The Disturbing Matter of Downward Causation

A Study of the Exclusion Argument and its Causal-Explanatory Presuppositions

Ph.D. Dissertation

Øistein Schmidt Galaaen

Program in Philosophy and
Humanistic Informatics

Faculty of Humanities
University of Oslo

2006

Series of dissertations submitted to the
Faculty of Humanities,University of Oslo
No. 299

ISSN 0806-3222

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS, Oslo, 2007.

Produced in co-operation with Unipub AS.
The thesis is produced by Unipub AS merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Unipub AS is owned by
The University Foundation for Student Life (SiO)

There is no twisted thought without a twisted molecule.

– Attributed to neurophysiologist Ralph Gerard

Table of Contents

Acknowledgments vi
1. Preface viii
2. Introduction 1
2.1. Downward Causation from Descartes to Jaegwon Kim 2
2.2. The Significance of Downward Causation 6
2.3. Exclusion Arguments and Their Presuppositions 10
2.4. Aims and Methods: How Should We Approach Downward Causation? 15
2.5. "The Physical" in Downward Causation 21
2.6. Overdetermination and Causal Competition 23
2.6.1. Synopsis of Paper (#1) 24
2.6.2. Synopsis Paper (#2) 27
2.7. Causal Closure and Physical Causes of Bodily Movements 31
2.7.1. Synopsis Paper (#3) 32
2.7.2. Synopsis Paper (#4) 34
2.8. Concluding Remarks on the Prospects of Exclusion Arguments 38
3. Paper (#1): Mechanisms Do Not Overdetermine Their Effects 40
3.1. Introduction 40
3.2. Loewer's Case against Production 43
3.3. Characterizing the Productive View 45
3.4. First Response to Loewer: Empirical Theories of Causation 49
3.5. Second Response to Loewer: Mechanisms 51
3.6. Conclusions 63
4. Paper (#2): Pace Burge: Some Empirical Warrant for Epiphobia 66
4.1. Introduction. Physicalist Constraints on Causal Relevance in Psychology 66
4.2. Burge's Theory of Causal Explanation and Causal Relevance 71
4.3. Burge's Arguments Against Physicalist Constraints on Causal Relevance 76
4.3.1. (P1) and the Call for Tight Relations between Mental and Physical Causes .. 77
4.3.2. (P2) and the Call for Mechanisms 79

4.3.3. (P3) and the Call for Locally Supervenient Properties 81
4.3.4. (P4) and the Call for Psychological Laws..... 82
4.4. Causal Relevance in Psychology and Neuroscience..... 84
4.4.1. Neural and Psychological Patterns of Events 85
4.4.2. Patricia Goldman-Rakic' Theory of Prefrontal Cortex..... 95
4.4.3. Mental Causation, Mechanisms and Part-Whole Relations..... 103
5. Paper (#3): Is there a Binding Problem of Behavior? E.J. Lowe on Causal Closure . 107
5.1. Introduction..... 108
5.2. The Case for Causal Closure..... 110
5.3. The Possibility of Invisible Mental Causation..... 112
5.4. Lowe's Argument for the Plausibility of Invisible Mental Causation..... 114
5.5. Against Invisible Mental Causation..... 118
5.5.1. Lowe's Argument is Inconclusive 119
5.5.2. Invisible Mental Causation Fits Ill with Neuroscientific Practice 121
5.5.3. Bodily Movements Do Not Appear to Be Rendered Coincidental..... 126
5.6. Conclusions..... 131
6. Paper (#4): What Is Closed in Causal Closure?..... 132
6.1. Introduction..... 132
6.2. The Significance of Causal Closure..... 133
6.3. The Empirical Status of Closure..... 136
6.4. Sturgeon's Challenge..... 138
6.5. Lessons from Sturgeon's Challenge 143
6.6. Neurobiological quasi-Closure 146
6.6.1. Neurobiological quasi-Closure and a Neural Exclusion Argument..... 147
6.6.2. Arguing for quasi-Closure 152
6.6.3. Extrapolating from these Models..... 167
6.7. Conclusions..... 169
References..... 170

Acknowledgments

I began working on this project in July 2003, after receiving a grant from the Faculty of Humanities at the University of Oslo. The present thesis is the result of three years work, primarily conducted at the Department of Philosophy at the University of Oslo. Throughout this period the department has provided me with a very friendly and open-minded workplace. I offer a huge general "thank you" to my friends and colleagues here for making these three years highly enjoyable, not only professionally, but also socially speaking.

Special thanks are due to my supervisor, Bjørn Torgrim Ramberg. Bjørn has patiently followed my explorations in the philosophy of mind since my Masters Thesis on emergentism. He always combines criticism with encouragement and charitable interpretation. Bjørn has the gift of being able to make old problems look more interesting by offering refreshingly new metaphilosophical perspectives. No doubt he is also partly responsible for sparking my interest in investigating how neuroscience can be brought to bear on the philosophy of mind.

