause, it has a sufficient physical cause" (Lowe, 2000, p.576). Since, in the case just considered, p_3 does not have a sufficient physical cause at the time of *m*'s occurrence, this new formulation of (2) would enable us to rule that case out. A case of overdetermination-free mental causation that satisfies even this formulation of (2) can, however, still be produced if one allows for the possibility of simultaneous causation (Lowe, 2000, p.576-7). Indeed, all one need do is modify the first case discussed above so as to make *m* and p_1 synchronic.



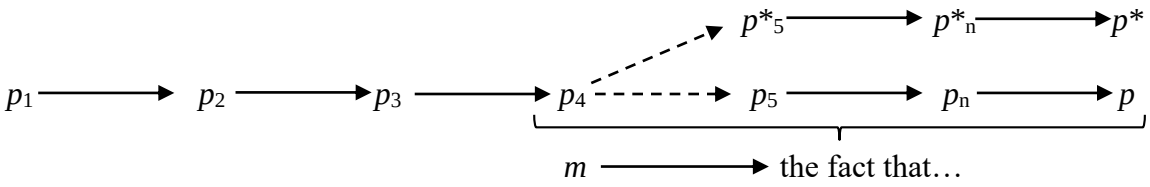
Here the transitivity of "…is a sufficient cause of…" again ensures that since p_1 is a sufficient cause of *m*, and *m* is a sufficient cause of p_2, p_1 is also a sufficient cause of p_2, thereby making this a situation that satisfies the formulation of (2) currently under consideration. But since the immediate cause of p_2 is the mental event *m*, and none of the events involved is overdetermined, the present case also entails (1) and (4*), and its frequent occurrence would be compatible with (3). The proposed strengthening of (2) thus again seems to leave us with a principle that is consistent with the conjunction of (1), (3), and (4*). Those who take simultaneous causation to be impossible will of course object to the case just presented, but as Lowe (2000, p.577) points out, it must then "be acknowledged that [the rejection of simultaneous causation] is a further substantive claim, without which [the formulation of (2) just proposed] cannot be used to mount an attack on interactionist dualism." The defense of this claim might, moreover, be made

difficult by certain oddities involved in the measurement of entangled quantum states, which seem to suggest that the measurement of a particle can have an instantaneous effect on the state of another, distantly located particle with which the former is entangled (Maudlin, 2011, ch.5). At any rate, until presented with a convincing refutation of the possibility of simultaneous causation, dualists are entitled to regard the present formulation of (2) as compatible with (1), (3), and (4*).

The previous three formulations of (2) having proven unable to generate the inconsistency between (1), (2), (3), and (4*) needed to make the Exclusion Problem a real problem for dualists, let us now consider one last potential formulation of (2) that seems stated in such a way that any stronger formulation of (2) would most likely collapse into the question begging CCP. The formulation is: "Every physical event contains only other physical events in its transitive causal closure," where the transitive causal closure of an event *p* is "the set of events consisting of the immediate causes of [*p*], the immediate causes of those causes, the immediate causes of *those* causes…and so on" (Lowe, 2000, pp.581-2). While this principle rules out the three cases discussed thus far, and may indeed seem to leave not even the slightest possibility for *any* non-physical causation of physical effects, Lowe (2000, pp.582-3) presents a situation that suggests otherwise. Consider a physical event *p* that contains the causal chain of physical events $p_1 \ldots p_n$ in its transitive causal closure. What's to prevent us, Lowe asks, from saying that a certain non-physical mental event *m* is the cause *of the fact that* certain of the events from *p*'s transitive causal closure cause the occurrence of *p*?



Since *m* is not the immediate cause of *p*, or of any of the events $p_1 \ldots p_n$ in *p*'s transitive causal closure, such a situation is completely compatible with the formulation of (2) currently under consideration. Likewise, since *m* does not cause any event that already has a sufficient cause (how could it, when it doesn't cause any event at all?), the proposed case also does not involve any overdetermination. And while the role that *m* plays in this case may strike some as superfluous, we can easily conceive of a similar situation in which it is clearly not by imagining that p_4 has a 50% chance of causing p_5 and a 50% chance of causing p^*_5, and that the occurrence of *m* suffices to determine p_4 in the direction of p_5, so that the fact that p_4 caused $p_5 \ldots p_n$ and *p* to occur (rather than $p^*_5 \ldots p^*_n$ and *p**) is properly attributable to the occurrence of *m*.[162]



While those who hold that causation is always a relation between events will object to the use of facts as causal *relata* in the cases just described, given that there are other seemingly legitimate instances of causation (e.g. causation by omission) that also fail to fit into the model of event causation, the burden of proof is on those who deny that facts can be caused to provide some reason for thinking this is so.[163] In the absence, then, of any more substantial reason for rejecting the above two cases, the correct response

[162] If quantum mechanics is any guide, then many physical causes are in fact like p_4, in that they do not fully necessitate their effects, but instead merely make them more or less probable.

[163] See Mellor (1995, ch.9, 11) for a defense of the view that facts can be causes and effects.

seems to be to regard the proposed formulation of (2) (which in this case looks to be the strongest formulation admissible) as again compatible with (1), (3), and (4*).

Summing up, it appears that providing a non-question begging formulation of the Causal Self-Sufficiency of the Physical that is truly inconsistent with Mind-Body Dualism, the Causal Efficacy of the Mental, and the Absence of Systematic of Overdetermination is a more difficult task than one might first have thought. In light of this fact, dualists seem entitled to regard the Exclusion Problem as posing no problem for their position until presented with a formulation of (2) that can rule out cases of the sort discussed above without collapsing into CCP.[164]

2. Do conservation laws pose a problem for dualism?

We have now seen that dualists may have reason to hold that (2) either begs the question against the possibility of non-physical causation of physical effects, or else is compatible with (1), (3), and (4*), and can hence be accepted by one who holds that non-physical mental causes can produce non-overdetermined physical effects. One might wonder, though: even assuming that (2) is compatible with (1), (3), and (4*), what reason is there to accept it anyway? Why shouldn't dualists save themselves the trouble of demonstrating the compatibility of (2) with their position and instead just reject (2) outright? The standard answer to these questions is that (2) follows or else draws strong

[164] Yates (2009, pp.116-7, 131) and Garcia (2014, pp.101, 107) provide two additional formulations of (2) that might be up to the task, but Garcia goes on to argue that "we have grounds for rejecting" the principle he formulates, and Yates argues that his formulation of (2) is effectively useless in any argument for physicalism, since it "costs almost as much, in evidential terms, as physicalism itself."

inductive support from certain fundamental and well-established laws of physics: viz., the conservation laws of energy and momentum.

In their colloquial form, the conservation laws of energy and momentum state that in a closed system (i.e., one that exchanges no matter or energy with its surroundings, and on which no external force acts), the total quantity of energy and momentum always remains the same. The idea that these laws pose a problem for interactionist dualism goes back at least as far as Leibniz (1714/1898, §80), who objected to Descartes' theory of mind-body interaction on the grounds that it violates the conservation of momentum (or as Leibniz (c.1691/1896, p.667) called it, the conservation of "common progress"), and despite periodic fluctuations in its popularity, acceptance of this idea is today still fairly widespread.[165] Of those who hold this view, some (e.g. Papineau (2001)) see the problem that conservation laws raise for interactionist dualism as lying in the support they lend to (2), whereas others (e.g. those cited in footnote 165) seem to see such laws as flatly inconsistent with any non-physical causation of physical effects. The aim of the present section will be to respond to these criticisms by showing that (2) does not follow from conservation laws, and that such laws are moreover consistent with interactionist dualism.

Let us begin with the question of whether conservation laws can be used as the basis for a direct argument against interactionist dualism that is independent of (2). (For ease of exposition, I'll henceforth focus primarily on the law of conservation of energy

[165] Proponents of this idea include Crane (2001, p.48), Dennett (1991, p.35), Fodor (1981, p.114), Putnam (1999, p.78-9) and Searle (2004, p.42). Other advocates are noted by Montero (2006, pp.384-5) and Gibb (2010, p.636fn1).

(CoE). The arguments and claims to be discussed can, however, be applied *mutatis mutandis* to other conservation laws as well.) A natural first step in attempting to construct such an argument might be to conjoin CoE with the additional premises that (i) the universe is a closed system, and (ii) that any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered. From these premises, one might then try to derive the conclusion that no change of bodily motion has a non-physical cause, and that interactionist dualism is therefore false. This yields Argument I:

Argument I

(CoE) Energy is conserved in any closed system.

(i) The universe is a closed system.

(ii) Any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered.

No change of bodily motion has a non-physical cause.

As many have noted, however, even granting premise (ii), the resulting argument against interactionist dualism invalid, for the possibility is left open that mental events might themselves possess some form of energy, which they transfer to those bodies whose motion they affect.[166] So long as the total amount of energy that exists before and after

[166] Hart (1988, ch.9, 10, 12) develops a fairly sophisticated form of interactionist dualism along these lines. In what follows, the ascription of energy to mental states and events is shorthand for the suggestion that the minds that are in such states or in which such events occur possess certain quantities of energy by virtue of instantiating the mental properties that such states and events are the instantiations of.

these exchanges turns out to be the same, non-physical, mental events could be capable of causally affecting bodily motion without violating CoE or any of the other premises of the argument just proposed.

To exclude the possibility that non-physical, mental events might causally interact with physical events in this way, a *valid* argument against interactionist dualism based on (i), (ii), and CoE must hence contain an additional premise stating either that (iii) nothing non-physical has energy (or at least none that is capable of being transferred to any physical body), or (iv) that the physical realm constitutes a closed system (i.e., one that exchanges no matter or energy with its surroundings, and on which no external force acts). The choice between these two further premises gives us the following two arguments in place of the one first proposed:

Argument II

(CoE) Energy is conserved in any closed system.

(i) The universe is a closed system.

(ii) Any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered.

(iii) Nothing non-physical has energy (or at least none that is capable of being transferred to any physical body)

No change of bodily motion has a non-physical cause.

Similarly, an attribution of energy to a physical state or event may serve as shorthand for the claim that the object involved in that state or event possesses a certain quantity of energy by virtue of instantiating the physical property that the state or event is the instantiation of.

Argument III

(CoE) Energy is conserved in any closed system.

(i) The universe is a closed system.

(ii) Any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered.

(iv) The physical realm is a closed system.

No change of bodily motion has a non-physical cause.

While both of these arguments are perfectly valid, I do not think either is ultimately all that compelling. First, as noted by Barbara Montero (2006), once (iii) or (iv) is added to the original CoE-based argument against interactionist dualism proposed above, the appeal to CoE becomes unnecessary[167], for the falsity of interactionist dualism can be deduced from (ii) and (iii), or (ii) and (iv) alone. The fact that CoE can be removed from the premises of Arguments II and III without in any way affecting their validity would seem to suggest that the real threat to interactionist dualism lies *not* in CoE, but rather in certain of the remaining premises of Arguments II and III, viz. (ii), (iii), and (iv).[168] This

[167] Note that (i) is thereby made redundant as well.

[168] See Broad (1925, pp.107, 109). Montero (2006, p.395) takes this to show that CoE in fact has "nothing whatsoever" to do with the defense of physicalism. This would be correct if Arguments II and III were the *only* arguments against interactionist dualism that CoE might figure into, but as will be shown below, that is not the case. Koksvik (2007b, p.579), e.g., points out that an additional argument against interactionist dualism can be constructed by conjoining CoE and (iv) with the assumptions that "[i]f a non-physical mind changes a physical system, it changes its energy level," and that "[i]f the energy level of a physical system is changed by a non-physical system, energy is not conserved in the physical world." Discussion of this argument will be postponed until section 2.iv.

would be good news for dualists, for since these other premises are not self-evident, nor are they, like CoE, well-established, fundamental laws of physics, contesting them should be significantly easier than calling CoE itself into question.

2.i. Can non-physical entities possess energy?

Starting, then, with Argument II, what reason might dualists have to accept premise (iii) – the claim that nothing non-physical has energy? One such reason might be found in Edward Averill and B.F. Keating's (1981, p.105) remark that "changes in the energy of non-physical things are undefined, i.e. there is no way of specifying the state of a non-physical thing in terms of the variables of physics." Here the idea seems to be that quantities of energy are attributed to things only under some physical description of their states and relations (i.e., some description given in terms of the proprietary laws and kinds of physics); therefore, energy cannot be properly attributed to any non-physical thing, because non-physical things cannot be described in physical terms. To this the dualist might respond by arguing that there is no reason why a non-physical entity could not be ascribed a physical quantity if such an ascription were warranted by certain effects that it was found to have upon some physical system.[169] If, e.g., the occurrence of a

[169] On the assumption that while non-physical entities might possess energy, they cannot possess mass, the energy possessed by non-physical entities would have to differ from that possessed by physical entities at least in the respect that when possessed by non-physical entities, it is not equivalent with mass. This may seem dubious, but any solution to the Exclusion Problem is bound to have some consequences that are difficult to accept. Note also that the kind of mass physicists are generally most interested in (viz. rest or invariant mass) is *not* equivalent with energy, since photons have energy but their rest mass is 0. In light of this point, dualists might suggest either (a) that if protons can have energy without rest mass, it may be possible for minds to have energy without any mass whatsoever, or (b) that it is only the ascription of

certain mental event was found to correlate with changes in the energy level of the brain in which that event was realized, and during these changes the energy in the brain's physical surroundings was known to remain constant, then rather than immediately rejecting CoE, it would not seem unreasonable to instead attribute the quantity of energy needed to account for these changes in the brain's energy level to the mental event itself. Averill and Keating are therefore wrong claim that "there is no way of specifying the state of a non-physical thing in terms of the variables of physics," for such specifications could be made on the basis of such a thing's measurable effects on physical systems.[170] The values of the relevant variables could be specified as required to account for those effects in a manner that is consistent with CoE.[171]

Assuming that ascriptions of energy to non-physical entities could indeed be justified in this way, wouldn't the attribution of a physical quantity to such an entity nonetheless deprive the latter of its non-physical status? In other words, if mental events can possess physical quantities, what grounds could there be for treating such events as non-physical? Here the dualist might reply that so long as mental events exhibit certain qualitative and/or intentional features that cannot be identified with, reduced to, or fully explained in terms of any physical quantities we might ascribe to them, there is no reason

---

rest mass to minds that is inconsistent with their position, and that non-physical minds might thus, like protons, have energy and relativistic mass (which *are* equivalent), but no rest mass.

[170] Averill and Keating's reluctance to attribute energy to non-physical states also seems to stand in tension with their suggestion (discussed further below) that the mind exerts an external, non-physical force on the brain, for it is unclear how something could exert a force without possessing energy.

[171] See Fair (1979, p.229), who notes that "the hypothesis that energy is a *conserved* quantity" has at times forced physicists to "revise [their] definition of energy," thereby leading to the discovery of "new forms and carriers of energy." Couldn't the same procedure at some point lead us to treat mental events as potential carriers of energy as well?

to view their possession of such quantities as somehow "turning them into" physical events. For property dualists, in particular, this idea should not sound overly strange, for if the property dualist is correct in thinking that physical entities can bear non-physical properties, there would seem to be no reason why non-physical entities could not likewise be capable of possessing certain physical quantities.

2.ii. Papineau's Argument from Fundamental Forces

Averill and Keating's remark thus does not appear to provide a compelling argument for (iii) – the claim that nothing non-physical has energy. A more promising case for the claim might, however, be drawn from two arguments developed by David Papineau (2001), which he calls, respectively, the Argument from Fundamental Forces, and the Argument from Physiology. These two arguments (which were summarized briefly in Chapter 2) draw upon the history of science as providing inductive support for the thesis that "there are no special [i.e., non-physical] mental or vital forces" (Papineau, 2001, p.27). Since energy is partly defined in terms of force (for energy is the capacity to do work or transfer heat, and work is the application of force to a body that results in the displacement of that body in the force's direction), it follows that if there are no special, non-physical forces, then nothing non-physical has energy (or at least none that it can use to do work). Papineau's arguments can thus be equally viewed as arguments for (iii), inasmuch as (iii) is entailed by the thesis that they purport to establish.

As formulated by Papineau, the Argument from Fundamental Forces "is that all apparently special forces characteristically *reduce* to a small stock of basic physical

forces that conserve energy. Causes of macroscopic accelerations standardly turn out to be composed out of a few fundamental physical forces that operate throughout nature. So, while we ordinarily attribute certain physical effects to…'mental causes,' we should recognize that these causes, just as all causes of physical effects, are ultimately composed of the few basic physical forces" (Papineau, 2001, p.27). In short, reflection on the history of science shows that physical effects that were at one point attributed to the operation of certain "special" forces have typically been found, upon closer analysis, to be fully explainable in terms of a small number of fundamental physical forces (e.g., gravity, electromagnetism, and strong and weak nuclear forces). This would seem to suggest that the same will likely hold true of any physical effects that are now commonly accounted for by appeal to mental causes.

Two points should be made regarding the relation between this argument, CoE, and (2). First, while it was noted above that CoE is not needed to deduce the falsity of interactionist dualism from (ii) and (iii), CoE might still be said have some relevance for such an argument against dualism, inasmuch as it *does* play a role in the Argument from Fundamental Forces, which (as previously mentioned) can be viewed as an argument for (iii). Papineau (2001, p.28) describes this role as consisting in a certain tension between CoE and the postulation of special forces that do not reduce to any fundamental physical forces, which is that "[a]n insistence on the independent existence of *sui generis* special forces inside bodies threatens to remove the reasons for believing in the conservation of energy in the first place. For there are no obvious grounds for expecting such *sui generis* forces to be conservative." In other words, the main reason we have for accepting CoE is that the few fundamental physical forces we know of appear to obey it, and all other

forces that have been "quantatively analyzed" thus far have turned out to reduce to these few forces. The postulation of special forces that do not reduce to these few forces would therefore undermine our justification for accepting CoE, for the simple reason that if the former forces are indeed distinct from the latter, we would have no reason to believe that they share the latter's obedience to CoE. Consequently, any reason we have for accepting CoE may *ipso facto* be viewed as a reason for rejecting the existence of special, non-physical forces (and *vice versa*). The fact, then, that we have very strong grounds for believing that CoE is true can therefore be seen as lending added support to the idea that all forces are reducible to a few basic physical forces that conserve energy.

On the assumption that all physical effects can be explained as due to the action (or inaction) of forces, one can easily see how the support that CoE lends to the idea just mentioned might also seem to make it into a source of support for (2). Put simply, if this assumption is correct, then the thesis that all forces reduce to a small number of basic physical forces entails that every physical effect has a sufficient physical cause, so any support that CoE lends to the former thesis will thereby count equally in favor (2) as well. This in fact appears to be the main way in which Papineau sees CoE as figuring into the case against interactionist dualism, viz. as helping to generate the Exclusion Problem by providing evidence for (2). Rather than arguing directly from CoE to the falsity of interactionist dualism, Papineau thus instead seems to have something like the following in mind:

> (a) If there are forces that do not reduce to a small stock of basic physical forces that conserve energy, then we have no good reason to accept CoE.
>
> (b) We have good reason to accept CoE.

(c) All forces reduce to a small stock of basic physical forces that conserve energy. (From (a) and (b))

(d) All physical effects can be fully explained as due to the action of forces.

(2) Every physical effect has a sufficient physical cause. (From (c) and (d))

(3) Overdetermination is rare.

~(4*) or ~(1) Mental events are physical events or mental events do not cause physical effects.

This argument I think captures the part that Papineau sees CoE as playing both in the Argument from Fundamental Forces and in the more general case against interactionist dualism.

There are at least two major criticisms that dualists might raise to the preceding arguments. The first, which has been made by Robert Garcia (2014, p.102) and Ole Koksvik (2007a, pp.133-4), is that the Argument from Fundamental Forces begs the question against dualism, because the inference it makes from the successful reduction of various *physical* forces (e.g. friction) to the conclusion that all forces will ultimately be so reduced relies on the very sort of presumed similarity between mental and physical entities that the dualist is apt to deny. Since the dualist maintains that the mind is importantly different from any physical thing, s/he will naturally be skeptical of the claim that any mental forces are likely to reduce to a small stock of conservative, physical forces just because certain physical forces have been so reduced in the past. On their view, mental and physical entities are *disanalogous*, so the fact that various physical forces have been reduced to a small number of conservative, physical forces gives us no

reason to expect that the same will hold true of mental forces (if such forces exist). The Argument from Fundamental Forces therefore fails to make a cogent case for (iii).

The second criticism is directed at the support that CoE is alleged to lend to the Argument from Fundamental Forces and thus also to (2). Here, again, Papineau's idea seems to be that our reasons for accepting CoE count in favor of the view that all forces reduce to a small number of fundamental physical forces in terms of which all physical effects can be explained, because we have no reason to suppose that any additional, non-physical forces would conserve energy like the known fundamental physical forces do. As he himself notes, though:

> [T]his is scarcely conclusive. Those thinkers who remain convinced…that there must be irreducible special forces inside living bodies, could still respect the universal conservation of energy, by maintaining that these extra forces must themselves operate conservatively. In support of this they could [offer] the alternative inductive argument that, because all the *other* fundamental forces examined so far have turned out to be conservative, we should infer that any extra…mental fundamental forces will be conservative too. (Papineau, 2001, p.29)

In short, the dualist might argue that the postulation of non-physical, mental forces is not at all in tension with acceptance of CoE, because the conservative nature of the basic physical forces we now know of gives us ample reason to think that any non-physical, mental forces there are will likewise obey CoE.

While this is, I think, the right thing for the dualist to say, there is nonetheless at least an apparent conflict between this argument and the reasons given above for rejecting the Argument from Fundamental Forces. How, one might ask, can the dualist cite the conservative nature of basic physical forces as evidence for the assumption that any non-physical, mental forces will likewise conserve energy, while at the same time

claiming that the mind is so unlike any physical thing that the reduction of various physical forces to a few basic physical forces gives us no reason to believe that mental forces (if such there are) will eventually be so reduced as well? The former idea seems to assume what the latter denies: viz. that physical and mental forces are sufficiently similar that traits possessed by the one (e.g., obedience to CoE and/or reducibility to a few basic physical forces) can be justifiably attributed to the other. Which is it then? Are mental and physical forces similar enough to warrant such inferences, or aren't they?

This dilemma can be resolved by noting that we have much greater reason to think that all forces conserve energy than we do to believe that all forces will reduce to a small stock of basic physical ones. CoE is, after all, a well-established scientific law, whereas the latter conjecture is not a law of any science. It is, moreover, somewhat misleading to present the history of science as providing unequivocal inductive support for the reducibility of all forces to a few physical ones, since scientists have also seen fit to *add* to the stock of fundamental forces in cases where certain newly discovered interactions could not be fully accounted for in terms of the forces then viewed as basic.[172] Given, then, that the number of forces thought to be basic and irreducible has been increased in the past, it seems reasonable to allow that further additions to this stock could be made in the future[173], and who's to say that these additions might not include

[172] The postulation of the strong and weak nuclear forces in order to account, respectively, for the phenomenon of beta decay and the coherence of the nucleus despite the electromagnetic repulsion between protons might be cited as examples of this point.

[173] A similar point is made by Popper in Popper and Eccles (1977, pp.542-3). This idea might also be supported by recent suggestions that the postulation of additional fundamental forces besides the usual four is needed to overcome certain difficulties that the Standard Model has in accounting for dark matter and the accelerated expansion of the universe. (See Reich (2010), Battersby (2013), Feng and Trodden (2014), and Dobrescu and Lincoln (2015).)

certain forces that are mental rather than physical in nature? The claim that all forces will ultimately be reduced to a few basic physical forces hence seems dubious enough on its own that dualists needn't commit themselves to the view that mental and physical forces are *completely* disanalogous in order to deny it. They can therefore allow that any mental forces are liable to at least be similar enough to known physical forces that the former's obedience to CoE can be plausibly inferred from that of the latter, without also having to concede that the reducibility of any mental forces to fundamental physical ones can be justifiably inferred from the fact that many physical forces have been so reduced in the past.

2.iii. Papineau's Argument from Physiology

The preceding considerations seem to show that there is no straightforward, decisive argument from CoE to either (iii) – the claim that nothing non-physical has energy – or (2) – the Causal Self-Sufficiency of the Physical. This, however, leaves Papineau's second argument for (iii), the Argument from Physiology, unaddressed. As presented by Papineau, the Argument from Physiology "is simply that there is no direct evidence for vital or mental forces. Physiological research reveals no phenomena in living bodies that manifest such forces. All organic processes in living bodies seem to be fully accounted for by normal physical forces" (Papineau, 2001, p.27). While this argument can be seen as "operating against the background provided by the…argument from fundamental forces," inasmuch as the assumption that all physical forces reduce to a few fundamental ones makes it much easier to say what sort of changes in living bodies

would count as evidence of the action of non-physical forces, (viz. any changes that cannot be explained in terms of those few physical forces deemed basic), it does not share the latter argument's dependence on CoE (Papineau, 2001, p.30). Here the point is merely that when we look in those places where additional non-physical forces would be most likely to manifest themselves, we fail to find any physical changes that might be attributed to their influence and thus taken as evidence of their presence, for all such changes can be fully explained in terms of the physical forces already at our disposal. The problem is thus not that the putative non-physical forces may fail to conserve energy, but rather that they don't seem to produce any effects.

Although the Argument from Physiology itself "has little to do with the conservation of energy," it nevertheless demands a response from dualists who are interested in using the arguments advanced above to develop an account of mental causation that rejects (2) without violating CoE (Papineau, 2001, p.30). To see why, recall that the basic idea behind the account proposed in section 2.i. was that mental causes might possess certain quantities of energy that enable them to exert a non-physical force on physical entities (viz. parts of the brain) and thereby alter the motion of such entities in ways that cannot be accounted for in terms of any purely physical events. Section 2.ii. sought to show that such an account does not stand in any necessary conflict with CoE. The dualist, however, does not seem entitled to rest content with this result, for even granting that CoE is consistent with the postulation of non-physical, mental forces that produce physical effects for which there are no sufficient physical causes, the question then naturally arises: What grounds do we have for thinking that such forces actually exist? The Argument from Physiology holds that the fact that we have found no

evidence of such forces gives us reason to believe that there are none. Unless the dualist can offer some rebuttal to this Argument, the work that has been done towards establishing the consistency of CoE with the account of mental causation proposed in section 2.i. will seem rather pointless. For nothing much follows from the observation that non-physical, mental forces *could* produce physical effects that lack sufficient physical causes without violating CoE if such forces don't exist.

While Papineau appears to view the Argument from Physiology as offering a more conclusive refutation of the existence of non-physical forces than the Argument from Fundamental Forces, the response to the former argument is actually much simpler than that given to the latter above. In this case, the dualist need only point out that while neuroscience has made astounding progress in the past century, there can be little doubt that our scientific understanding of the inner workings of the brain is currently still in its initial stages. As such, it is extremely premature to claim that all changes of acceleration that take place within the brain can be fully explained in terms of the action of physical forces. While many important neural processes (e.g., the transmission of signals across synapses, or the modification of neural pathways through learning or memory) have been analyzed into more basic bio-chemical processes that perhaps can be explained in such terms, further research may yet uncover circumstances in which the physical forces acting on the material constituents of a brain are unable to account for certain subsequent increases/decreases in the firing rate of certain neurons or the amounts of certain neurotransmitters being released. Until our understanding of the brain develops to the point where such possibilities can be ruled out, to dismiss them seems a bit hasty, to say

the least.[174] And although Papineau is right to point out that we have yet to find any positive evidence for the existence of non-physical forces in the brain, given, again, that neuroscience is still in its infancy, to take this as evidence *against* the existence of such forces is to argue from ignorance. The absence of such evidence at these early stages should, moreover, be expected, seeing as one must first have an understanding of what changes within a system *can* be explained by the action of physical forces before one can be in a position to recognize occurrences within the system that *cannot* be explained in such terms. It is hence only after we have acquired a more fully developed understanding of the biochemical mechanisms that account for the various forms of neural activity that take place in the brain that we can expect to be able to identify changes of acceleration in the brain that cannot be attributed to the action of physical forces. For these reasons, the Argument from Physiology fails to provide any compelling grounds for accepting (iii) – that nothing non-physical has energy.

2.iv. Is the physical realm a closed system?

As none of the arguments for (iii) have stood up to scrutiny, dualists seem entitled to regard the possession of energy by mental states as an open possibility, and so long as this possibility remains open, the argument against interactionist dualism from (iii) and (ii) – the claim that all changes of bodily motion involve a transference of energy between the cause of the change and the body whose motion is altered – is at best

[174] See Garcia (2014, p.103).

inconclusive. The argument from (ii) and (iv) – the claim that the physical realm is a closed system – has, however, yet to be answered. Fortunately for dualists, both of these premises are contentious. With regard to (iv), Montero (2006, pp.386-8) maintains that in contrast to the view that energy is conserved in the universe *as a whole*, the idea that energy is conserved "among the physical components of the universe" is "a philosophical principle rather than a law of physics," for "while physics gives us reason to believe the [former conjecture], it does not seem to give us reason to believe [the latter]." If this is correct, then given that CoE applies only to closed systems, the alleged fact that current science only supports the claim that energy is conserved in the universe as a whole implies that the only closed system that we have reason to believe exists is the universe itself, in its entirety. To assume, therefore, that the physical realm constitutes its own closed system would be to assume either that everything in the universe is physical (i.e., that physicalism is true), or that in addition to the universe in its entirety, there is also unique subsystem within the universe that likewise conserves energy. The former assumption begs the question against dualism, and the latter is, according to Montero, unsupported by current science. Either way, dualists would seem justified in accepting CoE while rejecting (iv), provided that Montero's contention is well founded.

Montero, however, offers little support for her claim that the restriction of CoE to the physical realm is "a philosophical principle" unsupported by current science, and reasons can indeed be given for thinking that it is instead under such an interpretation that CoE has the most evidential support. Thus, following Koksvik (2007b, pp.579-80), one might argue that since "the experimental evidence in favor of CoE has resulted from observations of entirely physical systems,…[t]he evidence we have for CoE is *only*

evidence for the restricted version," which "holds that energy is conserved among the physical components" of the universe. Given, moreover, that CoE only applies to closed systems, the fact (if it is a fact) that CoE applies to the physical realm would likewise imply (iv) – that the physical realm is a closed system. If this argument for (iv) is successful, then in addition to the argument from (iv) and (ii), interactionist dualism will also, as Koksvik (2007b, p.579) notes, be threatened by the following argument from (iv) and CoE:

Argument IV

(CoE) Energy is conserved in any closed system.

(iv) The physical realm is a closed system.

(v) "If a non-physical mind changes a physical system, it changes its energy level."

(vi) "If the energy level of a physical system is changed by a non-physical system, energy is not conserved in the physical [realm]."

"It is not the case that a non-physical mind changes a physical system."

In light of these challenges, the prospects for interactionist dualism would be greatly improved if grounds could be given for doubting (iv) and the corresponding claim that energy is conserved in the physical realm.

To provide such grounds, dualists might start by suggesting that the only reason that current evidence seems to support the restriction of CoE to the physical realm is that we have yet to fully map out all the various exchanges of energy that take place in those

circumstances in which the restricted CoE is most likely to be violated, viz. those wherein a certain brain state gives rise to or immediately follows the occurrence of some mental event. As long as this remains the case, the restricted CoE must be regarded as at best underdetermined, for a hypothesis can hardly be regarded as empirically confirmed if it has not been tested in the very circumstances where it stands the greatest risk of being falsified. Until such tests have been carried out, the dualist thus seems justified in holding that if Koksvik (2007b, pp.580) is correct in claiming that "[t]he evidence we have for CoE is *only* evidence for the restricted version," then the evidence for CoE is consequently much less conclusive than is typically thought. For the restricted formulation of CoE could very well be falsified when tested in those conditions where its failure seems most likely, thereby requiring us to either expand the scope of CoE and postulate certain non-physical entities that exchange energy with physical entities in such a way that energy is conserved in the universe as a whole, or else allow that energy is not conserved after all. At any rate, since current evidence does not exclude the possibility that violations of the restricted version of CoE may be discovered in the brains of minded beings, and it is this possibility that seems to pose the greatest threat to the idea that energy is conserved in the physical realm, current evidence seems insufficient to justify acceptance of CoE as restricted to the physical realm. If this is so, then any attempt to infer (iv) from the restricted formulation of CoE will be equally suspect.

In response, one could argue that while we are at this point still unable to determine whether or not violations of the restricted version of CoE occur in the brain, we nevertheless have enough evidence indicating that other kinds of physical systems do not undergo any losses or gains of energy that cannot be accounted for in terms of

compensatory exchanges of energy with their physical surroundings to enable us to infer that energy is conserved in the physical realm as a whole. Even if the restricted CoE has yet to be tested in the conditions under which it stands the greatest risk of being falsified, the wealth of other evidence in its favor might thus be thought to give us sufficient reason to accept it as true. Against this proposal, however, dualists can point out that since, on their view, the mind is (in certain significant respects) *unlike* any physical thing, those physical systems (viz. brains) that *interact* with minds should consequently be expected to behave in ways that other physical systems do not. It is, indeed, for this very reason that any evidence falsifying the restriction of CoE to the physical realm seems most likely to be found in the brains of minded beings, rather than in other types of physical systems; viz., because unlike other types of physical systems, brains (from the dualist's viewpoint) are capable of directly affecting and being affected by non-physical minds. If this is so, however, then brains are themselves importantly *disanalogous* with other types of physical systems, so the fact that other types of physical systems behave in a manner that is consistent with the restricted version of CoE gives us little reason to assume that the same will prove true of brains as well. To assume that brains are similar enough to other types of physical systems to enable us to infer that brains (and indeed *all* physical systems) behave in a manner consistent with the restricted CoE simply because other types of physical systems have been found to so do is thus to beg the question against the interactionist dualist, who is apt to deny the assumption on which this inference rests. In sum, until our understanding of the brain reaches the point where we can rule out the possibility that the energy levels of the brains of minded beings can change in ways that are inconsistent with the assumption that energy is conserved in the physical realm,

dualists seem entitled to maintain that our current evidence is inadequate to justify belief in the restricted version of CoE or the corresponding assumption that the physical realm is a closed system.

2.v. Does all causation of bodily motion involve transference of energy?

If the foregoing considerations succeed in showing that (iii) and (iv) are at least contestable, the dualist of course has no need to also call (ii) – the claim that all changes of bodily motion involve a transference of energy between the cause of the change and the body whose motion is altered – into question in order to diffuse the arguments against their position from (ii) and (iii), and (ii) and (iv).[175] Nevertheless, it at least worth noting that (ii) is rather dubious as well. The idea that any change in a body's motion requires some transference of energy between that body and the cause of the change is naturally viewed as an expression a more general Transference theory of causation of the sort advocated by David Fair (1979), Wesley Salmon (1997), and Phil Dowe (2000, ch.5), according to which causation is itself simply the transference of some conserved quantity (e.g., energy) from one thing to another. One of the more common objections to such theories is that the analysis of causation they propose seems applicable only to concrete physical entities, and is hence unduly narrow.[176] In particular, it is difficult to see how

[175] Hence Fair's (1979, p.237) remark that "the theory that causation is a matter of energy flow might be compatible with certain forms of dualist interactionism."

[176] Broad (1925, pp.107-8), however, provides an apparent counterexample to Transference theories involving a causal interaction between two purely physical entities.

Transference theories can accommodate instances of causation by omission, since absences (being non-existent) cannot have or transfer conserved quantities.[177] Assuming, therefore, that it could in certain circumstances be truly said that the death of a plant was caused by the gardener's failure to water it, or that a person's religious experience was caused by a lack of food, such cases would appear to constitute straightforward counterexamples to any Transference account of causation.[178] The same objection can likewise be applied directly to (ii), for the motion of a body would likely be significantly altered if it were to enter a void.[179] Due to the absence of surrounding air pressure, a human body placed in such conditions would quickly burst into bits. This would certainly constitute a change in the body's motion, and the surrounding void seems clearly to be the cause of this change, yet since the void does not exist, the effect it has on the body's motion cannot involve any transference of energy to or from the void itself. If this is correct, though, then (ii) is false.

2.vi. Is interactionist dualism compatible with (iii), (iv) and CoE?

Before concluding the present discussion, a few words are in order regarding a certain proposal of C.D. Broad's (1925, p.109), which warrants mention inasmuch as it may seem to give dualists a way of rendering their position compatible not only with

[177] Fair (1979, pp. 245-8) attempts to expand his Tranference account to cover cases of causation by omission, but requires the use of counterfactuals in order to do so.

[178] Not everyone is convinced that such cases do constitute genuine instances of causation. See, e.g., Beebee (2004).

[179] The example is drawn from Lewis (2004).

CoE, but also with (iii) – the claim that nothing non-physical has energy – and (iv) – the claim that the physical realm is a closed system. Having this option at their disposal would of course prove useful for dualists should the arguments offered against (iii) and (iv) above fail to convince. The basic outlines of Broad's proposal are as follows: Assuming both that the physical realm is a closed system in which energy is conserved, and that no energy is transferred between the mind and any physical thing, Broad maintains that the mind could still causally influence bodily motion by "determin[ing] that at a given moment so much energy shall change from the chemical form to the form of bodily movement…without altering the total amount of energy in the physical world."[180] Since the kind of situation Broad envisions seems consistent with (iii), (iv), and CoE, it appears that further assumptions are needed to derive a valid argument against interactionist dualism from these premises. If the forgoing discussion is any indication, we might expect it to be rather difficult to find a set of assumptions that are strong enough to rule out Broad's hypothesis when conjoined with (iii), (iv), and CoE without also being at least as questionable as (iii) and (iv) were shown above to be. Thus, while Broad's model of psychophysical causation could be excluded by adopting a Transference theory of causation such as that expressed in (ii), such theories are, as previously noted, open to a number of criticisms that would have to be answered before they could be reasonably accepted as grounds for rejecting Broad's proposal. One might

[180] A similar suggestion is made by Larmer (1986, pp.281-2).

naturally wonder whether the same might not be true of any other assumption added to (iii), (iv), and CoE with this end in mind.[181]

There is, however, at least one potential criticism of Broad's model of psychophysical causation that does not seem open to this objection, which is that while consistent with CoE, Broad's hypothesis still stands in at least apparent conflict with the law of conservation of momentum. This is because it is difficult to see how the mind could alter the distribution of energy in a physical system, or the rate or time at which it is transferred or converted from one body or form to another, without also changing the speed or direction of the motion of certain bodies within the system. But if the consequent changes in the momentum of these bodies is indeed caused by some non-physical mental state (and thus cannot be fully accounted for in terms of the interaction of these bodies with other bodies), then it seems inevitable that those bodies whose motion is altered by mental causes will undergo changes of momentum that are not accompanied by the kind of compensatory changes in the momentum of other bodies needed to ensure that momentum is conserved.

Averill and Keating (1981) respond to this objection by arguing that since the law of conservation of momentum only requires that momentum be conserved in systems that are subject to no external force, Broad's proposal does not violate that law, for even if his hypothesis does imply that momentum is not conserved in the brain, or in the physical

[181] To this point, Gibb (2010) offers up two premises that *would* yield a valid argument against interactionist dualism if conjoined with (iii), (iv), but finds these premises "dubious," primarily because they "cannot be established by appealing to the energy transference theory of causation" and "if these premises are instead to be inferred from facts within physics then it is unclear what these facts are" (pp.376, 382).

realm as a whole, the brain and hence the physical realm as a whole are, on Broad's view, subject to certain external, non-physical forces[182], and so momentum needn't be conserved in those systems in order for the law of conservation of momentum to hold valid. While Averill and Keating provide no explicit argument for their claim that physical systems are, on Broad's view, acted upon by external, mental forces, their reasoning seems to be that since acceleration is by definition directly proportional to force, anything that causes an acceleration must do so by exerting some force on the accelerated object.[183] Hence, insofar as Broad's account entails that mental events cause accelerations in certain bodies, it must also require that the physical realm is subject to certain non-physical, mental forces whereby mental events produce such accelerations.

While this reasoning shows that Broad's model of psychophysical causation is consistent with both the law of conservation of momentum and CoE, it also shows that upon closer examination, Broad's model turns out to be inconsistent with both (iii) and (iv). For to allow that changes in bodily motion have non-physical causes is to allow that the physical realm is subject to non-physical forces, in which case (iv) is false. And as it is difficult to see how anything could exert a force without possessing energy, to allow that non-physical, mental events exert forces is (*pace* Averill and Keating) to allow that such events have energy, in which case (iii) is false. If this is correct, then contrary to first appearances, Broad's proposal cannot be adopted without rejecting both (iii) and (iv).

---

[182] See also Larmer (1986, p.282). Given the crucial role that this premise plays in Averill and Keating's argument, it is somewhat puzzling why Gibb (2010, p.379) should cite them as providing a forceful defense of Broad's proposal against the objection noted above, while also claiming that Broad's proposal entails that "an entity can cause...redistribution [of energy and momentum] without exerting a force."

[183] Or, in the case of accelerations caused by omission, by failing to exert such a force.

While this does not mean that Broad's proposal is false, it does mean that dualists will have to look elsewhere if they wish to find a formulation of their position that enables them to accept the conservation of energy and momentum along with (iii) and (iv).

In light of the various objections we have now considered to the arguments against interactionist dualism from CoE, (ii), (iii), and (iv), the following conclusions seem warranted:

- CoE does not provide clear support for (2), as there are formulations of interactionist dualism that reject (2) without violating CoE.
- Any direct argument against interactionist dualism from CoE will likely end up being invalid unless it makes use of certain additional premises (viz. (ii), (iii), and/or (iv)), which, once added, render CoE's role in the argument redundant.
- (ii), (iii), and (iv) are disputable, and are not clearly supported by current science.

If these conclusions are correct, then interactionist dualism does not necessarily violate conservation laws, and it may likewise be possible for dualists to reject (2) without thereby setting themselves at odds with current science. Should the various attempts to respond to the Exclusion Argument by rejecting (3) prove inadequate, interactionist dualists can hence take comfort in the fact that an alternative response to the Argument may be available to them that is at least as consistent with current science as any of the assumptions on which the Argument itself is based.

3. Two strategies for rejecting the causal self-sufficiency of the physical

Thus far in the present chapter we have considered a number of strategies that dualists might employ to show that their position is consistent with (2), or to question whether support for (2) can indeed be drawn from conservation laws. Aside from a prior commitment to interactionist dualism, however, no positive reason has yet been given for thinking that (2) is in fact false. And while there is of course no need for dualists to dispute (2) if the arguments of section 1 succeed in showing that any non-question begging formulation of (2) is compatible with their position, it would nonetheless be a major asset for them to have a direct argument against (2) at their disposal. For if (2) can be shown to be false, then dualists needn't be constrained to show that the models of psychophysical causation they propose can be made consistent with it, and the Exclusion Problem would likewise be made much more tractable, since non-overdetermining mental causes could then be posited for those physical effects that lack a sufficient physical cause. With these points in mind, the present section will briefly lay out two arguments that to me seem to provide the most promising grounds for rejecting (2). The first appeals to certain discoveries in quantum physics, and the second makes use of certain arguments developed by Nancy Cartwright (1994) regarding the scope and status of scientific laws.

3.i. The argument from quantum mechanics

There are two main pieces of evidence that quantum physics seems to offer against (2), one of which concerns the dual wave-particle nature of quantum phenomena and the probabilistic character of quantum state descriptions, and the other of which concerns the interpretation of the effect of measurement on the wave function that defines a quantum state.[184] Both of these pieces of evidence can be understood through a close analysis of the famous "double slit" experiment, wherein a stream of particles (e.g. photons or electrons) is aimed at a screen containing two slits, behind which another screen is placed that detects the location of those particles that pass through the slits in the first screen at the point where they make contact with the second. It is important for this experiment that the size of the slits in the first screen be small enough to cause the waves that pass through them to diffract, and the slits must also be placed close enough to one another that the two diffracted waves exiting the slits overlap and interfere with one another. The effect of this interference is then revealed on the second screen by an increased density of particles in those regions where constructive interference occurs, and a decreased density of particles in regions of destructive interference.

[184] The following discussion relies heavily on Chalmers (1996, ch.10), Ghirardi (2005), and Maudlin (2011).

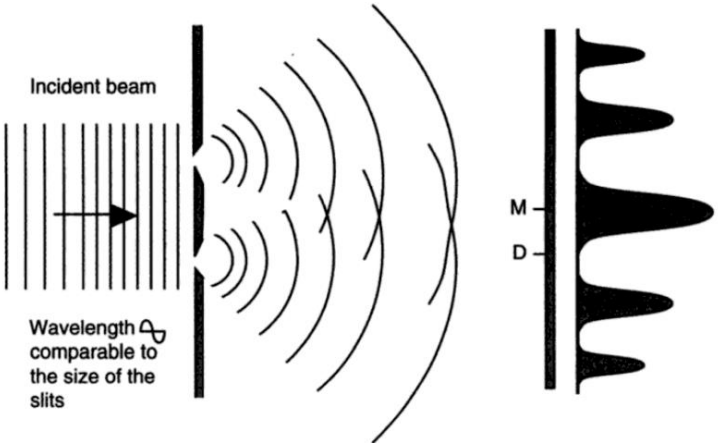


Figure 6. Experimental set-up for the double slit experiment (Ghirardi, 2005, p.14). M and D respectively identify areas of constructive and destructive interference between the two diffracted waves produced by the two slits in the first screen.

The odd thing is that this pattern of interference still manifests itself when the incident beam of particles is made weak enough that only one particle passes through the experimental setup at a time (so that, e.g., a particle would not arrive at the first screen until enough time had passed for the one that preceded it to pass through one of the slits and strike the second screen). This is strange, for if each particle that makes it past the first screen is the only particle between the first and second screens during its passage from the former to the latter, there seems to be nothing that could interfere with its trajectory so as to make it more likely to strike the second screen in the regions defined by the area of constructive interference. Yet as particle after particle strikes the second screen, a pattern emerges that displays the tell-tale signs of such interference.

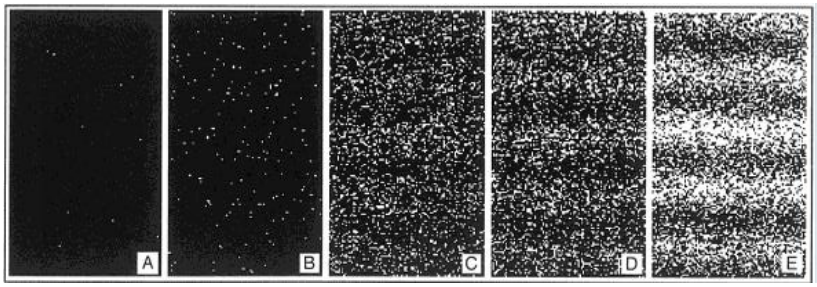


Figure 7. Gradual accumulation of particles on the second screen indicating the effects of interference in the double-slit experiment (Ghirardi, 2005, p.50).

The quantum mechanical solution to this oddity is to describe the state of a particle that makes it past the first screen as being in the *superposition* of the two diffracted waves that emerge from each slit. Under such a description, it is illegitimate to refer to the particle as passing either through one slit or the other (or through both or neither). Its state is instead defined in terms of a wave function that is the sum of the wave functions of the waves emerging from each slit, "reduced by a factor of $1\sqrt{2}$" (Ghirardi, 2005, p.91). This function can be represented by the following equation, where Ψ is the wave function that defines the particle's state as it passes the first screen, and the notation "|x>" identifies x as a property of the system being defined:

$$|\Psi> = (1\sqrt{2})[|\text{in slit 1}> + |\text{in slit 2}>]$$

The square of the wave function indicates the probability of obtaining a certain result upon measurement of the particle while it is in the state defined by the function. In the case at hand, the wave function thus tells us that any measurement of the particle's position at the moment it passes through the first screen will have a 50% chance of locating it in slit 1, and a 50% chance of locating it in slit 2.

This feature of the wave function becomes extremely significant when coupled with another basic principle of orthodox quantum theory, which is that the wave function of any quantum system provides a complete description of that system's state at a particular time. All that can be truly said about the state of any quantum system at a given point in time is hence contained in the wave function that defines it. This means that the probabilistic element in the quantum mechanical description of a given system cannot be attributed to some merely epistemic factor, such as our ignorance of certain "hidden variables" that would, if known, enable us to give a more exact, non-statistical description of its state[185], but must instead be taken as representing a real feature of the state being described. Consequently, although the wave function that defines a quantum system evolves according to strictly deterministic laws (which are encapsulated in the Schrödinger equation that describes the evolution of such functions over time), the quantum mechanical description of any system (as expressed in its wave function) is nonetheless inherently statistical; it tells us only the probability of obtaining a certain result if we were to make a certain measurement of some property of the system at a particular time.

[185] Einstein, Podolski, and Rosen's famous 1935 paper arguing for the incompleteness of quantum mechanics inspired a number of "hidden variable" theories (e.g. Bohm's "pilot wave" theory), which sought to account for the probabilistic elements in quantum state descriptions by postulating additional *epistemically* inaccessible variables whose values, if known, would allow the results of any measurement to be predicted with absolute certainty. Bell's inequality, however, raised a serious problem for such theories, by demonstrating that "a deterministic and local theory with predictive capability equal to mechanics cannot exist" (Ghirardi (2005, p.237). While this does not exclude the possibility of a hidden variable completion of quantum mechanics, it does mean that any such theory must necessarily be non-local, which seems at odds with the "classical" motivations for developing such theories in the first place. Such theories also face major difficulties in being made consistent with relativity. (See Chalmers (1996, pp.343-5), Ghirardi (2005, pp.381-3, 434-5), and Maudlin (2011, pp.215-6).)

At this point, one may begin to see how the statistical character of quantum mechanics might lend itself to an argument against (2). If the complete description of any quantum state necessarily contains certain ineliminable statistical elements, then it seems that whenever a measurement shows a quantum system *S* as having a certain definite property *P* at a certain time *t*, and according to *S*'s wave function at the preceding moment *t*-1, the probability that *S* would have been found upon measurement at *t* to possess *P* is < 1, the physical event consisting of the possession of *P* by *S* at *t* will lack a sufficient physical cause. To put this in terms of the double-slit experiment, given that the wave function of a particle at the moment it passes through the first screen is a superposition of the states |in slit 1> and |in slit 2>, if a measurement of the particle's position immediately after it passes through the first screen were to place it immediately behind slit 1, the event of its being there would lack a sufficient physical cause, since there is nothing in the particle's preceding state that could guarantee that this measurement would produce that result (as there was only a 50% chance that a measurement of the particle's position would locate it in slit 1). The basic principles of quantum mechanics hence seem to entail that (2) is false.

One might wonder, though, whether the quantum evidence against (2) considered thus far is really any help to dualists, for while this evidence indicates that there are certain physical events that lack a sufficient physical cause, these are not the type of physical events that dualists are most eager to posit mental causes for. When dualists claim that mental events cause physical effects, the effects they have in mind are not things such as the position or momentum of individual particles, but rather certain features or instances of animal behavior. It is hence unclear what consolation dualists can

draw from the falsity of (2) if the only counterexamples to (2) all appear at the quantum level, for the only kinds of counterexamples that would seem to be of real use to them would be ones involving the bodily motions of complex organisms. Moreover, the evidence presented above gives us no reason to suppose that those physical events that have no sufficient physical cause must instead be caused by some non-physical mental event. Equally possible is that such events are simply uncaused, in which case they would not qualify as counterexamples to (2) at all, (since (2) asserts merely that every physical *effect* has a sufficient physical cause).

There is, however, a second piece of evidence that quantum physics furnishes against (2) that might provide the basis for a response to at least the latter of the two concerns just raised, inasmuch as it may be taken to indicate not only that (2) is false, but also that certain mental events are needed to produce those physical effects that have no sufficient physical cause. This second piece of evidence concerns the interpretation of the effect of measurement on the state of a quantum system and the wave function that describes it. Opinions on this issue vary greatly, but one fairly standard view is that the act of measurement causes an instantaneous "collapse" of the wave function that defines a quantum system, wherein the system takes on a determinate value for the measured property, even if prior to the act of measurement, the system (as regards that property) was in a superposition of multiple distinct states.[186] Thus, in the case of the double slit experiment discussed above, if a particle's position were measured as it was passing

[186] Other prominent interpretations of quantum mechanics, e.g. Bohm's "pilot wave" theory, Everett's "many worlds" interpretation, and Albert and Loewer's "many minds" interpretation, deny that there is any such collapse or reduction of wave functions.

through the first screen, thereby locating the particle in slit 1, that act of measurement would be said to cause the wave function defining the particle's state to change instantly from:

$$|\Psi\rangle = (1\sqrt{2})[|\text{in slit 1}\rangle + |\text{in slit 2}\rangle]$$

to:

$$|\Psi\rangle = |\text{in slit 1}\rangle$$

While this interpretation of the relation between measurement and the state of a system is, again, fairly common, those who favor it must answer the difficult question of just what it is about acts of measurement that causes this striking, instantaneous reduction of superposed states to fully determinate ones. Although by no means universally held, one prominent view (often associated with Eugene Wigner (1961)) is that the collapse of the wave function is caused by the interaction of the measured system with the conscious mind of the observer carrying out the measurement. On this view, not only does the mind have an impact on the course of physical events, it is ultimately responsible for any determinacy we encounter in the physical world, for if there were no conscious mind to perceive it, the world would be left to develop into a massive, indeterminate tangle of superposed states.

There are, however, a number of potential problems with such "consciousness causes collapse" (CCC) interpretations of the effect of measurement on the wave function. First, it seems highly improbable that prior to the emergence of the first conscious being, the universe was completely indeterminate, existing solely as a vast, complex superposition of an infinite number of different possible states, and that at the moment of the first conscious perception, a momentous collapse of the universe's wave

function suddenly occurred, marking the first appearance of any kind of determinacy in the physical world. Adherents of CCC, however, seem forced to accept this as a consequence of their view.[187] A second problem for CCC is raised by the fact that the borderline between consciousness and lack of consciousness is itself unclear, making it doubly difficult for advocates of CCC to date the first appearance of determinacy in the universe, since they will first have to determine what "level" or "degree" of consciousness is needed in order to cause a wave function collapse, what physical conditions must be satisfied in order for such a consciousness to emerge, and how these conditions could be satisfied in the highly indeterminate, superposed state of the universe prior to the first conscious perception.

Lastly, with respect to our present concern, it is also uncertain whether CCC supplies dualists with the materials needed to develop a satisfactory solution to the Exclusion Problem, for while it clearly allows for mental causation of physical effects (since according to CCC, virtually every instance of conscious perception causes the physical system being perceived to take on certain determinate properties that it previously did not possess), it is still unclear how the kind of psychophysical causation that CCC entails could be used to provide mentalistic causal explanations of animal behavior. The main reason for this is that the physical effects that CCC attributes to mental causes are not features or instances of animal behavior, but rather things such as the position, spin direction, momentum, or polarization of quantum particles, and while the bodily motions involved in animal behavior are of course constituted by changes in

[187] See Ghirardi (2005, p.403) and Chalmers (1996, pp.339-41), who also raises some additional objections to CCC.

the states of the particles that the moving body is made up of, some work would nonetheless have to be done to show how the mind's alleged ability to affect the state of quantum systems translates into an ability to cause things like utterances, jumping-jacks, or chess moves. The notorious difficulty in understanding the relation between microscopic and macroscopic phenomena and the quantum/statistical or classical/deterministic laws that govern them suggests that this would be no easy task.

A second reason why CCC may not be all that helpful to dualists is that the physical effects that CCC posits mental causes for are typically external to the body of the minded being that produces those effects. Thus, in the case of the double slit experiment, the observer's conscious perception of the particle's location (or of the output of the measuring device that indicates its location) is said to produce an immediate effect on the state of that particle, even though that particle is not part of the observer's own body. The dualist, however, is most interested in explaining how an animal's mind might have some causal impact on its own bodily states. According to CCC, such cases would seem to require some form of self-measurement or observation, whereby an individual's conscious perception of his/her own body causes the wave function that defines its quantum constituents to collapse in certain ways. But this would mean that I cause my body to move only by perceiving it as moving, which sounds strange, to say the least. The oddity is compounded by the fact that in order to affect the movement of my body, my mind must surely first produce some effect on the state of my brain, but it is especially unclear how I could have the kind of conscious perceptions of my own neural states needed for my mind to directly influence my brain states in the manner suggested

by CCC. [188] It seems that to obtain the requisite perceptions, I would either have to be constantly monitoring a readout of a brain scan of my own brain, or else consciousness itself would (as Spinoza thought) have to be a kind of self-monitoring activity that generates nothing but perceptions of the internal states of one's own body (or brain). The former proposal, however, is absurd, and the second has the implausible consequence that I am never directly conscious of anything outside of my own skin.

In sum, while quantum physics seems to provide clear counterexamples to (2), it remains uncertain whether these counterexamples are the sort that would be useful to dualists in developing a theory of how non-physical mental states cause the physical effects involved in animal behavior. The only interpretation of quantum mechanics that gives the mind a direct causal role in producing those physical events that quantum theory treats as lacking any sufficient physical cause (viz. CCC) is, moreover, open to a number of objections, and it is at any rate still unclear whether the kind of psychophysical causation that this interpretation posits is the kind that dualists are most interested in providing for. Although quantum physics may help undermine the implicit faith in (2) that is engendered by a more classical, deterministic worldview, it thus appears inadequate to provide dualists with a fully satisfactory solution to the Exclusion Problem on its own.

3.ii. The argument from Cartwright's challenge to the universality of physics

[188] See Chalmers (1996, p.157).

The second type of argument that dualists might advance against (2) appeals to certain ideas proposed by Cartwright (1994) in order to contest the assumption that the laws of physics are universally valid.[189] This constitutes a challenge to (2), inasmuch as the plausibility of that thesis relies heavily on the idea that the scientific laws that describe the behavior of physical phenomena hold quite generally. For if this were not the case (if, e.g., the various physical laws that explain the motion of bodies only applied under certain restricted conditions), then while our current scientific laws may be capable of providing a complete physical explanation for any physical effect that occurs under conditions where they do apply, it is difficult to see what justification there could be for assuming that there must also be a sufficient physical cause for any physical effect that occurs under conditions that fall outside their scope of application. Dualists can hence raise doubts about whether there is in fact sufficient evidence for (2) by questioning whether the laws of physics are indeed universal.

The reasons that Cartwright (1994, p.282) supplies for questioning the universality of physics start from the simple observation that in order for the laws of a particular theory to be applied to any given situation, it is necessary to "first produce a model[190] of the situation in terms the theory can handle." Thus, in order to explain the trajectory of a moving body in terms of classical mechanics, one must first construct a model wherein the various forces acting on the body over the course of its trajectory are

[189] The possibility of using Cartwright's ideas to argue against (2) is also noted by Vicente (2006, pp.154, 156-7).

[190] Here the term "model" is used to refer to abstract representations of concrete systems, not (as in the discussion of Bickle's work in Chapter 5) to the concrete systems that satisfy or serve as instances of a particular theory.

quantified, thereby making it possible to use Newton's second law to account for the object's acceleration at each moment by dividing the vector sum of forces then acting upon it by the object's mass. Given this fact, the trouble Cartwright identifies with the presumed universality of physics is that the kinds of situations that can be adequately modeled in terms of current physics are fairly limited, and are indeed largely restricted to the sorts of highly controlled and idealized conditions that can be produced in a laboratory setting. This raises the question, though, of what justification we could have for assuming that the physical laws that are tested and confirmed under such conditions also hold true in more complex situations for which we have no suitable physical models. As an empiricist, Cartwright's answer is: none. For without the requisite models, the laws of physics cannot be applied to such situations or used to generate predictions that can be tested against observation of how the entities involved in those situations actually behave. Until such models are produced, the claim that the laws of physics hold everywhere and under all conditions is hence, for Cartwright (1994, pp.284-5), a mere "expression of…faith," which rests on the unwarranted assumption that there is "in principle" a physical model for any physically possible situation we might imagine. All we can justifiably assert of the laws of physics is thus that they are true ("literally true") for those types of situations that can be adequately modeled in physical terms.

Cartwright's conclusion is a striking one, and if true, undermines much of (2)'s plausibility. Even granting that her reasoning is sound, though, one might still wonder whether Cartwright's argument can really be all that useful to dualists in responding to the Exclusion Problem, for while her argument clearly suggests that there may be physical effects that have no sufficient physical cause, it does not seem to supply any

grounds for thinking that such effects may instead be caused by non-physical, mental events. Such grounds can, however, be found in the positive, "patchwork" picture of laws that Cartwright (1994, pp.281) offers in place of the "fundamentalist" doctrine that the laws of physics (alone) "hold everywhere and govern in all domains." On this alternative view, (which she dubs "metaphysical nomological pluralism"), "nature is [instead] governed in different domains by different systems of laws not necessarily related to each other in any systematic or uniform way: by a patchwork of laws" (Cartwright, 1994, pp.288-9). Appealing to this idea, the dualist might suggest that while most physical systems are fully governed by the laws of physics, there are nonetheless certain physical systems (e.g. those that make up the body of a minded being) that cannot be adequately modeled except in psychological terms. These systems, the dualist might say, are consequently more properly treated as subject to the laws of psychology than those of physics. And inasmuch as these laws ascribe some of the physical effects that occur within such systems to mental causes, we have reason to believe that there are certain physical effects whose causes are mental rather than physical in nature. By conjoining Cartwright's metaphysical nomological pluralism with the existence of psychological laws that attribute physical effects to mental causes, and physical systems that are best modeled in terms of the psychological sciences to which such laws belong, the dualist can thus mount an argument that purports to show not only that (2) is false, but that (1) and (4*) are true.

The most likely objection to Cartwright's proposals concerns her inference from the current lack of any good physical models for certain situations to the conclusion that there may very well be no such models, and that the laws of physics consequently may

not apply in such cases. While few would deny that a close-fitting physical model has yet to be produced for every physically possible situation imaginable, those who would take this as grounds for seriously questioning the universality of physical laws are probably equally small in number. As Cartwright (1994, p.284) notes, the thinking behind this orthodoxy appears to be that the "[t]he successes of [physics] in situations that it can model accurately" provides sufficient inductive evidence for the idea that even those situations for which no adequate physical model has yet been given nonetheless have such a model "in principle," "albeit probably a very complicated one that we may never succeed in constructing." To this, however, Cartwright is I think justified in responding that such successes "show only that [physics] is true in its domain, not that its domain is universal." After all, from the standpoint of a strict empiricism, the only sure way of demonstrating that physics is universal is to show that the laws of physics yield observationally accurate predictions in every type of situation, and to do this one must first be able to provide a physical model for all the various types of situations that can arise. Until such models are constructed, the supposed universality of physics will hence at best be empirically underdetermined. Those dualists who are willing to accept Cartwright's rather stringent empiricist standards of evidence are therefore entitled to maintain that, pending the construction of a complete physical model describing the bodily behavior of a healthy, mature human being, many of the physical effects involved in such behavior are more reasonably viewed as governed by psychological rather than physical laws, since it is psychological rather than physical models that currently provide the most accurate and reliable predictions of how people act.

Of the two strategies we have now considered for challenging (2) the Causal Self-Sufficiency of the Physical, the argument from Cartwright's views I think holds more promise for dualists seeking a solution to the Exclusion Problem that is consistent with (1) and (4*). This is chiefly because, as noted above, her proposals can be used not only to undermine the credibility of (2), but also to support the idea that certain components or features of animal behavior are caused by mental, rather than physical events. This makes the appeal to Cartwright's views more useful to dualists than the argument from quantum mechanics, for as previously mentioned, it is rather unclear how the apparent counterexamples that quantum physics provides to the Causal Self-Sufficiency of the Physical might be developed into an argument for the view that animal behavior has non-physical, mental causes. Consequently, while Cartwright's suggestion that the laws of physics hold only in certain restricted domains may be hard for some to swallow, it nonetheless strikes me as the best way for dualists to go about rejecting (2).

The present chapter has presented a variety of ways in which dualists might respond to the Exclusion Problem by putting pressure on (2), the Causal Self-Sufficiency of the Physical. In section 1, it was shown that depending on how it is formulated, (2) either begs the question against interactionist dualism, or else is compatible with the conjunction of (1) the Causal Efficacy of the Mental, (3) the Absence of Systematic of Overdetermination, and (4*) Mind-Body Dualism. Section 2 then questioned the support that (2) is often thought to draw from conservation laws, and contended that direct conservation-based arguments against interactionist dualism are either invalid or else rely on suspect premises. Lastly, section 3 provided two strategies for rejecting (2) outright,

one of which draws on certain discoveries in and interpretations of quantum physics, and the other of which appeals to Nancy Cartwright's empiricist skepticism about the universality of physics. Having thus considered a range of arguments that collectively seem to show (a) that (2) is not clearly supported by evidence typically adduced in its favor, (b) that there are moreover positive reasons to think that (2) is false, and (c) that even if true, any non-question begging formulation of (2) is still compatible with interactionist dualism and the assumption that the effects of mental causes are not overdetermined, dualists can I think reasonably claim that the alleged Causal Self-Sufficiency of the Physical raises no insurmountable problems for their position. Taking these points in conjunction with the previous chapter's arguments showing that (3) the Absence of Systematic Overdetermination is likewise either dubious or compatible with interactionist dualism (regardless of whether (2) is true or false), the dualist is now equipped with a range of strategies for responding to the Exclusion Problem without having to abandon either (1) or (4*). While the apparent tension between (1), (2), (3), and (4*) has thus been resolved, nothing has yet been done to specify precisely what is meant in saying that non-physical, mental events cause physical effects, or to demonstrate that such instances of causation do in fact occur. It seems to me that the best way to address this issue is to provide an account of causation under which (given our current understanding of the relations between the mental states and behavior of animals) mental events would qualify as causes of certain bodily states. It is to this task that I turn in the next and final chapter of the dissertation.

CHAPTER 9

THE CAUSAL EFFICACY OF THE MIND

Our discussion of (2) the Causal Self-Sufficiency of the Physical and (3) the Absence of Systematic Overdetermination in the preceding two chapters has shown that, first appearances notwithstanding, these theses do not pose any insurmountable threat to interactionist dualism. With respect to (2), we have seen that (a) the alleged causal self-sufficiency of the physical derives no clear support from conservation laws or any other well-established scientific principles, and (b) that (2) is moreover consistent, on some formulations, with the conjunction of interactionist dualism and (3). In regard to (3), it has been shown that depending on how overdetermination is defined, the claim that it is rare is either implausible or else consistent with the conjunction of interactionist dualism and (2). In light of these conclusions, it seems that interactionist dualists can regard the Exclusion Problem as effectively solved, for if the reasoning of the previous two chapters is sound, then the apparent tension between (2), (3), and the conjunction of (1) the Causal Efficacy of the Mental and (4*) Mind-Body Dualism can be reasonably resolved without rejecting (1) or (4*), but by instead either denying (2), denying (3), or denying that (1), (2), (3), and (4*) are in fact incompatible.

While dualists should certainly be encouraged by this result, critics may wonder whether the view has yet been fully vindicated. For it is one thing to show that (2) and (3) do not preclude the possibility that non-physical, mental events cause physical effects, but it is another thing entirely to provide positive reasons for thinking that the latter claim is actually true. Those swayed by the forgoing chapters may consequently remain

reluctant to allow that non-physical mental events can cause physical effects, for even granting that the Exclusion Problem does not provide any adequate reason for us to reject this possibility, some might nevertheless find it simply incomprehensible how something non-physical could do something like alter the motion of a body through space. This charge of incomprehensibility is perhaps the oldest and most deep-seated source of the skepticism with which dualist interactionism is now typically received. The fact that it persists even after the Exclusion Problem has seemingly been resolved should make it clear, however, that the challenge it raises to interactionist dualism is distinct from that raised by (2) and (3). Put simply, the problem that the Exclusion Problem poses for interactionist dualism is not that the causation of physical effects by non-physical causes is incomprehensible, but rather that such causes would be somehow superfluous or redundant. The objection raised by the charge of incomprehensibility, on the other hand, is not that non-physical causation of physical effects is superfluous or redundant, but rather that it is simply unintelligible, if not outright incoherent.

The only interpretation of the latter objection I can think of under which it does not simply beg the question against interactionist dualism is to treat it as a request for a plausible theory or definition of causation that allows for the causation of physical effects by non-physical, mental causes. Viewed in this way, the objection is quite reasonable, and the present chapter aims to satisfy the request it conveys by providing two different theories of causation under which non-physical, mental events can qualify as causes of physical effects without running afoul of the Exclusion Problem. This should, it is hoped, dispel any lingering doubts as to the ability of dualists to provide a satisfactory account of mental causation by showing not only that it is in principle possible for non-physical,

mental events to cause physical effects without generating any objectionable form of overdetermination, but that clear sense can also be made of the claim that they really do so.

1. Option 1: Mental causation as energy transference.

The first way in which interactionist dualists might explain how mental causes produce physical effects is to adopt a Transference theory of causation of the sort discussed in the previous chapter, according to which for one thing to cause another is for some conserved quantity (e.g. energy or momentum) to be transmitted from one to the other, and then propose that energy is sometimes transmitted between physical and non-physical, mental events. On such a view, non-physical, mental events could be said to cause physical effects by virtue of transmitting energy to, or receiving it from, physical events. If the conclusions of the previous chapter are sound, then there is nothing in known conservation laws, or indeed in any law or principle of current physics that prohibits this possibility, so long as the total energy prior to and after any transfers of energy between mental and physical events remains constant.

Though the ascription of energy to non-physical, mental states may seem outlandish, the view just described at least has the virtue of being empirically testable (in principle). For, as noted in the previous chapter, if the energy in the body of a mentally active human being was found to increase or decrease without its having lost or gained any energy from the surrounding physical environment, rather than taking this as a falsification of the law of conservation of energy, it would seem reasonable to instead

view the result as confirming the hypothesis that energy can be transmitted between physical and non-physical, mental states (especially if the subject reported the onset or cessation of some psychological process or event at roughly the same time that the amount of energy in his/her body was found to change). While certain empirical findings would thus tell rather strongly in favor of the proposed view, it must be conceded that the aforementioned hypothesis could also be developed in such a way as to render it immune to empirical falsification, thus making it (by Popperian standards) merely pseudoscientific. For even if no changes were ever detected in the total amount of energy in the body of a mentally active human being besides those due to exchanges of energy with its surrounding physical environment, an adherent of the hypothesis might still refuse to accept this as a refutation of their view by instead postulating that while energy is constantly being transferred between the subject's mind and body, the total amount of energy in the subject's body is unaffected by such exchanges for the simple reason that all losses of energy from the body due to the transference of energy from the body to the mind are perfectly compensated for by simultaneous gains due to the transference of energy from the mind to the body.

By making use of this stratagem, one wedded to the idea that physical states exchange energy with non-physical, mental states could always fix the quantities of energy transmitted between an individual's mind and body in such a way that no empirical discovery concerning changes of energy in the bodies of minded beings could ever prove that idea false. Properly regarded, however, such irrefutability would be quite damaging to the present proposal, for compatibility with any potential empirical finding is usually more indicative of a certain vacuity and lack of explanatory power in one's

hypothesis than of its proximity to the truth. Proponents of the present proposal would hence do well to seek some independent measure for the quantities of energy possessed and transmitted by mental states that does not simply ascribe to a person's various mental states whatever quantities of energy happen to be most consistent with whatever is known about the gains or losses of energy in his or her body.

Following up on an idea proposed by W.D. Hart (1988), a potential method for determining the quantities of energy possessed by intentional states would be to use the resources of decision theory to derive measures for an individual's various subjective credences and utilities, and then ascribe greater amounts of energy to those beliefs and desires whose corresponding credences or utilities are higher. Such an approach could then be expanded to other intentional states by analyzing such states in terms of beliefs and desires and then ascribing to them the sum or product of the quantities of energy possessed by the beliefs and desires in terms of which they are analyzed. Thus, the state of fearing that *p* might be analyzed as the conjunction of the belief that *p* (or possibly *p*) and the desire that ~*p*, with the quantity of energy ascribed to that fear in a given individual being determined by the weighted sum or product of the individual's subjective credence for *p* and their subjective utility for ~*p*. A measure for phenomenal states could in turn be derived from introspective reports of or third person tests for intensity and phenomenal salience, with greater quantities of energy being ascribed to those phenomenal states that are more intense, noticeable, and phenomenally salient.

While such methods may fall short of enabling us to specify the exact quantity of energy possessed by each of an individual's mental states, together they would at least provide us with an independent measure for the relative quantities of energy possessed by

his or her various phenomenal and intentional states across time and with respect to one another[191], and that seems to be enough to render the hypothesis that physical states exchange energy with non-physical, mental states empirically falsifiable. For with these measures in place, one cannot simply attribute whatever quantities one wishes to an individual's mental states; one must ensure that the relations between the quantities of energy one ascribes to those mental states are isomorphic to the relations between the credences, utilities, and degrees of phenomenal intensity or salience associated with each such state. Given the measures for mental energy proposed above (and assuming that energy is conserved), the hypothesis that energy is transmitted between physical and non-physical, mental states would hence be falsified if, e.g., the amount of energy in the body of a minded being were found to remain constant across some stretch of time during which it lost neither more nor less energy to its surrounding physical environment than it gained, while the credences, utilities and degrees of phenomenal intensity or salience associated with the mental states of that individual underwent a net increase or decrease. For in such a situation, the proposed measures for mental energy would require us to say that the individual's mind had gained or lost a certain quantity of energy that could not have been acquired from or transmitted to the individual's body. Conjoined with these measures, the suggestion that energy can be exchanged between mind and body thus becomes an empirically contentful hypothesis that is (in principle) open to both confirmation and refutation through testing.

[191] Note, however, that since the measures proposed for phenomenal and intentional states are distinct, they do not enable one to determine the relation between the amount of energy possessed by a phenomenal state and that possessed by an intentional state (unless both quantities are 0).

Having seen how specific (relative) quantities of energy might be assigned to each of a person's various mental states independently of the energy possessed by the states of his or her body, a few conjectures can now be made as to how mental and bodily states might causally interact with one another. First, in cases of intentional action, the beliefs and desires that figure into an accurate intentional explanation of an agent's behavior might be said to cause (at least some of) the bodily motions involved in that behavior by virtue of transmitting certain quantities of energy to various motor centers in the agent's brain, which are thereby stimulated in such a way as to initiate the ensuing action. (Alternatively, the energy transferred to the brain's motor centers might be transmitted from the agent's operative desires alone, while his/her beliefs merely determine *which* motor areas the energy that the motivating desires supply is transmitted to, thereby determining what means the agent employs to try and satisfy the desires that s/he acts on.) The beliefs and desires that cause an agent's intentional actions in the sense just mentioned could also perhaps be identified by first assessing the amount of energy possessed, at the time of the action, by (a) each of his/her desires D_1, D_2,…D_n, (b) his/her beliefs, for each of those desires and a range of potential actions φ_1, φ_2,…φ_n, that a given desire D_i can be satisfied by φ_k-ing, and (c) his/her desire not to φ_k,[192] and then combining the quantity of energy possessed by each desire D_i with that possessed by each belief $B\varphi_1$, $B\varphi_2$,…$B\varphi_n$ that φ_k-ing will lead to the satisfaction of D_i and that possessed by

[192] This last quantity may itself be determined by a complex weighted sum or product of the energy possessed by (d) the agent's intrinsic dislike of φ_k-ing, (e) his/her belief that φ_k-ing will lead to certain consequences that s/he desires not to occur, and (f) his/her desire that those consequences not occur.

each desire $D{\sim}\varphi_1$, $D{\sim}\varphi_2$,…$D{\sim}\varphi_n$ not to φ_k in a weighted sum or product of the following form:

$$xD_i + yB\varphi_k - zD{\sim}\varphi_k$$

or

$$(xD_i \times yB\varphi_k) \div zD{\sim}\varphi_k.$$[193]

It may then be that the desire that causes an intentional action always turns out to be identical with the D_i that figures in the largest such sum, while the belief that determines the means that the agent employs to satisfy that desire is identical with the $B\varphi_k$ that figures in that same sum.

In similar fashion, instances of involuntary action (e.g. the reflexive scratching of an itch, or wincing in response to pain) that are directly caused by phenomenal states without any intermediating beliefs or desires might be said to consist in the transmission of a certain quantity of energy from the phenomenal state that causes the action to the motor centers whose stimulation triggers the bodily motions involved in the action. In such cases, it may also be that, of the agent's various phenomenal states, the state that causes the action is always the one that possesses the most energy at the time of the action.

Cases involving the causation of conscious perceptual states through sensory stimulation might be described as follows: First, an external stimulus transmits a certain quantity of energy to a perceiver's sense organs, which respond to this stimulus by transmitting some pattern of energy to his/her sensory cortex. In processing the signal

[193] *x*, *y*, and *z* are presumed to be psychological constants, whose values can be discovered through empirical research.

received from the peripheral sense receptors, the sensory cortex then transmits certain quantities of energy to the various phenomenal states involved in the perceiver's perceptual experience. A further quantity of energy is then transmitted from the perceiver's brain to some perceptual belief, where the content of this belief and the amount of energy transferred to it depend upon both (a) the strength and nature of the phenomenal states caused by the activity in the perceiver's sensory cortex, and (b) the relative strength and content of the perceiver's memories and background beliefs. As the content and strength of the resulting perceptual belief depends (in part) on the strength and nature of the phenomenal states involved in the associated perceptual experience, which are in turn caused by a certain type of cortical activity, which itself depends upon the character of the external stimulus, the strength and nature of all the various phenomenal and intentional states involved in the perceiver's conscious perception would thereby depend ultimately on the nature and strength of the original stimulus, (and this in such a way that the strength of (i.e., energy possessed by) the relevant phenomenal states and perceptual belief will in general be directly proportional to the strength of the external stimulus that causes them, judged in terms of how much energy the latter transmits to the perceiver's sense organs).

Having sketched how some central cases of mental-to-physical and physical-to-mental causation might be described within the present framework, the same approach can be extended to instances of mental-to-mental causation as well. Consider, e.g., a thinker who acquires a new belief that *q* by reasoning from his/her prior beliefs that *p* and that *if p, then q*. It would seem natural, in such cases, to describe the thinker's newly acquired belief that *q* as having been caused by his/her beliefs that *p* and that *if p, then q*

(as well as perhaps his/her belief that *modus ponens* is a valid form of inference). Put in terms of the present model, this would be to say simply that the thinker's prior beliefs that *p* and that *if p, then q* transmit a certain quantity of energy to his/her new belief that *q*, thereby raising his/her credence in the proposition that is the object of that belief.

Further reflection on the case just described, however, reveals a potential problem for the account of mental causation currently under consideration. Put simply, the difficulty is that on the proposed account, mental states often[194] cause things by virtue of transmitting energy to their effects. But if, as previously suggested, the energy possessed by a mental state is proportional to its strength (measured in terms of its associated credence, utility, or phenomenal intensity and salience), the strength of any such mental cause will end up being depleted by the transference of its own energy to its effect(s). This is particularly worrisome in the case just considered, because (a) it seems empirically false that whenever a thinker acquires a new belief through inference from certain prior beliefs, the strength of those prior beliefs is thereby decreased, and (b) a rational thinker's credence in a conclusion arrived at through *modus ponens* should presumably be equivalent to the product of his/her credences in the premises of the inference, but this requirement seems impossible to satisfy if one cannot make an inference without thereby weakening the strength of one's belief in the premises. Thus, if one acquires the belief that *q* by inference from one's prior beliefs that *p* and that *if p, then q*, one's credence for *q* should be equivalent to the product of one's credences for *p* and *if p, then q*. Suppose that prior to making the inference, one's credences for *p* and *if*

[194] One may also want to allow for potential cases wherein a mental cause produces an effect on some system or state by "siphoning off" some quantity of energy from the latter.

p, then q were both 0.5, while one's credence for *q* was 0. The resulting credence for *q* should then be 0.25. Now assuming that any decrease in one's credence in a certain proposition *r* due to a transference of energy from one's belief that *r* to another belief is matched by a corresponding increase in one's credence in the proposition that is the object of the latter belief (and *vice versa*), it follows that in order for one's beliefs that *p* and that *if p, then q* to transmit enough energy to the belief that *q* to ensure that the resulting credence for *q* is 0.25, one's credences for *p* and *if p, then q* will between them have to be decreased by exactly 0.25. Suppose that each is decreased by 0.125. The resulting credence in *q* will then be 0.25, as it should. But now in the course of inferring *q* from *p* and *if p, then q*, the transference of energy from the beliefs that *p* and that *if p, then q* to the belief that *q* has led one's credences for *p* and *if p, then q* to drop from 0.5 to 0.375. The product of these credences is now *less* than 0.25, which is the resulting credence for *q*. Generalizing from this example, it appears that in any case wherein a new belief is acquired through inference from prior beliefs, the present account (as developed thus far) will not permit a rational assignment of credences to the propositions that are the objects those beliefs.

The problems raised for the present account by the loss of energy that mental causes would have to undergo in transmitting energy to their effects are not restricted to instances of mental-to-mental causation. Similar difficulties also arise for the descriptions of mental-to-physical causation offered above. For if the strength of one's desires and phenomenal states (judged in terms of their associated subjective utilities, or phenomenal salience and intensity) is directly proportional to the amount of energy they possess, some portion of which is transmitted to the brain whenever they cause a physical effect, it

follows that the strength of one's desires and phenomenal states must always be decreased by one's acting on them. But this seems empirically false. For while an initiation of bodily motion that is caused by a certain desire or phenomenal state may be attended in some cases by an immediate decrease in the strength of that desire or phenomenal state, it seems clear that on other occasions, the initiation of bodily motion does nothing to decrease the strength of the desire or phenomenal state that caused it, and may sometimes even *increase* its strength (as when, e.g., the anticipated satisfaction of a certain desire or the sense that one is getting closer to its object leads to a strengthening of that desire upon the initiation of a course of action aimed towards its satisfaction).

To accommodate such cases, as well as instances of mental-to-mental causation of the sort described above, the proposed model might be modified to allow those mental causes whose strength is not depleted in the act of producing their effects to derive further energy from their physical realizers while they are in the process of transmitting energy to their effects. This would enable the energy that such causes lose to their effects to be immediately replenished, and in some cases exceeded by the energy that they simultaneously acquire from the brain, so that the strength of the mental cause (as measured by its associated credence or utility, or its phenomenal salience and intensity) need not be diminished by the loss of energy it undergoes in causing the effects that it does. When causal processes of this sort persist long enough for the mental cause to go on transmitting energy to its effect after it has already transferred to it all of the energy it originally possessed, the mental cause may then take on the role of a kind of conduit, which exerts a continued causal influence over its effect by channeling energy from its

own physical realizers to some other part of the brain or some other mental state.[195] The more complex exchanges of energy that are postulated as taking place in such cases may sometimes be difficult, if not impossible to test for (as it would, e.g., be impossible to distinguish empirically between the conjecture that a certain brain state P_1 transfers a certain quantity of energy to some mental state *M*, which simultaneously transmits an equivalent amount of energy to another brain state P_2, and the conjecture that that same quantity of energy is simply transferred directly from P_1 to P_2). If, however, the more general hypothesis that physical states exchange energy with non-physical, mental events is confirmed under the kinds of experimental conditions previously discussed, and a subject's credences, utilities, and reports of phenomenal intensity and salience prove a reliable measure of the energy possessed by his/her various mental states at a given time, one might take this as reason to also accept the existence of more complex, less easily detectable exchanges of energy of the sort just described in those cases wherein mental causes transmit energy to their effects without thereby growing weaker.

2. Option 2: Mental causation as model-relative difference making.

The forgoing section has provided an answer to the objection that interactionist dualism is simply incomprehensible by furnishing a theory of causation under which non-physical, mental events may qualify as causes of physical effects. One may, however, be

[195] In the latter case, the situation will be further complicated by the fact that (assuming mind-body supervenience) the physical realizer of the mental state to which energy is transferred will also have to undergo some corresponding change, which will require that it too receive some quantity of energy from either the mental cause of the mental state it realizes or some other part of the brain.

understandably dubious about the prospects for interactionist dualism should this be the only account of mental causation available to its proponents, for (a) the idea that non-physical, mental events exchange energy with physical states is sufficiently unorthodox that few will be likely to consider, let alone accept it without very compelling reasons to do so, and (b) the Transference theory of causation that the account relies on also faces a number of potential problems, including (as previously noted) its inability to allow for instances of causation by omission. In view of this, the case for dualism would be much improved if an additional account of mental causation could be provided that avoided the drawbacks of the previous one while still allowing for the causation of physical effects by non-physical causes. Such an account can I think be drawn from the work of Peter Menzies (2003; 2004), who has developed a theory of causation that defines causes as things that "make a difference" to an effect relative to a certain contextually specified model.

The notion of difference-making that lies at the heart of Menzies' theory of causation is spelled out in terms of certain relations of counterfactual dependence. More specifically, an event *c* is said to "make a difference" to another event *e* iff, if *c* were to occur, then *e* would also occur, and if *c* were to not occur, then *e* wouldn't either.[196] The use of such counterfactuals in Menzies' definition of causation identifies his proposal as one of a family of theories that seek to analyze causation in terms of the counterfactual dependence of effects on their causes. The possibility of using such an analysis of causation to allow for non-reductive mental causation of physical effects has been the

[196] Note that only one of these conditionals will actually be counterfactual in any particular case of difference making.

subject of much debate.[197] The primary attraction that an account of this sort holds for those interested in defending both the causal efficacy and irreducibility of mental properties is that, unlike the theory of mental causation developed in the previous section, it does not require one to deny (2), the causal self-sufficiency of the physical, nor accept any contentious and speculative ideas about exchanges of energy between physical and non-physical states. For if causation is simply some form of counterfactual dependence, all one needs to do to establish the mind's causal efficacy is identify certain physical events that would not have occurred if some prior mental event hadn't, and there seems no reason why one should have to identify mental causes with physical events, deny that the physical effects of mental causes lack sufficient physical causes, or attribute energy to non-physical, mental states in order to do this.

Moreover, if the conditions on overdetermination (O1) and (O2) discussed in Chapter 7 are valid, it would also seem fairly easy to develop a non-reductive, counterfactual theory of mental causation that is consistent with both (2) *and* (3), the Absence of Systematic Overdetermination. For if overdetermining causes must be capable of occurring independently of one another, and causation is again merely some

[197] Some, e.g. LePore and Loewer (1987; 1989, pp.188-90), Horgan (1989, pp.56-64), and Loewer (2007, pp.255-9), maintain that the counterfactual dependence of physical events on mental events is enough to establish the causal relevance of the latter to the former. Others, e.g. Kim (1998, pp.67-72; 2007), Fodor (1989, pp.70-3), and Kallestrup (2006, pp.475-9) regard relations of counterfactual dependence as too weak to justify the postulation of any real causal relations between events (in part because the counterfactual dependence of physical events on mental events seems consistent with epiphenomenalism). Kim (2007, p.236) thus argues that "mere counterfactual dependence is not enough to sustain the causal relation involved in our idea of acting upon the natural course of events and bringing about changes so as to actualize what we desire and intend," and that the vindication of our sense of agency hence requires that there be some more substantial connection between mental causes and their effects, such that the former actually *generate* or *produce* the latter (by, e.g., transmitting energy to them). The Transference account of mental causation developed in the previous section and the hybrid Counterfactual/Transference account developed below are meant to satisfy those who are sympathetic to Kim's point.

form of counterfactual dependence, one might suggest that any physical effect *e* that is counterfactually dependent on, and thus caused by, some prior mental event *m* also depends counterfactually on, and is thus caused by, some other physical event *p*, which itself could not have occurred without *m* (and this not because *p*=*m*, but because there is some metaphysically necessary law that renders events of type *p* sufficient for the synchronic occurrence of events of type *m* under the type of conditions surrounding the occurrence of *e*). Counterfactual theories of causation thus also appear to offer dualists a simple way of implementing the solution to the Exclusion Problem proposed in Chapter 7, which would enable them to affirm the causal efficacy of the mind without having to contest either (2) or (3). Given the potential advantages of such an approach, it is unsurprising that counterfactual theories such as Menzies' are often viewed as the best option for dualists seeking an account of mental causation that is consistent with their position.

The *locus classicus* for counterfactual theories of causation is David Lewis' (1973a) paper "Causation," which provides a useful backdrop for understanding the details of Menzies' theory. Lewis' original proposal (which he later modified in order to handle certain difficulties noted below) was that an event *c* causes another event *e* iff *c* stands in the ancestral of counterfactual dependence to *e*, meaning that there is a chain of events running from *c* to *e*, each member of which depends on its predecessor in such a way that if its predecessor had not occurred, it wouldn't have occurred either. In accordance with the standard reading of counterfactual conditionals developed by Stalnaker (1968) and Lewis (1973b), an assertion of counterfactual dependence of the form $\sim c \ \Box\!\!\rightarrow \sim e$ qualifies as non-vacuously true iff some (accessible) possible world

wherein both *c* and *e* fail to occur is more similar to the actual world than any (accessible) world wherein *e* occurs but *c* does not. Since his proposed definition of causation makes use of such counterfactual conditionals, Lewis' truth conditions for statements of the form ⌜ *c* causes *e* ⌝ (hereafter "causal statements") can be seen to rely on an ordering of possible worlds with respect to their similarity to the actual world. This ordering, for Lewis, remains for the most part[198] constant, so that in normal cases (under what Lewis (1979, p.457) calls the "standard resolution of vagueness" of counterfactuals) all counterfactual conditionals, and thus all causal statements are to be assessed relative to the same ordering of worlds.

To ensure that his proposed definition of causation doesn't end up treating effects as causing their causes (as many Regularist theories of causation do), Lewis (1979) suggests that the closest world in which a given event *e* fails to occur will always be one whose history is exactly the same as the actual world's up until shortly before the time of *e*'s occurrence, at which point a small, localized "divergence miracle" prevents *e* from occurring. From that point on, the course of events in that world then proceeds as if it were henceforth governed by the same laws as those that obtain in the actual world, and will consequently diverge further and further from the actual course of events due to the non-occurrence of *e*. This proposal effectively rules out all "backtracking"

[198] *Pace* Menzies (2004, p.141), Lewis (1979, p.465) allows that "[d]ifferent resolutions of the vagueness of overall similarity [to the actual world] are appropriate in different contexts," and again, that "the appropriate similarity relation [to use when evaluating counterfactuals] will differ from context to context." (See also Stalnaker (1968, pp.109-10).) That said, Lewis does seem to regard all normal contexts in which causal statements are made as being governed by the same similarity relation, so that, in his view, it is only in very unusual contexts that the counterfactuals underlying such statements should ever be assessed relative to a different, "non-standard" ordering of worlds (and even in such cases, the relevant orderings will likely still be "centered" on the actual world).

counterfactuals, which assert the counterfactual dependence of the past on the future, by guaranteeing that all statements of the form $\sim c \ \Box\!\!\rightarrow \sim e$ wherein *c* actually occurs *after e* will turn out false, since the closest possible world wherein *c* fails to occur will be perfectly similar to the actual world until shortly before the time of *c*'s actual occurrence, and will therefore contain the occurrence of *e*. The asymmetry of causation is thus established under Lewis' analysis by the fact that the ordering of worlds against which counterfactual conditionals are assessed itself contains a certain temporal asymmetry, such that similarity with respect to the past ends up being given more weight in determining proximity to the actual world than similarity with respect to the future. Conjoined with the standard Stalnaker/Lewis semantics for counterfactual conditionals, this feature of the relation of similarity between worlds suffices to ensure that the past does not depend counterfactually on the future, and thus that prior causes are not counterfactually dependent upon (and hence, by Lewis' definition, not caused by) their effects.

While a number of problems have been noted in Lewis' theory, the most noticeable is that it seems far too permissive. Thus, if causation is the ancestral of counterfactual dependence, then my death will be caused by my birth, Wilt Chamberlain's 100 point game was caused by the Big Bang, and if, e.g., someone were to pull a switch diverting an incoming train to the left at a split in the tracks, and these split tracks were to reconvene shortly thereafter, uniting once more into a single track to which a poodle has been tied, the person's pulling the switch would be a cause of the poodle's demise, even though it would have been run over by the train whether the person pulled

the switch or not.[199] To explain why this abundance of causes is apt to strike us as excessive, Lewis (1986, pp.215-6) suggests that while causes are indeed as plentiful as his definition of causation implies, among the various causes of a given effect, we typically focus only on those that are most salient or relevant to our current explanatory or practical interests, and thereby pragmatically single out one event or small collection of events as *the* cause of the effect, relegating the remaining causes (which are in truth no less causes than the one(s) we've singled out) to the status of mere background conditions that enable *the* cause to do its work.

Although Lewis is thus able to account for the intuitions that weigh against the proliferation of causes under his proposed definition of causation by suggesting that for various pragmatic reasons, we customarily speak of only some of the actual causes of an effect as having actually caused it, this explanation requires us to accept that the seemingly quite objective (i.e., mind-independent) distinction between the causes of an effect and the conditions that enable those causes to produce it is purely pragmatic, and relative to human interests. This would perhaps be an acceptable consequence of Lewis' view if no more substantial basis for the distinction between causes and conditions seemed forthcoming. Menzies' alternative counterfactual theory of causation seems, however, to retain all the advantages of Lewis' theory while also providing a basis of this very sort. The central difference between Menzies and Lewis' respective definitions of causation is that, under Menzies' definition, the ordering of worlds against which the counterfactuals that determine the truth or falsity of a causal statement are assessed is not

[199] This last example is from Collins, Hall, and Paul (2004, p.40).

the same in all cases (or even in all "normal" cases). This is because, on Menzies' analysis, possible worlds are ordered not with respect to their similarity to the actual world, but instead with respect to their similarity to a *model* whose content is determined by the context in which a given causal statement is made. In addition to providing a more objective foundation for the distinction between causes and conditions, and thereby limiting, to some extent, the inordinate spread of causation that attends Lewis' theory, Menzies' model-relative analysis of causation is also well suited to enable dualists to account for mental causation of physical effects without having to deny either (2) or (3), for reasons that will shortly be made clear.

In the sense relevant to Menzies' theory of causation, a model can be understood as a trajectory in an *n*-dimensional state space, each of whose dimensions corresponds to a certain determinable property that admits of a range of determinate values. The properties that make up the dimensions of a given model's associated state space determine what kind of system the model is a model of. A model of a strict Newtonian system may thus consist of a series of points in a 5-dimensional state space, whose dimensions correspond to mass, velocity, and location along three mutually perpendicular spatial dimensions. In addition to the properties that determine the kind of system being modeled, a set of laws are also needed to specify how the components of the modeled system (represented by points in the system's associated state space) evolve over time. The trajectories of the points representing the components of a strict Newtonian system will thus, e.g., be determined by plugging their initial values for each dimension of the relevant state space into Newton's three laws of motion and law of universal gravitation.

The movement of these points through state space will then provide a model of the changes that the represented system undergoes through time.

With this background in place, we are now in a position to understand Menzies' analysis of causes as things that make a difference to an effect relative to certain model. The notion of model-relative difference making that lies at the heart of this analysis is defined by Menzies (2004, p.166) as follows:

> *C makes a difference to E* in an actual situation relative to the model *M* of the situation if and only if every most similar *C*-world generated by the model is an *E*-world and every most similar ~*C*-world generated by the model is a ~*E*-world.

Or, more simply put:

> *C makes a difference to E* in an actual situation relative to the model *M* if and only if $C \; \Box\!\!\rightarrow_M E$ and $\sim C \; \Box\!\!\rightarrow_M \sim E$.[200]

In determining the similarity of worlds relative to a certain model, greater weight should, Menzies claims, be given to the initial conditions described in the model (which specify the values that each component of the modeled system has for those properties that determine the kind of system being modeled) and the laws that govern their subsequent evolution than to the supposed absence of any interfering factors that would alter the model's trajectory from that prescribed by its governing laws. For any event *e*, the most similar *e*-worlds generated by a model *M* in which *e* does *not* occur will hence be those that contain a system of the same kind with the same initial conditions evolving in accordance with the same laws as the system modeled by *M*, but in which, due to some external intervention in the system's evolution, *e does* occur. Conversely, the most

[200] Here, again, only one of these conditionals will actually be counterfactual; which one will depend on whether or not *C* actually obtains.

similar ~*e*-worlds generated by a model *M* in which *e does* occur will be those that contain a system of the same kind with the same initial conditions evolving in accordance with the same laws as the system modeled by *M*, but in which, due to some external intervention in the system's evolution, *e* does *not* occur. The most similar *e*-worlds generated by a model *M* in which *e does* occur will, on the other hand, simply be those that contain a system that is perfectly modeled by *M* (and likewise for the most similar ~*e*-worlds generated by a model in which *e* does *not* occur).

On Menzies' view, then, when someone makes a statement to the effect that an actual event *c* caused another actual event *e*, the context in which the statement is made will dictate that the truth of the statement be judged relative to a certain model. While this model must be of a system of the same kind as one of[201] the systems to which *c* and *e* actually belong, it need not be of a system containing *c* or *e* as components. In order for the causal statement to be true, however, the modeled system must either contain both *c* and *e*, or neither *c* nor *e*. In the former case, the most similar ~*c*-worlds generated by the model must not, if the statement is true, include any *e*-worlds, and in the latter case, the most similar *c*-worlds generated by the model must include the actual world (in which *e* does occur) but no ~*e*-worlds. If these conditions are met, then the statement is true. Otherwise, it is false.

How, then, does Menzies' relativization of causation to models provide a more objective basis for the distinction between causes and conditions and reduce the surplus of causes generated by Lewis' theory? The answer to this question lies in the fact that

[201] Note that there may be many different kinds of systems to which an event or collection of events simultaneously belongs. This will prove important later on.

many of the events x_1, x_2, x_3,…x_n that stand in the ancestral of counterfactual dependence to a given effect *e* (and which, under Lewis' original proposal, thereby count as causing it) will not qualify as causes of *e* under Menzies' account, because the most similar x_1-worlds, x_2-worlds, x_3-worlds,…and x_n-worlds generated by the models relative to which statements about what caused *e* are best interpreted are not all worlds wherein *e* occurs. This identifies the events x_1, x_2, x_3,…x_n as mere conditions, rather than actual causes of *e*.

The following example may help illustrate this point (Menzies, 2004, pp.171-2). An indication of the seemingly overly permissive nature of Lewis' theory of causation can be found in the fact that whenever a person *S* develops lung cancer as a result of smoking, Lewis' theory requires us to count *S*'s having lungs as having also caused his/her contraction of lung cancer, since *S*'s having lungs and *S*'s smoking both stand in the ancestral of counterfactual dependence to *S*'s contraction of lung cancer. Menzies is able to avoid this counterintuitive result by suggesting that in most contexts, the statement ⌜ *S*'s smoking caused *S* to develop lung cancer ⌝ is best assessed relative to a model describing "a person living according to the laws of normal healthy functioning" (Menzies, 2004, p.171). Relative to such a model, *S*'s smoking makes a difference to *S*'s contraction of lung cancer, because in the most similar "non-smoking" worlds generated by *M, S* does not develop lung cancer, whereas in the most similar "smoking" worlds generated by *M, S does* develop lung cancer. In contrast, *S*'s having lungs does *not* make a difference to *S*'s contraction of lung cancer, because while in the most similar "lungless" worlds generated by *M, S* of course does *not* develop lung cancer, in the most similar "lunged" worlds generated by *M, S* also does not develop lung cancer (which is at

least partly due to the fact that in the most similar "lunged" worlds generated by *M*, *S* does not smoke).

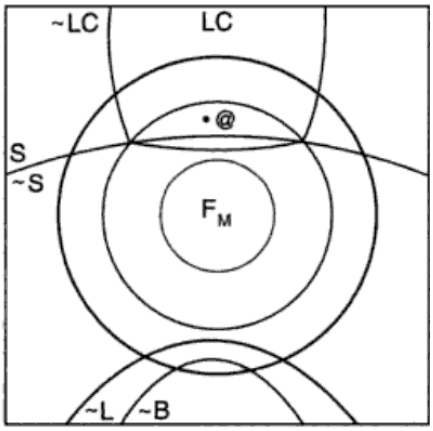


Figure 8. Diagrammatic representation of the causation of lung cancer (Menzies, 2004, p.171). F_M represents the field of worlds containing systems perfectly modeled by M. The surrounding concentric circles encompass worlds of varying proximity/similarity to F_M. @ represents the actual world. LC and ~LC encompass those worlds in which S does and does not develop lung cancer. S and ~S encompass those worlds in which S does and does not smoke. ~L and ~B encompass those worlds in which S has no lungs or is not born.

On Menzies' account, then, while *S*'s contraction of lung cancer depends counterfactually on both his/her smoking and his/her possession of lungs, only the former is eligible (in most contexts) to count as a cause of *S*'s cancer, for unlike *S*'s possession of lungs, *S*'s smoking also makes a difference to his/her contraction of cancer (relative to *M*). *S*'s having lungs is, on the other hand, a mere condition, rather than a cause of *S*'s cancer (again, relative to *M*), because while *S*'s contraction of cancer depends counterfactually on his/her possession of lungs, the latter makes no difference to the former (relative to *M*).

While Menzies' account thus seems to provide a more principled basis for the distinction between causes and conditions than Lewis', which treats conditions as merely pragmatically demoted causes, one might worry that the means by which Menzies achieves this result end up making causation itself too "subjective" or relative to human interests, inasmuch as the truth of any causal statement must, according to Menzies, be assessed relative to some model that is determined by the context in which the statement is made. This feature of Menzies' theory of causation may also lead one to wonder whether his solution to the undue proliferation of causation under Lewis' theory is not purely cosmetic, for given that the truth conditions for causal statements are, for Menzies, context sensitive, the distinction between causes and conditions will presumably be so as well. In which case, any events that in one context count as mere conditions for a certain effect *e* may in another context qualify as causes of *e*, thereby leaving the total number of causes of *e*, when counted up across all possible contexts, no different from what it would be under Lewis' theory.

In response to these worries, Menzies (2004, p.159) notes, first, that while context plays a role in determining what model should be used in evaluating a causal statement about some situation, once the kind of system being modeled and the laws governing it are fixed, it is then "a completely mind-independent matter whether some factor in the situation makes a difference to another." This follows from the fact that the ordering of worlds that specifies what factors make a difference to an effect is fully determined by the laws and initial conditions of the model that generates said ordering. Second, and perhaps more importantly, while Menzies (2004, p.159) suggests, quite reasonably, that "a plausible metaphysics is likely to allow that any particular spatiotemporal region

instantiates several kinds of systems," this does not mean that "any causal model of a situation is as good as any other, or more specifically, [that] any kind of system is just as natural as any other for determining causal relations." In particular, the models available for the proper assessment of any given causal statement should presumably be restricted to those that are of a *natural kind* of system that is of the same kind as some system that is actually instantiated in the relevant situation, meaning that the determinable properties corresponding to the dimensions of the model's state space must constitute natural kinds, the evolution of the model must be governed by actual laws of nature, and the events about which the causal statement is made must be components of a system that is of the same kind and governed by the same laws as the system that the model represents. Given these constraints on the range of models that are available for use in evaluating any causal statement, it would seem likely that there will be certain events that stand in the ancestral of counterfactual dependence to a given effect, but which nevertheless fail to make a difference to that effect (and which hence fail to qualify as anything more than mere conditions of it) under any of the appropriate models. If this is so, then the total number of causes (across all available models) will be fewer under Menzies' account than under Lewis' after all.[202]

Having seen some of the advantages that Menzies' counterfactual theory of causation has over that of Lewis, it remains to see what advantages it offers to dualists seeking an analysis of causation that will enable them to substantiate the mind's ability to

[202] The set of events that count as causes as opposed to mere conditions is further restricted by Menzies' additional requirement (discussed below) that causes must also be connected to their effects by a process that is "picked out" by the model-relative counterfactual dependence of the latter on the former.

cause physical effects. As previously noted, counterfactual theories of causation hold many attractions for those interested in finding a solution to the Exclusion Problem that is consistent with both (1), the Causal Efficacy of the Mental and (4*), Mind-Body Dualism. Foremost among these is the fact that the same event can depend counterfactually on a wide variety of other events. Hence, if causation is just some form of counterfactual dependence, there is no apparent reason why a physical event could not have both physical and non-physical, mental causes, since its occurrence might easily turn out to depend counterfactually on events of both types. This, however, is an advantage that dualists might derive from pretty much any counterfactual theory of causation. What sets Menzies' proposal apart from other such theories is that it has the added benefit of dispelling any suspicion that the causal efficacy of mental events is merely secondary to or derived from their physical realizers by allowing that in some contexts, it may in fact be *false* to attribute a certain physical effect to anything other than a non-physical, mental cause, (even though that same effect may, in other contexts, be truly ascribed to physical causes).

This further feature of Menzies' theory is a consequence of the aforementioned fact that the same situation can instantiate many different kinds of systems. This means that for any given situation, there may very well be a range of different kinds of models (corresponding to the different kinds of systems the situation instantiates) relative to which causal statements made about that situation can be properly assessed. Which of these models is relevant to the assessment of any particular causal statement will depend on the context in which the statement is made. Given, then, that certain events involved in the situation may qualify as causes of another event relative to some of these models, but

not others, it follows that an event may in one context be truly said to cause another, whereas in another context it cannot.

The types of situations that are of interest to us are of course those involving some sort of mind-body interaction. In such cases, we may suppose the relevant situation to instantiate at least two different kinds of systems, governed by distinct sets of laws: a neurophysiological system (modeled by vectors in a state space whose dimensions correspond to determinable neurophysiological properties and whose evolution is governed by various neurophysiological laws), and a psychological system (modeled by vectors in a state space whose dimensions correspond to determinable psychological and behavioral properties and whose evolution is governed by various psychological laws). Judged in terms of Menzies' theory, we should expect that statements about such situations will in different contexts be best assessed relative to different types of models. In contexts where a psychological explanation for some bodily motion is sought (e.g., when asking a person's reasons for acting as s/he did), causal statements about that motion should be evaluated relative to some psychological model, whereas in contexts where a neurophysiological explanation for the motion is sought (as, e.g., in the case of a neurologist seeking the cause of a chronic muscle spasm), causal statements about that motion should be evaluated relative to some neurophysiological model. This opens up the possibility that in contexts of the former sort, the only events that make a difference to the relevant bodily motion are non-physical, mental events, whereas in contexts of the latter sort, the primary difference-makers are physical in nature. Should this be the case, then it will in certain contexts (viz. those wherein causal statements about some situation

are best assessed relative to a psychological model) be *false*, under Menzies' theory, to attribute certain physical effects to anything *but* non-physical, mental causes.

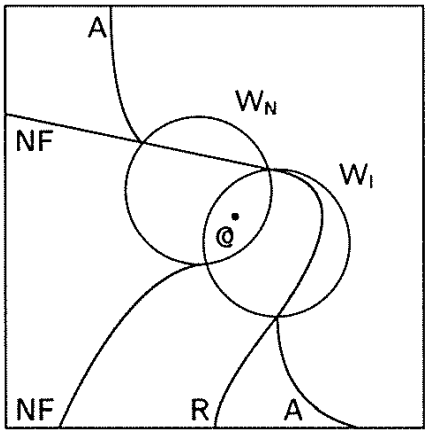


Figure 9. Diagrammatic representation of the causation of intentional action (Menzies, 2003, p.209). W_N and W_I represent the sets of most similar worlds generated, respectively, by a certain neurophysiological model N and an intentional/psychological model I. @ represents the actual world, which as the diagram indicates, contains a situation instantiating systems of both kinds. A encompasses those worlds wherein a certain action a (e.g. the raising of an arm) is performed by an agent S. NF encompasses those worlds wherein a certain pattern of neural firing occurs in the S's brain. R encompasses those worlds wherein the agent has certain reasons for performing a. Note that whereas S's reasons make a difference to A relative to I, the neural firing in S's brain does not. As will shortly be made clear, this is due to the fact that S's reasons are multiply realizable at the neurophysiological level.

The primary reason to think this is indeed the case (i.e., that judged relative to the appropriate psychological models, there are some bodily motions to which mental events alone make a difference) is that mental difference-makers are multiply realizable (Menzies, 2003, p.221). Since psychological models are specified solely in terms of psychological/behavioral properties, and are indifferent to the neurophysiological realizers of the systems they represent, the most similar worlds generated by any such

model will include a variety of worlds, each of which contains a different neurophysiological realization of the same psychological system. Thus, if a certain psychological model *M* represents a system wherein a given mental event *m* gives rise to a certain behavioral effect *e*, the most similar *m*-worlds generated by that model will include a variety of worlds containing psychological systems that are perfectly modeled by *M*, and which differ from one another only with respect to the neurophysiological realization of *m*. In one of these worlds, e.g., *m* might hence be realized by a certain neurophysiological event p_1, while in another, *m* is realized by a different type of neurophysiological event p_2, and in another by p_3, etc. Yet all of these worlds will still be contained in the set of most similar *m*-worlds generated by *M* inasmuch as they all contain different neurophysiological realizations of the same psychological system that *M* represents.

Assuming, then, that *m* makes a difference to *e* relative to *M* (i.e., $m \; \Box\!\!\rightarrow_M e$ and $\sim m \; \Box\!\!\rightarrow_M \sim e$), it follows that the same cannot be true of *m*'s actual neurophysiological realizer $p_@$ (which is typically *m*'s main physical "competitor" for the role of *e*'s cause). For given that the occurrence of *m* is included in the very content of *M*, and *e* occurs in every most similar *m*-world generated by *M*, not all the most similar $\sim p_@$-worlds generated by *M* will be $\sim e$ worlds (i.e. $\sim(\sim p_@ \; \Box\!\!\rightarrow_M \sim e)$). This is because the most similar $\sim p_@$-worlds generated by *M* will all be worlds wherein *m* still occurs, but is instead realized by some other neurophysiological event besides $p_@$, and since *m*'s occurrence is sufficient for *e* in all of the most similar worlds generated by *M*, all of these worlds will be worlds wherein *e* occurs as well. The multiple realizability of mental properties thus ensures that, under Menzies' account, there will likely be certain contexts wherein the

only things that can be truly described as causes of a given physical effect are mental in nature.[203]

By allowing mental events to act as causes in contexts where their realizers cannot, Menzies' theory of causation helps allay the worry that the only causal efficacy that mental events "possess" in truth belongs entirely to the physical states on which they depend. This feature of Menzies' account does not, however, render it inconsistent with (2) the Causal Self-Sufficiency of the Physical, for it may be suggested that for any situation containing a physical effect *e* that is in some contexts attributable only to a non-physical, mental cause, there must also be some appropriate *physical* model *P* such that (a) there are other contexts in which causal statements about *e* would be best assessed relative to *P*, and (b) some physical event qualifies as causing *e* relative to *P*. Put in terms of the previous example, while $p_{@}$ may fail to make a difference to *e* relative to *M*, (since $\sim(\sim p_{@} \,\Box\!\!\rightarrow_{M} \sim e)$), there may nonetheless be certain contexts wherein causal statements about *e* are most naturally evaluated in relation to a certain physical (or

[203] Note that if, as proposed in Chapter 7, the psychophysical laws that render the physical realizers of a mental state sufficient for that state are metaphysically necessary, then while the multiple realizability of mental states enables them to make a difference to physical effects (relative to some appropriate psychological model) without their realizers' also doing so, the actual neurophysiological realizer of a given mental state may in contrast be incapable of making a difference to a physical effect (relative to some appropriate neurophysiological model) unless the mental event it realizes does so as well. This is because for any neurophysiological model *M* of an actual system containing a certain neurophysiological realizer *p* of a given mental state *m*, if the psychophysical law linking *p* to *m* is metaphysically necessary (so that every *p*-world is an *m*-world), and *p* makes a difference to a physical effect *e* relative to *M*, it follows, first, that the most similar *m*-worlds generated by *M* will also be *e*-worlds. For those worlds containing systems that are perfectly modeled by *M* will all contain both *p* and *e*, and given that every *p*-world is an *m*-world, it follows that these worlds will all contain *m* as well. Second, the most similar ~*m*-worlds generated by *M* will likewise be ~*e*-worlds, for the nearest ~*m*-worlds (relative to *M*) will simply be the closest worlds wherein *p* fails to occur, and given that *p* makes a difference to *e* relative to *M*, it follows that these worlds will all be ~*e*-worlds. Thus, assuming the metaphysical necessity of psychophysical laws, if *p* makes a difference to *e* relative to *M*, then *m* must do so as well. (Note that this also means that, relative to *M*, *m* and *p* will end up making a difference to one another. This, however, doesn't entail that they likewise *cause* one another, for as will be discussed further below, making a difference is, according to Menzies, necessary but not sufficient for being a cause.)

neurophysiological) model *P*, relative to which $p_{@}$ *does* make a difference to *e* (i.e. $p_{@} \; \Box\!\!\rightarrow_{P} e$ and $\sim p_{@} \; \Box\!\!\rightarrow_{P} \sim e$). As there is nothing in Menzies' account that prohibits this from holding quite generally, it appears that Menzies' theory of causation gives dualists the ability to accept (2) while also maintaining that there are some physical effects that only non-physical, mental events can, in certain circumstances, be truly said to cause.

A further benefit that dualists stand to gain by applying Menzies' theory to instances of mental causation is that it also corroborates the main thesis of Chapter 7: that (3) the Absence of Systematic Overdetermination is either implausible or else consistent with the conjunction of interactionist dualism and (2). Which of these is the case depends on how overdetermination is defined within the framework of Menzies' account, and in particular on whether overdetermining causes of an effect *e* must cause *e* relative to the same kind of model, or whether *e* can also be overdetermined by causes which only qualify as causing it relative to models of different kinds of systems. In the latter case, (3) is likely to be false, for as noted above, a single situation will often instantiate a variety of different kinds of systems, each of which is represented by a different kind of model, and relative to these different kinds of models, the same event will often have many different causes. Thus, elaborating on our previous example, a situation involving an instance of psychophysical causation will presumably instantiate not only certain psychological and neurophysiological systems, but also a range of different biological, chemical, and quantum systems. Relative to models of these different kinds of systems, the same event *e* may consequently turn out to have not only distinct mental and neurophysiological causes, but also various biological, chemical, and quantum causes as well. While context will determine which of these causes can be truly described as *the* cause of *e* on any

particular occasion by specifying the kind of model relative to which causal statements about *e* are then most naturally assessed, if different causes of an effect needn't qualify as causing it relative to the *same* kind of model in order for it to be overdetermined by them, then *e* would seem to be massively overdetermined by causes at a variety of different mental and physical levels. Nor will such overdetermination be restricted to events like *e* that have mental causes, for those physical effects that have no mental cause relative to any available model will still typically have a variety of distinct quantum, chemical, and other non-mental causes relative to certain quantum, chemical, and other non-psychological models.

It appears, then, that unless one requires overdetermined effects to have distinct causes relative to the *same* kind of model, nearly all physical effects will be overdetermined, and (3) will consequently be false. If, however, one accepts this added constraint on instances of overdetermination, then the conjunction of (2) and interactionist dualism is perfectly consistent with the claim that overdetermination is rare, for assuming that mental properties are multiply realizable, it follows (as shown above) that under Menzies' account, mental events and their physical realizers cause the same effects only relative to different kinds of models. Hence, if an effect must be produced by distinct causes relative to the *same* kind of model in order for it to be overdetermined by those causes, then mental causes and their physical realizers do not overdetermine their joint effects (Menzies, 2003, p.220-1). With this constraint in place, standard instances of mental causation will hence not qualify as instances of overdetermination, even if every physical effect produced by a mental cause is also caused by the latter's physical realizer. As promised, Menzies' theory of causation thus enables dualists to make the case that (3)

is either implausible (if overdetermining causes do not have to cause the effects they overdetermine relative to the same kind of model) or else consistent with the conjunction of interactionist dualism and (2) (if they do).[204]

Putting all this together, Menzies' analysis of causation in terms of model-relative difference making can be seen to offer dualists a way of accounting for the causation of physical effects by non-physical, mental causes in a manner that is consistent with (2), while also resolving the Exclusion Problem by showing that, depending on how overdetermination is defined within the framework of his account, (3) is either implausible or else consistent with the conjunction of (1), (2), and (4*). For those sympathetic to mind-body dualism, this makes for quite an attractive package. One issue, however, remains to be dealt with, for while Menzies' counterfactual theory of causation has been shown capable of preventing the over-proliferation of causes that such theories threaten to give rise to, there is another well-known objection to counterfactual theories of causation that has yet to be addressed. This objection concerns the ability of such theories to handle instances of so-called causal "preemption." In contrast to the problem of over-proliferation, which challenges the *sufficiency* of counterfactual dependence for causation by suggesting that any attempted analysis of causation solely in terms of counterfactual dependence will end up generating too many causes, instances of causal preemption are often taken to show that counterfactual dependence is *unnecessary* for

[204] When *does* an effect qualify as overdetermined under Menzies' account? We might say that an effect e is overdetermined by causes c_1 and c_2 iff (a) it is possible for c_1 to occur without c_2, and *vice versa*, (b) c_1 makes a difference to e relative to some model M_1 that abstracts away from the occurrence of c_2, and (c) c_2 makes a difference to e relative to some model M_2 that abstracts away from the occurrence of c_1. The plausibility of (3) will then depend largely on whether or not M_1 and M_2 must also be models of the same kind of system.

causation, because there are certain effects that do not seem to depend counterfactually on their apparent causes. While causal preemption comes in a variety of different forms[205], all instances of causal preemption involve two or more distinct events that tend towards an effect, where only one of these events actually produces the effect, thereby relegating the other event(s) to the status of "backup" or "would-be" cause(s). A common example of such preemption is that of two rocks being thrown at a bottle, one of which, *R*1, strikes the bottle and breaks it, while the other, *R*2, would have done the same if *R*1 hadn't already broken the bottle by the time it arrived at the place where the bottle used to be. The difficulty that such cases pose for counterfactual theories of causation is that we would typically identify the event that produced the effect as having caused it, even though the latter does not counterfactually depend on the former, since had the former failed to occur, one of the "backup" causes would have been sufficient to produce the effect anyways. Thus, in the example just given, while we would presumably identify the throwing of *R*1 as the cause of the bottle's breaking, the latter doesn't counterfactually depend on the former, since if *R*1 hadn't been thrown, the bottle still would've broken on account of its being struck by *R*2.

---

[205] Lewis' (1973a) original analysis of causation is only capable of handling instances of so-called "early" preemption. His subsequent (2000) proposal attempts to resolve the difficulties raised by remaining forms of preemption (e.g. "late" preemption, "trumping," and "preemptive prevention") by defining causation as the ancestral of "influence," where one event *C* influences another event E "iff there is a substantial range C_1, C_2,… of different not-too-distant alterations of *C* (including that actual alteration of *C*) and there is a range E_1, E_2,… of alterations of *E*, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on" (p.190). Problems with Lewis' revised theory have, however, been pointed out by Collins (2000, pp.230-1), Kvart (2001), Schaffer (2001), Strevens (2003), and Bigaj (2012), who provide a number of counterexamples showing (a) that it is possible, and indeed quite common for one thing to cause another without influencing it, and (b) that in some cases of preemption, the preempted cause actually exerts *more* influence over the effect than its actual cause, and is thus incorrectly identified as more of a cause of the effect than the actual cause under Lewis' revised account.

While Menzies' treatment of causes as model-relative difference makers provides the basis for an answer to this problem, its full solution requires us to concede that "[t]he notion of counterfactual dependence does not constitute the whole of the concept of causation…[Rather,] the existence of a counterfactual dependence is merely the surface marker of a deeper phenomenon that really counts as causation" (Menzies, 2003, p.209). More specifically, Menzies suggests that to accommodate instances of causal preemption, we should view the model-relative relations of counterfactual dependence involved in his notion of difference making as helping merely to define the *functional role* of causation, whereas causation *itself* is to be identified with the process that occupies that role, similarly to the way in which one might treat the various superficial (inessential) properties of water as merely specifying the functional role that picks out what water *itself* really is, viz. H_2O (Menzies, 2003, p.214). While counterfactual dependence thus continues to play a central role in Menzies' theory of causation, inasmuch as it figures among the characteristics that pick out the processes in which causation itself consists, to deal with instances of causal preemption, one must, Menzies' claims, allow that causes are more than *just* things on which their effects counterfactually depend. Given, however, that counterfactual dependence is still needed to identify both what causation really is, and which events qualify as causes, Menzies' theory of causation is still, I think, best viewed as a certain kind of counterfactual account.

How, then, does this distinction between causation and the functional role it occupies enable us to handle instances of causal preemption? The answer can be given in two parts. First, it is suggested that in cases of causal preemption, the relations of counterfactual dependence that are relevant to the allotment of causal responsibility can

only be brought out by assessing such situations relative to models that abstract away from preempted "backup" causes (Menzies, 2003, p. 211-2). This abstraction may be justified by the fact that in contexts wherein one is seeking an explanation for why an effect *actually* occurred, the presence of such "backup" causes has no bearing on the main point of interest, which is the series of events connecting the pertinent effect *e* to its contextually most salient cause *c*. Once all preempted causes have been removed from the picture, though, *c will* end up making a difference to *e*, and the latter hence *will* end up counterfactually depending upon the former, for without the presence of any "backup" causes to ensure that *e* still occurs in *c*'s absence, *e* would not have occurred without *c*. In other words, given that an effect's "backup" causes (if it has any) are of no interest to those who wish to understand why that effect *actually* occurred, the contexts in which causal statements are made will always call for such statements to be assessed relative to models that abstract away from any such causes. Relative to such models, though, effects *do* counterfactually depend on their causes, even when the relevant cause and effect are constituents of an instance of causal preemption, for even effects with "backup" causes counterfactually depend on their *actual* causes when assessed relative to models from which all "backup" causes have been removed. Thus, in the rock-throwing case, any model relative to which *R*1 is identified as the cause of the bottle's breaking will omit the throwing of *R*2 from its initial conditions, since the fact that *R*2 was thrown is irrelevant to why the bottle actually broke. Relative to any such model, though, the bottle's breaking depends counterfactually on *R*1, since in the most similar ~*R*1-worlds generated by these models, there is no *R*2 waiting in the wings to break the bottle in *R*1's place, and the bottle will hence remain unbroken. Menzies' model-relative treatment of causation

thus seems to enable him to maintain that, relative to those models wherein they qualify as having an actual cause, effects always counterfactually depend on their causes, even in cases of causal preemption.

This, however, only gives rise to another question, viz. "how does the fact that this counterfactual dependence exists in a *hypothetical* situation that abstracts away from the presence of [any preempted causes] help to identify the causal relations in the *actual* situation in which [such causes] are very much present?" (Menzies, 2003, p. 212). Menzies answers this question by proposing that in cases of causal preemption, the counterfactual dependence of the pertinent effect *e* on its cause *c* as defined relative to a model wherein all preempted "backup" causes have been removed picks out a certain process (i.e. "a temporally ordered sequence of events") connecting *c* to *e*, which process also connects *c* to *e* in the *actual* situation wherein all the preempted "backup" causes do occur. "The existence of this process" in the actual situation can then be seen as giving us "a very good reason for thinking that the two events in question are causally related in the actual circumstances" (Menzies, 2003, p. 212). The reason being that if *c* makes a difference to *e* in a certain hypothetical situation by virtue of being connected to *e* in that situation by a certain series of events, and that same series of events connects *c* to *e* in an actual situation, then insofar as *c*'s connection to *e* in the hypothetical situation would give us reason in that situation to identify *c* as a cause of *e*, the fact that *c* and *e* are actually linked by the exact same chain of events as they are in the hypothetical situation should give us reason to say that *c* causes *e* in the actual situation as well, *even if* in the actual situation, *c* does *not* make a difference to *e*, due to the presence of various preempted "backup" causes.

Incorporating this idea into the framework outlined above, Menzies' theory of causation can thus be made capable of dealing with instances of causal preemption by supplementing his analysis of causes as difference makers with the further requirement that causes must also be actually linked to their effects by a process that is "picked out" by a model relative to which they make a difference to them. The notion of a model's "picking out a process" is defined by Menzies as follows:

> The counterfactual dependence of *E* on *C* relative to the model *X picks out a process* (a temporally ordered sequence of events) if and only if the process is present in all the most similar *C*-worlds generated by the model that are *E*-worlds and is absent in all the most similar ~*C*-worlds generated by the model that are ~*E*-worlds. (Menzies, 2003, p.212)

This captures the sense in which the difference that one event makes to another (relative to a certain model) depends on a certain process connecting the two, inasmuch as that process is present in all the closest worlds wherein the one brings about the other, but absent in all the closest worlds wherein both fail to occur. With this notion in place, a complete analysis of causation can now be given that is capable of handling cases of causal preemption:

> *C* is *a cause* of the distinct state *E* relative to the model *X* of an actual situation if and only if
>
> (1) *E* counterfactually depends on *C* relative to the model;
> (2) this counterfactual dependence picks out a process;
> (3) this process connects *C* and *E* in the actual situation. (Menzies, 2003, p.212)

According to Menzies' final definition of causation, being a difference maker is thus necessary but not sufficient for being a cause. In addition to making a difference to an effect relative to an appropriate contextually determined model, causes must also be actually linked to their effects by a process that is "picked out" by that same model. The

causation of any one event by another can then be equated with the actual process, whatever it is, that connects them in this way, while the relations of counterfactual dependence that pick out this process pertain to the functional role that the causal relation between the two events occupies, and through which it is identified.

Far from undermining its suitability for dualists seeking a solution to the Exclusion Problem, these further additions to Menzies' analysis of causation instead open up two different ways in which dualists might use it to provide an account of mental causation that meets their needs. These two options correspond to two different ways in which the processes picked out by the model relative counterfactual dependence of physical events on mental events might be construed. One option is to hypothesize that these processes consist in transmissions of energy from mental to physical states of the sort described in section 1 above. Pursuing this option would enable one to unify the two strategies developed in the present chapter by suggesting that mental causation consists in a process of energy transmission between mental and physical events, which is picked out, in Menzies' sense, by the model-relative counterfactual dependence of the latter on the former.

Combining the Transference-based account of mental causation developed in the previous section with Menzies' theory in this way helps address at least one of the potential problems with simple Transference accounts, which is that such accounts are unable to allow for cases of causation by omission. Like most counterfactual theories of causation, Menzies' theory is well equipped to handle such cases, for there is no reason why omissions cannot make a difference to certain effects relative to certain models (since there is nothing to prevent an effect's occurrence from counterfactually depending

on the *failure* of some other event to occur), nor is there any reason why the model-relative counterfactual dependence of an effect on an omission could not also pick out a temporally ordered sequence of events running from the omission to the effect, which process, according to Menzies' analysis, thereby constitutes the actual causation of the latter by the former.[206] Subsuming the previously developed Transference account of mental causation under Menzies' model would consequently enable dualists to maintain that while standard instances of mental causation consist in the transference of energy from mental causes to their physical effects (since this happens to be the kind of process that is typically picked out by the model-relative counterfactual dependence of the latter on the former), not *all* instances of causation consist in such exchanges of energy, for in some cases (e.g. those wherein an effect is caused by an omission), the process picked out by the model-relative counterfactual dependence of the effect on its cause may not involve any transference of energy between the two. This highlights one of the more interesting advantages of Menzies' distinction between the functional role of causation and causation itself, which is that since the kind of process picked out by the model-relative counterfactual dependence of an effect on its cause may differ from case to case, one can remain non-committal about the precise nature of the process with which causation is to be identified in any particular instance. It may consequently be that in some cases, causation turns out to consist in an exchange of energy between cause and effect, while in other cases it doesn't, for what matters is that the relevant process

[206] The ability of Menzies' model-relative analysis of causation to prevent the over-proliferation of causes that besets other counterfactual theories also helps to dispel Beebee's (2004) worry that if omissions could be causes, then there would be far more instances of causation by omission than common sense suggests.

occupies the right kind of functional role, and processes that involve energy transfers needn't be the only ones capable of doing this.

Inasmuch as this hybrid account of mental causation requires one to accept that there are changes in the energy levels of certain physical entities (viz. those with non-physical, mental causes) that cannot be accounted for in strictly physical terms, dualists who adopt it must deny (2), the Causal Self-Sufficiency of the Physical. The resistance that many have towards doing this should be weakened to some extent by the observation, defended in the previous chapter, that (2) is neither a well-established law of any science, nor does it follow or derive clear inductive support from any such law (including the laws of conservation of energy and momentum). Further grounds for rejecting (2) can also be obtained by integrating the hybrid Transference/Counterfactual theory of mental causation currently under consideration with Nancy Cartwright's "patchwork" theory of laws (also discussed in the previous chapter). This would enable one to suggest, following Cartwright, that since the only events that we have reason to believe can be explained in terms of the laws of a given science are those that are components of a situation for which that science can provide an adequate model, the fact that we currently have no adequate physical models of situations wherein a person's beliefs, desires, or sensations lead them to behave in certain ways means that we have no reason to believe that such behavior can be fully explained in strictly physical terms. Tying this argument into Menzies' model-relative analysis of causation, one can then maintain that if there are indeed certain kinds of behavior that can only be adequately represented by way of certain psychological models, relative to which the physical events involved in such behavior have only non-physical, mental causes, it follows that the

physical effects of those mental causes have no sufficient physical causes, and hence that (2) is false. Here, then, is one account of mental causation that dualists might give, which resolves the Exclusion Problem by rejecting (2).

A second, less adventurous way for dualists to implement Menzies' theory would be to treat the processes picked out by the model-relative counterfactual dependence of physical effects on mental causes as mere sequences of mental and behavioral events between which no energy is exchanged. Thus, the process connecting a desire for mangosteens with certain mangosteen-seeking behavior might be said to consist of a series of judgments about the respective merits of the various ways of obtaining mangosteens, followed by the formation of a belief that one of these ways is currently the best available option, which is in turn followed by certain bodily movements that are initiated with the intent of carrying out that course of action that has been deemed the best. Likewise, the process connecting a sensation of pain with a certain aversive response might be said to consist in the occurrence of various feelings of anxiety, surprise, or fear, followed by the formation of a desire that the pain should cease, which is in turn immediately followed by certain behavioral expressions of discomfort (e.g. wincing or groaning) and bodily movements aimed at removing the source of pain.[207]

Since physical events need never undergo any changes in energy that cannot be accounted for in strictly physical terms if the processes connecting mental causes with their effects are conceived in this way, dualists who opt for this alternative

[207] The initiation of these bodily movements may in some cases be preceded by the formation of a belief that these motions would be the best way to alleviate the sensation of pain, but this kind of situation is likely atypical.

implementation of Menzies' theory do not have to deny (2), but may instead openly accept it, by allowing that for any physical event *e* that has a mental cause *m* relative to some psychological model *M*, there is also some available physical model *P* relative to which that same event is assigned a physical cause. The distinct nature of the causes and causal processes picked out by *M* and *P* will be ensured by the fact that *m* and the process linking it to *e* (relative to *M*) are both multiply realizable at the neurophysiological and other more basic physical levels (Menzies, 2003, p.217-8). Thus, while the physical cause *p* that *P* assigns to *e* and the process it picks out between the two may realize *m* and the process that *M* picks out between *m* and *e* in the actual situation surrounding *e*'s occurrence, the two causes and processes (viz. *p* and *m*, and the processes connecting *p* to *e* and *m* to *e*) cannot be identified, for just as there are other physical events that could've realized *m* besides *p* (which accounts for why the physical cause that *P* assigns to *e* will *not* make a difference to *e* relative to *M*), there are likewise other physical processes that could have realized the process linking *m* to *e* besides the one *P* picks out between *p* and *e* (which accounts for why the process picked out by *P* is *not* also picked out by *M*). Having thus blocked any attempt to identify mental causes and the processes linking them to their effects with the physical events and processes that realize them, dualists can then make use of the arguments offered above to suggest that, depending on whether or not overdetermining causes must produce the effects they overdetermine relative to models of the same kind of system, (3) the Absence of Systematic Overdetermination is either implausible or else consistent with interactionist dualism and (2). By treating the processes picked out by the model-relative counterfactual dependence of physical effects on mental causes as mere sequences of mental and behavioral events between which no

energy is exchanged, dualists can thus derive an alternative account of mental causation from Menzies' theory that resolves the Exclusion Problem by showing that if overdetermination is defined in such a way that (3) is not implausible, then the apparent incompatibility of (1), (2), (3), and (4*) is illusory.

3. Conclusion

Having now identified two different theories of causation according to which non-physical mental events can coherently be said to cause physical effects, the lingering worry that such causal interaction between physical and non-physical entities is incomprehensible can be dismissed as groundless. As shown above, these two different theories of causation, one of which equates causation with the transmission of conserved quantities, and the other of which identifies it with processes picked out by model-relative relations of counterfactual dependence, yield at least three potential accounts of mental causation that dualists can use to respond to the Exclusion Problem either by rejecting (2) or by maintaining that (3) is either implausible or else consistent with (1), (2), and (4*). On the one hand, dualists can maintain that mental states transmit energy to physical states, and that in doing so they thereby qualify as causing the latter, either because (a) for one thing to cause another just is for some conserved quantity (e.g. energy) to be transmitted from one to the other, or because (b) for one thing to cause another is for it to be linked to its effect by some process that is picked out by the model-relative counterfactual dependence of the latter on the former, and processes of energy transference from mental to physical states happen to be picked out by the counterfactual

dependence, relative to certain psychological models, of physical states on those mental states from which they receive energy. Dualists who take either of these options can resolve the Exclusion Problem by denying (2), on the grounds that the amount of energy possessed by physical states that are affected by the mind may change in ways that cannot be accounted for in terms of any compensatory change in the energy possessed by other physical states.

On the other hand, dualists can also maintain that relative to certain psychological models, physical states often counterfactually depend on mental states that they are connected to by way of sequences of mental and behavioral events between which no energy is exchanged. This fact may then be said to qualify such physical states as being caused by the mental states on which they counterfactually depend, because the sequences of events connecting the two types of states are picked out by the model-relative relations of counterfactual dependence between them, and for one thing to cause another just is for it to be connected to its effect by some process that is picked out in this way. Dualists who take this option can then resolve the Exclusion Problem without rejecting (2), by instead insisting that (3) is either implausible (if overdetermining causes of an effect needn't produce that effect relative to the same kind of model), or else compatible with (1), (2), and (4*) (if they do).

With these three accounts of mental causation at their disposal, dualists seem justified in affirming that mind-body dualism cannot be reasonably rejected on the grounds that it renders the mental causation of physical effects incomprehensible or inconsistent with (2) and (3). For as was shown in Chapter 8, (2) is not a well-established principle of any current science, nor does it follow or derive inductive support from any

such principle, and as argued in Chapter 7 and again in the present chapter, depending on how causal overdetermination is defined, (3) is likely either false or else consistent with interactionist dualism and (2). Given, then, that (2) and (3) are far less incontrovertible than is often supposed, dualists are free to construct theories of mental causation (such as those described above) that reject (2) and/or entail that (3) is either implausible or compatible with (1), (2), and (4*).

The challenge raised by the Exclusion Problem having thus been answered, the burden now falls upon the physicalist to provide some alternative reasons for denying the existence of non-physical, mental properties and events. This task will be made more difficult by the fact that if these reasons are to justify the rejection of dualism, they will have to be sufficiently compelling to outweigh the arguments offered in Chapters 4, 5, and 6, wherein the multiple realizability of mental properties and the phenomenal and intentional features exhibited by their instances was shown to provide substantial support for the view that mental properties and events are not identical with, reducible to, or fully explainable in terms of physical properties and events. As long as these arguments in favor of dualism remain unmatched by any equally forceful reasons for rejecting it, and no more convincing grounds can be given for thinking that dualism makes mental causation incomprehensible or inconsistent with some indisputable scientific or metaphysical principle, the position should remain, at the very least, on the list of viable options for those seeking to develop a satisfactory theory of mind.

BIBLIOGRAPHY

Abramov, Israel, and James Gordon. 1994. “Color Appearance: On Seeing Red-or Yellow, or Green, or Blue.” *Annual Review of Psychology* 45: 451–85.

Adams, Frederick, and Kenneth Aizawa. 1994. “‘X’ Means X: Fodor/Warfield Semantics.” *Minds and Machines* 4: 215–31.

Aizawa, Kenneth. 2009. “Neuroscience and Multiple Realization: A Reply to Bechtel and Mundale.” *Synthese* 167: 493–510.

———. 2013. “Multiple Realization by Compensatory Differences.” *European Journal for Philosophy of Science* 3: 69–86.

Aizawa, Kenneth, and Carl Gillett. 2009a. “Levels, Individual Variation and Massive Multiple Realization in Neurobiology.” In *The Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 539–82. Oxford University Press.

———. 2009b. “The (Multiple) Realization of Psychological and Other Properties in the Sciences.” *Mind and Language* 24: 181–208.

———. 2011. “The Autonomy of Psychology in the Age of Neuroscience.” In *Causality in the Sciences*, eds. Phyllis McKay Illari Federica Russo, 202–23. Oxford University Press.

Antony, Louise. 2003. “Who’s Afraid of Disjunctive Properties?” *Philosophical Issues* 13: 1–21.

Antony, Louise, and Joseph Levine. 1997. “Reduction with Autonomy.” *Philosophical Perspectives* 11: 83–105.

Armstrong, David. 1968. *A Materialist Theory of the Mind*. Routledge.

Audi, Paul. 2013. “How to Rule Out Disjunctive Properties.” *Noûs* 47: 748–66.

Averill, Edward, and Bernard Keating. 1981. “Does Interactionism Violate a Law of Classical Physics?” *Mind* 90: 102–7.

Aydede, Murat. 2000. “An Analysis of Pleasure Vis-a-Vis Pain.” *Philosophy and Phenomenological Research* 61: 537–70.

Baker, Lynne Rudder. 1993. “Metaphysics and Mental Causation.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

———. 1995. *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge University Press.

Balog, Katalin. 1999. “Conceivability, Possibility, and the Mind-Body Problem.” *Philosophical Review* 108: 497–528.

Battersby, Stephen. 2013. “Chasing Shadows.” *New Scientist*, May 11.

Bechtel, William, and Jennifer Mundale. 1999. “Multiple Realizability Revisited: Linking Cognitive and Neural States.” *Philosophy of Science* 66: 175–207.

Beebee, Helen. 2004. “Causing and Nothingness.” In *Causation and Counterfactuals*, eds. John Collins, Ned Hall, and L. A. Paul, 291–308. MIT Press.

Bennett, Karen. 2003. “Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It.” *Noûs* 37: 471–97.

———. 2008. “Exclusion Again.” In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, eds. Jakob Hohwy and Jesper Kallestrup. Oxford University Press.

Berkeley, George. 1710/1977. *A Treatise Concerning the Principles of Human Knowledge*. Bobbs-Merrill.

Bickle, John. 1998. *Psychoneural Reduction: The New Wave*. MIT Press.

Bigaj, Tomasz. 2012. “Causation Without Influence.” *Erkenntnis* 76: 1–22.

Bird, Alexander. 2005. “Laws and Essences.” *Ratio* 18: 437–61.

———. 2007. *Nature’s Metaphysics*. Oxford University Press.

Black, Robert. 2000. “Against Quidditism.” *Australasian Journal of Philosophy* 78: 87–104.

Block, Ned. 1978. “Troubles with Functionalism.” *Minnesota Studies in the Philosophy of Science* 9: 261–325.

———. 1990. “Inverted Earth.” *Philosophical Perspectives* 4: 53–79.

———. 1996. “Mental Paint and Mental Latex.” *Philosophical Issues* 7: 19–49.

———. 1997. “Anti-Reductionism Slaps Back: Mental Causation, Reduction and Supervenience.” *Philosophical Perspectives* 11: 107–32.

———. 2003. “Do Causal Powers Drain Away?” *Philosophy and Phenomenological Research* 67: 133–50.

———. 2007. “Max Black’s Objection to Mind-Body Identity.” In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, eds. Torin Alter and Sven Walter, 249–306.

Block, Ned, and Robert Stalnaker. 1999. “Conceptual Analysis, Dualism, and the Explanatory Gap.” *Philosophical Review* 108: 1–46.

Bloj, M. G., D. Kersten, and A. C. Hurlbert. 1999. “Perception of Three-Dimensional Shape Influences Colour Perception through Mutual Illumination.” *Nature* 402: 877–79.

Bontly, Thomas. 2002. “The Supervenience Argument Generalizes.” *Philosophical Studies* 109: 75–96.

Boyaci, Huseyin, Katja Doerschner, and Laurence Maloney. 2004. “Perceived Surface Color in Binocularly Viewed Scenes with Two Light Sources Differing in Chromaticity.” *Journal of Vision* 4: 664–79.

Braddon-Mitchell, David. 2003. “Qualia and Analytical Conditionals.” *Journal of Philosophy* 100: 111–35.

Breitmeyer, Bruno. 1985. “Problems with the Psychophysics of Intention.” *Behavioral and Brain Sciences* 8: 539–40.

Brentano, Franz. 1874. *Psychology From an Empirical Standpoint*. Routledge.

Broad, C. D. 1925. *The Mind and Its Place in Nature*. Routledge & Kegan Paul.

Brodmann, Korbinian. 1909. *Localisation in the Cerebral Cortex*. trans. L. Garey. Springer.

Burge, Tyler. 1979. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4: 73–122.

———. 1993. “Mind-Body Causation and Explanatory Practice.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

———. 2010. *Origins of Objectivity*. Oxford University Press.

Cartwright, Nancy. 1994. “Fundamentalism vs. the Patchwork of Laws.” *Proceedings of the Aristotelian Society* 94: 279–92.

Chalmers, David. 1996. *The Conscious Mind*. Oxford University Press.

Churchland, Patricia S. 1986. *Neurophilosophy: Toward A Unified Science of the Mind-Brain*. MIT Press.

Churchland, Paul M. 1981. “Eliminative Materialism and the Propositional Attitudes.” *Journal of Philosophy* 78: 67–90.

———. 1984. *Matter and Consciousness*. MIT Press.

———. 1985. “Reduction, Qualia and the Direct Introspection of Brain States.” *Journal of Philosophy* 82: 8–28.

Clapp, Leonard. 2001. “Disjunctive Properties: Multiple Realizations.” *Journal of Philosophy* 98: 111–36.

Collins, John. 2000. “Preemptive Prevention.” *Journal of Philosophy* 97: 223–34.

Collins, John, Ned Hall, and L. A. Paul. 2004. “Counterfactuals and Causation: History, Problems, and Prospects.” In *Causation and Counterfactuals*, eds. John Collins, Ned Hall, and L. A. Paul. MIT Press.

Corry, Richard. 2011. “Can Dispositional Essences Ground the Laws of Nature?” *Australasian Journal of Philosophy* 89: 263–75.

Crane, Tim. 2001. *Elements of Mind*. Oxford University Press.

Davidson, Donald. 1963. “Actions, Reasons, and Causes.” *Journal of Philosophy* 60: 685–700.

———. 1969. “The Individuation of Events.” In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher, 216–34. Reidel.

———. 1970. “Mental Events.” In *Experience and Theory*, eds. Lawrence Foster and J.W. Swanson. Duckworth.

———. 1987. “Knowing One’s Own Mind.” *Proceedings and Addresses of the American Philosophical Association* 60: 441–58.

———. 1993. “Thinking Causes.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

———. 1995. “Laws and Cause.” *Dialectica* 49: 263–79.

Delk, John, and Samuel Fillenbaum. 1965. “Differences in Perceived Color as a Function of Characteristic Color.” *The American Journal of Psychology* 78: 290–93.

Dennett, Daniel. 1988. “Quining Qualia.” In *Consciousness in Contemporary Science*, eds. A. Marcel and E. Bisiach. Oxford University Press.

———. 1991. *Consciousness Explained*. Penguin.

———. 1995. “The Unimagined Preposterousness of Zombies.” *Journal of Consciousness Studies* 2: 322–26.

DeValois, Russell. 1960. “Color Vision Mechanisms in the Monkey.” *The Journal of General Physiology* 43: 115–28.

Dobrescu, Bogdan, and Don Lincoln. 2015. “Mystery of the Hidden Cosmos.” *Scientific American,* July.

Dowe, Phil. 2000. *Physical Causation*. Cambridge University Press.

Dretske, Fred. 1988. *Explaining Behavior*. Cambridge, Mass: MIT Press.

———. 1995. *Naturalizing the Mind*. MIT Press.

Eccles, John. 1985. “Mental Summation: The Timing of Voluntary Intentions by Cortical Activity.” *Behavioral and Brain Sciences* 8: 542.

Einstein, Albert, Boris Podolsky, and Nathan Rosen. 1935. “Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?” *Physical Review* 47: 777–80.

Ekroll, Vebjørn, and Franz Faul. 2012a. “Basic Characteristics of Simultaneous Color Contrast Revisited.” *Psychological Science* 23: 1246–55.

———. 2012b. “New Laws of Simultaneous Contrast?” *Seeing and Perceiving* 25: 107–41.

Ellis, Brian, and Caroline Lierse. 1994. “Dispositional Essentialism.” *Australasian Journal of Philosophy* 72: 27–45.

Enç, Berent. 1983. “In Defense of the Identity Theory.” *Journal of Philosophy* 80: 279–98.

Endicott, Ronald. 1998. “Collapse of the New Wave.” *Journal of Philosophy* 95: 53–72.

———. 2012. “Resolving Arguments by Different Conceptual Traditions of Realization.” *Philosophical Studies* 159: 41–59.

Fair, David. 1979. “Causation and the Flow of Energy.” *Erkenntnis* 14: 219–50.

Feigl, Herbert. 1958. “The ‘Mental’ and the ‘Physical.’” *Minnesota Studies in the Philosophy of Science* 2: 370–497.

Feng, Jonathan, and Mark Trodden. 2014. “Dark Worlds.” *Scientific American*, August.

Ferrier, David. 1880. *The Functions of the Brain*. G.P. Putnam’s Sons.

Feyerabend, Paul. 1962. “Explanation, Reduction and Empiricism.” In *Scientific Explanation, Space, and Time*, eds. Herbert Feigl and Grover Maxwell.

Fine, Kit. 1994. “Essence and Modality.” *Philosophical Perspectives* 8: 1–16.

———. 2012. “Guide to Ground.” In *Metaphysical Grounding*, eds. Fabrice Correia and Benjamin Schnieder, 37–80. Cambridge University Press.

Fodor, Jerry. 1968. *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House.

———. 1974. “Special Sciences (or: The Disunity of Science as a Working Hypothesis).” *Synthese* 28: 97–115.

———. 1981. “The Mind-Body Problem.” *Scientific American*, January.

———. 1987. *Psychosemantics*. MIT Press.

———. 1989. “Making Mind Matter More.” *Philosophical Topics* 17: 59–79.

———. 1994. *The Elm and the Expert*. MIT Press.

———. 1997. “Special Sciences: Still Autonomous After All These Years.” *Philosophical Perspectives* 11: 149–63.

Fodor, Jerry, and Ernest Lepore. 1991. “Why Meaning (Probably) Isn’t Conceptual Role.” *Mind & Language* 6: 328–43.

Fodor, Jerry, and Zenon Pylyshyn. 1988. “Connectionism and Cognitive Architecture.” *Cognition* 28: 3–71.

Frankish, Keith. 2007. “The Anti-Zombie Argument.” *Philosophical Quarterly* 57: 650–66.

Funkhouser, Eric. 2006. “The Determinable-Determinate Relation.” *Noûs* 40: 548–69.

Garcia, Robert. 2014. “Closing in on Causal Closure.” *Journal of Consciousness Studies* 21: 96–109.

Gates, Gary. 1996. “The Price of Information.” *Synthese* 107: 325–47.

Gertler, Brie. 1999. “A Defense of the Knowledge Argument.” *Philosophical Studies* 93: 317–36.

Ghirardi, Giancarlo. 2005. *Sneaking a Look at God’s Cards: Unraveling the Mysteries of Quantum Mechanics*. Princeton University Press.

Gibb, Sophie. 2010. “Closure Principles and the Laws of Conservation of Energy and Momentum.” *Dialectica* 64: 363–84.

Gillett, Carl. 2002. “The Dimensions of Realization: A Critique of the Standard View.” *Analysis* 62: 316–23.

———. 2003. “The Metaphysics of Realization, Multiple Realizability, and the Special Sciences.” *The Journal of Philosophy* 100: 591–603.

———. 2010. “Moving Beyond the Subset Model of Realization: The Problem of Qualitative Distinctness in the Metaphysics of Science.” *Synthese* 177: 165–92.

———. 2011. “Multiply Realizing Scientific Properties and Their Instances.” *Philosophical Psychology* 24: 727–38.

———. 2013. “Understanding the Sciences Through the Fog of ‘Functionalism(s).’” In *Functions: Selection and Mechanisms,* ed. Philippe Huneman, 159–81. Springer.

Gillett, Carl, and Bradley Rives. 2001. “Does the Argument From Realization Generalize? Responses to Kim.” *Southern Journal of Philosophy* 39: 79–98.

Godfrey-Smith, Peter. 1993. “Functions: Consensus Without Unity.” *Pacific Philosophical Quarterly* 74: 196–208.

Goldman, Alvin. 1970. *A Theory of Human Action*. Prentice-Hall.

Goldstein, E. Bruce. 2014. *Sensation and Perception*. 9th ed. Wadsworth, Cengage Learning.

Goodman, Nelson. 1955. *Fact, Fiction & Forecast*. University of London.

Grice, H. P. 1957. “Meaning.” *Philosophical Review* 66: 377–88.

Hardin, C. L. 1988. *Color for Philosophers*. Hackett.

Harman, Gilbert. 1990. “The Intrinsic Quality of Experience.” *Philosophical Perspectives* 4: 31–52.

Hart, W. D. 1988. *The Engines of the Soul*. Cambridge University Press.

Helmholtz, Hermann von. 1911. *Physiological Optics*. Dover.

Hering, Ewald. 1920. *Outlines of a Theory of the Light Sense*. Harvard University Press.

Hilbert, David, and Mark Kalderon. 2000. “Color and the Inverted Spectrum.” In *Vancouver Studies in Cognitive Science*, ed. Steven Davis, 187–214. Oxford University Press.

Hill, Christopher. 1997. “Imaginability, Conceivability, Possibility, and the Mind-Body Problem.” *Philosophical Studies* 87: 61–85.

Hill, Christopher, and Brian McLaughlin. 1999. “There Are Fewer Things in Reality Than Are Dreamt of in Chalmers’s Philosophy.” *Philosophy and Phenomenological Research* 59: 445–54.

Honderich, Ted. 1982. “The Argument for Anomalous Monism.” *Analysis* 42: 59–64.

Hooker, Clifford. 1981a. “Towards a General Theory of Reduction. Part I: Historical and Scientific Setting.” *Dialogue* 20: 38–59.

———. 1981b. “Towards a General Theory of Reduction. Part III: Cross-Categorical Reduction.” *Dialogue* 20: 496–529.

———. 1981c. “Towards a General Theory of Reduction. Part II: Identity in Reduction.” *Dialogue* 20: 201–36.

Horgan, Terence. 1989. “Mental Quausation.” *Philosophical Perspectives* 3: 47–74.

Horgan, Terence E., and John L. Tienson. 2002. “The Intentionality of Phenomenology and the Phenomenology of Intentionality.” In *Philosophy of Mind: Classical and Contemporary Readings*, ed. David Chalmers. Oxford University Press.

Hume, David. 1739/2003. *A Treatise of Human Nature*. Oxford University Press.

Hurvich, Leo, and Dorothea Jameson. 1957. “An Opponent-Process Theory of Color Vision.” *Psychological Review* 64: 384–404.

Jackson, Frank. 1982. “Epiphenomenal Qualia.” *Philosophical Quarterly* 32: 127–36.

———. 1998. *From Metaphysics to Ethics*. Oxford University Press.

Jackson, Frank, and Philip Pettit. 1990a. “Program Explanation: A General Perspective.” *Analysis* 50: 107–17.

———. 1990b. “Causation in the Philosophy of Mind.” *Philosophy and Phenomenological Research Supplement* 50: 195–214.

Jung, Richard. 1985. “Voluntary Intention and Conscious Selection in Complex Learned Action.” *Behavioral and Brain Sciences* 8: 544–45.

Kallestrup, Jesper. 2006. “The Causal Exclusion Argument.” *Philosophical Studies* 131: 459–85.

Kim, Jaegwon. 1973. “Causation, Nomic Subsumption, and the Concept of Event.” *Journal of Philosophy* 70: 217–36.

———. 1974. “Noncausal Connections.” *Noûs* 8: 41–52.

———. 1976. “Events as Property Exemplifications.” In *Action Theory,* eds. M. Brand and D. Walton, 310–26. D. Reidel.

———. 1984. “Epiphenomenal and Supervenient Causation.” *Midwest Studies in Philosophy* 9: 257–70.

———. 1989a. “Mechanism, Purpose, and Explanatory Exclusion.” *Philosophical Perspectives* 3: 77–108.

———. 1989b. “The Myth of Non-Reductive Materialism.” *Proceedings and Addresses of the American Philosophical Association* 63: 31–47.

———. 1992a. “‘Downward Causation’ in Emergentism and Nonreductive Physicalism.” In *Emergence or Reduction?: Prospects for Nonreductive*

Physicalism, eds. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim. De Gruyter.

———. 1992b. “Multiple Realization and the Metaphysics of Reduction.” *Philosophy and Phenomenological Research* 52: 1–26.

———. 1993a. “Can Supervenience and ‘Non-Strict Laws’ Save Anomalous Monism?” In *Mental Causation*, eds. John Heil and Alfred Mele, 19–26. Oxford University Press.

———. 1993b. *Supervenience and Mind*. Cambridge University Press.

———. 1993c. “The Nonreductivist’s Trouble with Mental Causation.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

———. 1998. *Mind in a Physical World*. MIT Press.

———. 2003. “Blocking Causal Drainage and Other Maintenance Chores with Mental Causation.” *Philosophy and Phenomenological Research* 67: 151–76.

———. 2007. “Causation and Mental Causation.” In *Contemporary Debates in Philosophy of Mind*, eds. Brian McLaughlin and Jonathan Cohen, 227–42. Blackwell.

———. 2011. *Philosophy of Mind*. Westview Press.

Kitcher, Patricia. 1980. “How to Reduce a Functional Psychology.” *Philosophy of Science* 47: 134–40.

———. 1982. “Genetics, Reduction and Functional Psychology.” *Philosophy of Science* 49: 633–36.

Koksvik, Ole. 2007a. “Conservation of Energy Is Relevant to Physicalism.” *Dialectica* 61: 573–82.

———. 2007b. *In Defence of Interactionism*. Masters Thesis. Monash University.

Kripke, Saul. 1980. *Naming and Necessity*. Harvard University Press.

Kroedel, Thomas. 2013. “Dualist Mental Causation and the Exclusion Problem.” *Noûs* 47: 1–19.

Kvart, Igal. 2001. “Lewis’s ‘Causation as Influence.’” *Australasian Journal of Philosophy* 79: 409–21.

Larmer, Robert. 1986. “Mind-Body Interactionism and the Conservation of Energy.” *International Philosophical Quarterly* 26: 277–85.

Latto, Richard. 1985. “Consciousness as an Experimental Variable: Problems of Definition, Practice, and Interpretation.” *Behavioral and Brain Sciences* 8: 545.

Leibniz, Gottfried Wilhelm. 1686/1989. “Discourse on Metaphysics.” In *Philosophical Essays*, trans. R. Ariew and D. Garber. Hackett.

———. 1689/1989. “Primary Truths.” In *Philosophical Essays*, trans. R. Ariew and D. Garber. Hackett.

———. 1691. “Essay on Dynamics on the Laws of Motion.” In *New Essays Concerning Human Understanding Together with an Appendix Consisting of Some of His Shorter Pieces*, trans. A. Langley. The Macmillan Company.

———. 1714. “Monadology.” In *The Monadology and Other Philosophical Writings*, trans. R. Latta. The Clarendon Press.

LePore, Ernest, and Barry Loewer. 1987. “Mind Matters.” *Journal of Philosophy* 84: 630–42.

———. 1989. “More on Making Mind Matter.” *Philosophical Topics* 17: 175–91.

Lewis, David. 1966. “An Argument for the Identity Theory.” *The Journal of Philosophy* 63: 17–25.

———. 1973a. “Causation.” *The Journal of Philosophy* 70: 556–67.

———. 1973b. *Counterfactuals*. Blackwell.

———. 1979. “Counterfactual Dependence and Time’s Arrow.” *Noûs* 13: 455–76.

———. 1980. “Mad Pain and Martian Pain.” In *Readings in the Philosophy of Psychology*, ed. Ned Block, 216–32. Harvard University Press.

———. 1986. “Causal Explanation.” In *Philosophical Papers*. Vol. II. Oxford University Press.

———. 1990. “What Experience Teaches.” In *Mind and Cognition*, ed. William Lycan, 29–57. Blackwell.

———. 2000. “Causation as Influence.” *The Journal of Philosophy* 97: 182–97.

———. 2004. “Void and Object.” In *Causation and Counterfactuals*, eds. John Collins, Ned Hall, and L. A. Paul, 277–90. MIT Press.

Libet, Benjamin. 1985. “Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action.” *Behavioral and Brain Sciences* 8: 529–39.

Livingstone, Margaret, and David Hubel. 1984. “Anatomy and Physiology of a Color System in the Primate Visual Cortex.” *The Journal of Neuroscience* 4: 309–56.

Loar, Brian. 1997. “Phenomenal States.” In *The Nature of Consciousness: Philosophical Debates*, eds. Ned Block, Owen Flanagan, and Güven Güzeldere. The MIT Press.

———. 2003. “Phenomenal Intentionality as the Basis of Mental Content.” In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, eds. Martin Hahn and B. Ramberg. MIT Press.

Loewer, Barry. 2007. “Mental Causation, or Something Near Enough.” In *Contemporary Debates in Philosophy of Mind*, eds. Brian McLaughlin and Jonathan Cohen, 243–64. Blackwell.

Lowe, E. J. 2000. “Causal Closure Principles and Emergentism.” *Philosophy* 75: 571–85.

Lycan, William. 1996. *Consciousness and Experience*. MIT Press.

Maudlin, Tim. 2011. *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Blackwell Publishers.

McCall, Storrs. 2013. “Does the Brain Lead the Mind?” *Philosophy and Phenomenological Research* 86: 262–65.

McLaughlin, Brian. 1993. “On Davidson’s Response to the Charge of Epiphenomenalism.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

Mellor, D. H. 1995. *The Facts of Causation*. Routledge.

Melnyk, Andrew. 2006. “Realization and the Formulation of Physicalism.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 131: 127–55.

Menzies, Peter. 2003. “The Causal Efficacy of Mental States.” In *Physicalism and Mental Causation*, eds. Sven Walter and Heinz-Dieter Heckmann. Imprint Academic.

———. 2004. “Difference-Making in Context.” In *Causation and Counterfactuals*, eds. John Collins, Ned Hall, and L. A. Paul. MIT Press.

Merricks, Trenton. 2001. *Objects and Persons*. Clarendon Press.

Millikan, Ruth. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

———. 1989. “Biosemantics.” *Journal of Philosophy* 86: 281–97.

Molnar, George. 2003. *Powers: A Study in Metaphysics*. Oxford University Press.

Montero, Barbara. 2006. “What Does the Conservation of Energy Have to Do with Physicalism?” *Dialectica* 60: 383–96.

Mumford, Stephen. 2004. *Laws in Nature*. Routledge.

———. 2006. “The Ungrounded Argument.” *Synthese* 149: 471–89.

Murphy, Timothy, and Dale Corbett. 2009. “Plasticity During Stroke Recovery: From Synapse to Behaviour.” *Nature Reviews Neuroscience* 10: 861–72.

Näätänen, Risto. 1985. “Brain Physiology and the Unconscious Initiation of Movements.” *Behavioral and Brain Sciences* 8: 549.

Nagel, Ernest. 1961. *The Structure of Science*. Harcourt, Brace & World.

Nemirow, Laurence. 1980. “Review of Nagel’s Mortal Questions.” *Philosophical Review* 89: 473–77.

Noordhof, Paul. 1999. “Micro-Based Properties and the Supervenience Argument: A Response to Kim.” *Proceedings of the Aristotelian Society* 99: 115–18.

Papineau, David. 2001. “The Rise of Physicalism.” In *Physicalism and Its Discontents*, eds. Carl Gillett and Barry Loewer. Cambridge University Press.

———. 2007. “Phenomenal and Perceptual Concepts.” In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism,* eds. Torin Alter and Sven Walter, 111–44. Oxford University Press.

Peacocke, Christopher. 1983. *Sense and Content: Experience, Thought, and Their Relations*. Oxford University Press.

Pereboom, Derk, and Hilary Kornblith. 1991. “The Metaphysics of Irreducibility.” *Philosophical Studies* 63: 125–45.

Place, Ullin T. 1956. “Is Consciousness a Brain Process?” *British Journal of Psychology* 47: 44–50.

Polger, Thomas. 2007. “Realization and the Metaphysics of Mind.” *Australasian Journal of Philosophy* 85: 233–59.

Polger, Thomas, and Lawrence Shapiro. 2008. “Understanding the Dimensions of Realization.” *Journal of Philosophy* 105: 213–22.

Popper, Karl, and John Eccles. 1977. *The Self and Its Brain*. Springer.

Putnam, Hilary. 1967. “Psychological Predicates.” In *Art, Mind, and Religion*, eds. W.H. Capitan and D.D. Merrill. University of Pittsburgh Press.

———. 1975. “The Meaning of ‘Meaning.’” *Minnesota Studies in the Philosophy of Science* 7: 131–93.

———. 1999. *The Threefold Cord: Mind, Body and World*. Columbia University Press.

Pylyshyn, Zenon. 1984. *Computation and Cognition*. MIT Press.

Quine, Willard V.O. 1969. “Natural Kinds.” In *Ontological Relativity and Other Essays*, eds. Jaegwon Kim and Ernest Sosa, 114–38. Columbia University Press.

Quine, W. V. 1960. *Word and Object*. MIT Press.

Raymont, Paul. 1999. “The Know-How Response to Jackson’s Knowledge Argument.” *Journal of Philosophical Research* 24: 113–26.

Reich, Eugenie. 2010. “Chameleon Cosmos.” *New Scientist*, July 31.

Reid, Thomas. 1788/2010. *Essays on the Active Powers of Man*. Edinburgh University Press.

Rey, Georges. 2001. “Physicalism and Psychology: A Plea for a Substantive Philosophy of Mind.” In *Physicalism and Its Discontents*, eds. Carl Gillett and Barry Loewer. Cambridge University Press.

Richardson, Robert. 1979. “Functionalism and Reductionism.” *Philosophy of Science* 46: 533–58.

———. 1982. “How Not to Reduce a Functional Psychology.” *Philosophy of Science* 49: 125–37.

Rinner, Oliver, and Karl Gegenfurtner. 2000. “Time Course of Chromatic Adaptation for Color Appearance and Discrimination.” *Vision Research* 40: 1813–26.

Salmon, Wesley. 1997. “Causality and Explanation: A Reply to Two Critiques.” *Philosophy of Science* 64: 461–77.

Schaffer, Jonathan. 2001. “Causation, Influence, and Effluence.” *Analysis* 61: 11–19.

Schaffner, Kenneth. 1967. “Approaches to Reduction.” *Philosophy of Science* 34: 137–47.

Searle, John. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

———. 2004. *Mind: A Brief Introduction*. Oxford University Press.

Shagrir, Oron. 1998. “Multiple Realization, Computation and the Taxonomy of Psychological States.” *Synthese* 114: 445–61.

Shapiro, Lawrence. 2000. “Multiple Realizations.” *Journal of Philosophy* 97: 635–54.

———. 2004. *The Mind Incarnate*. MIT Press.

Shapiro, Lawrence, and Elliott Sober. 2007. “Epiphenomenalism - the Do’s and the Don’ts.” In *Thinking about Causes: From Greek Philosophy to Modern Physics*, eds. G. Wolters and P. Machamer, 235–64. University of Pittsburgh Press.

Shoemaker, Sydney. 2001. “Realization and Mental Causation.” In *Physicalism and Its Discontents*, eds. Carl Gillett and Barry Loewer, 23–33. Cambridge University Press.

Sider, Theodore. 1993. “Van Inwagen and the Possibility of Gunk.” *Analysis* 53: 285–89.

———. 2003. “What’s so Bad About Overdetermination?” *Philosophy and Phenomenological Research* 67: 719–26.

Sklar, Lawrence. 1993. *Physics and Chance*. Cambridge University Press.

Smart, J. J. C. 1959. “Sensations and Brain Processes.” *Philosophical Review* 68: 141–56.

Sober, Elliott. 1999. “The Multiple Realizability Argument Against Reductionism.” *Philosophy of Science* 66: 542–64.

Sosa, Ernest. 1984. “Mind-Body Interaction and Supervenient Causation.” *Midwest Studies in Philosophy* 9: 271–81.

———. 1993. “Davidson’s Thinking Causes.” In *Mental Causation*, eds. John Heil and Alfred Mele. Oxford University Press.

Spinoza, Baruch. 1677/1955. *Ethics*. trans. R.H.M. Elwes. Dover.

Stalnaker, Robert. 1968. “A Theory of Conditionals.” In *Studies in Logical Theory*, ed. Nicholas Rescher, 98–112. Blackwell.

Stamm, John. 1985. “The Uncertainty Principle in Psychology.” *Behavioral and Brain Sciences* 8: 553.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science*. MIT Press.

Strevens, Michael. 2003. “Against Lewis’s New Theory of Causation: A Story with Three Morals.” *Pacific Philosophical Quarterly* 84: 398–412.

Sturgeon, Scott. 1998. “Physicalism and Overdetermination.” *Mind* 107: 411–32.

Suppe, Frederick. 1974. “The Search for Philosophic Understanding of Scientific Theories.” In *The Structure of Scientific Theories*, ed. Frederick Suppe, 3–241. University of Illinois Press.

Suppes, Patrick. 1960. “A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences.” *Synthese* 12: 287–301.

———. 1967. “What Is a Scientific Theory?” In *Philosophy of Science Today*, ed. Sidney Morgenbesser, 55–67. Basic Books, Inc.

Svaetichin, Gunnar, and Edward F. MacNichol. 1958. “Retinal Mechanisms for Chromatic and Achromatic Vision.” *Annals of the New York Academy of Sciences* 74: 385–404.

Swoyer, Chris. 1982. “The Nature of Natural Laws.” *Australasian Journal of Philosophy* 60: 203–23.

Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press.

———. 1999. “Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion.” *Mind* 108: 705–25.

———. 2003. “A Theory of Phenomenal Concepts.” In *Minds and Persons*, ed. Anthony O’Hear. Cambridge University Press.

———. 2009. *Consciousness Revisited: Materialism Without Phenomenal Concepts*. MIT Press.

Unger, Peter. 1979. “There Are No Ordinary Things.” *Synthese* 41: 117–54.

Van Fraassen, Bas. 1980. *The Scientific Image*. Oxford University Press.

———. 1987. “The Semantic Approach to Scientific Theories.” In *The Process of Science*, ed. Nancy Nersessian, 105–24. Springer.

van Inwagen, Peter. 1990. *Material Beings*. Cornell University Press.

Vicente, Agustín. 2006. “On the Causal Completeness of Physics.” *International Studies in the Philosophy of Science* 20: 149–71.

Walter, Sven. 2005. “Program Explanations and Causal Relevance.” *Acta Analytica* 20: 32–47.

———. 2006. “Multiple Realizability and Reduction: A Defense of the Disjunctive Move.” *Metaphysica* 7: 43–65.

———. 2008. “The Supervenience Argument, Overdetermination, and Causal Drainage: Assessing Kim’s Master Argument.” *Philosophical Psychology* 21: 673–96.

Wandell, Brian. 1995. *Foundations of Vision*. Sinauer Associates.

White, Stephen. 2007. “Property Dualism, Phenomenal Concepts, and the Semantic Premise.” In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, eds. Torin Alter and Sven Walter. Oxford University Press.

Wigner, Eugene. 1961. “Remarks on the Mind-Body Problem.” In *The Scientist Speculates,* ed. I. J. Good. Heineman.

Wilson, Jessica. 1999. “How Superduper Does a Physicalist Supervenience Need to Be?” *Philosophical Quarterly* 50: 33–52.

———. 2005. “Supervenience-Based Formulations of Physicalism.” *Noûs* 39: 426–59.

———. 2009. “Determination, Realization and Mental Causation.” *Philosophical Studies* 145: 149–69.

———. 2010. “What Is Hume’s Dictum, and Why Believe It?” *Philosophy and Phenomenological Research* 80: 595–637.

———. 2011. “Non-Reductive Realization and the Powers-Based Subset Strategy.” *The Monist* 94: 121–54.

Witmer, D. Gene. 2001. “Sufficiency Claims and Physicalism: A Formulation.” In *Physicalism and Its Discontents,* eds. Carl Gillett and Barry Loewer. Cambridge University Press.

Yablo, Stephen. 1992a. “Cause and Essence.” *Synthese* 93: 403–49.

———. 1992b. “Mental Causation.” *Philosophical Review* 101: 245–80.

———. 1993. “Is Conceivability a Guide to Possibility?” *Philosophy and Phenomenological Research* 53: 1–42.

Yates, David. 2009. “Emergence, Downwards Causation and the Completeness of Physics.” *Philosophical Quarterly* 59: 110–31.

Young, Thomas. 1802. “The Bakerian Lecture: On the Theory of Light and Colours.” *Philosophical Transactions of the Royal Society of London* 92: 12–48.

Zeki, Semir. 1983a. “Colour Coding in the Cerebral Cortex: The Reaction of Cells in Monkey Visual Cortex to Wavelengths and Colours.” *Neuroscience* 9: 741–65.

———. 1983b. “Colour Coding in the Cerebral Cortex: The Responses of Wavelength-Selective and Colour-Coded Cells in Monkey Visual Cortex to Changes in Wavelength Composition.” *Neuroscience* 9: 767–81.