The participants at the department's Dissertation Seminar should all be thanked both for providing feedback on the two occasions when I have presented drafts there, as well as for allowing me to learn about their own work. Four of my fellow Ph.D. students – Jon Anders Lindstrøm, Gry Oftedal, Lars Bjarne Reinholdtsen and Anders Strand – deserve special mention. I have greatly enjoyed the informal colloquia we have organized as well as our trips to conferences. In their various ways these four have all influenced my philosophy.

My interest in mental causation was originally kindled by Carsten Martin Hansen's seminar on Jaegwon Kim's philosophy in the spring term of 2000. I have since had the pleasure of discussing philosophy with him on many occasions, and have benefited enormously from his very perceptive comments on my work. In the same spirit I thank Nils Roll-Hansen for penetrating comments on, and discussions of, physicalism. Thanks also to Lene Bomann-Larsen and Jakob Elster for commenting on the introduction to this thesis. Jakob also provided some very perceptive advice about my

criticism of Tyler Burge. Anders Nes should be credited for some helpful suggestions about dualism.

My work has also benefited from contact with researchers abroad. I thank John Bickle for discussions of neuroscience, guidance in my tentative explorations into the scientific literature of that field, as well as for commenting on parts of my work. Iris Oved came up with some really useful suggestions about productive causation and about my arguments against E.J. Lowe's dualism. Similarly, I thank Stathis Psillos for commenting on my ideas about causation and causal concepts. Jaegwon Kim, Barry Loewer and David Papineau have all offered valuable clarifications of their positions in response to my e-mails. I thank them for that.

A generous grant from the U.S.-Norway Fulbright Foundation for Educational Exchange enabled me to stay as a visiting scholar at Rutgers University in the U.S. from September 2005 to February 2006. The Department of Philosophy and the Center for Cognitive Science at Rutgers proved to be very exciting places to visit indeed, and I learnt a lot from attending lectures and from discussing with faculty and students there. In particular, I would like to thank Brian McLaughlin for meeting with me to discuss philosophy even though he was actually on leave at the time. Thanks also to Ernest Lepore, who kindly provided me with office space at the Center for Cognitive Science.

I have presented parts of the present thesis and related work of mine at the *Fifth European Congress for Analytic Philosophy (ECAP5)* in Lisbon, Portugal (August 2005), the *Department Colloquium* at the Department of Philosophy at the University of Oslo (March 2006) and at the conference *Computers & Philosophy (i-C&P 2006)* in Laval, France (May 2006). I thank the audience at these venues for useful feedback.

"Last, but not least," as they say, I want to thank my wife Pernille Grindaker. Definitely not least. Her patience, encouragement and flexible attitude were decisive factors in making this thesis and my stay at Rutgers possible.

The thesis is dedicated to my daughter Aurora.

Oslo, September 2006

1. Preface

The present thesis is submitted for the degree Philosophiae Doctor (Ph.D.) as a collection of papers. It consists of four individual papers, which are motivated, summarized and compared in the introduction. These papers are:

(#1) "Mechanisms Do Not Overdetermine Their Effects"
(#2) "Pace Burge: Some Empirical Warrant for Epiphobia"
(#3) "Is there a Binding Problem of Behavior? E.J. Lowe on Causal Closure"
(#4) "What's Closed in Causal Closure?"

Traditionally, most Norwegian theses in philosophy have been monographs. Nevertheless, the Faculty of Humanities accepts submissions of paper based theses; provided the individual papers are related to each other and these relations are accounted for.[1]

Papers (#1)-(#4) are all thematically related insofar as they concern the so-called "Exclusion Argument" in the philosophy of mind. They are all largely dedicated to discussions of responses that have been made to this argument. Papers (#1) and (#2) discuss problems with the so-called "No Overdetermination" premise, which plays a key role in the Exclusion Argument. Similarly, papers (#3) and (#4) concern problems with another premise, known as "Causal Closure." Finally, all papers share a common methodology and aim, insofar as they are attempts to see how considerations of explanatory practice, and in particular of neuroscience, can be brought to bear on problems like the Exclusion Argument. Their interrelations and implications are discussed further in the introduction. In spite of these unifying factors, the papers were written independently and with different aims in mind, making a paper based presentation natural. The following papers are all under submission to journals. However, I have taken the liberty of using the format of a thesis to develop more fully some of the points

[1] Cp. the guidelines for such theses, adopted by the Faculty's Research Committee, 30. August, 2004.

made in the papers, in particular with regard to the themes that connect them to each other. The versions of the papers presented in the thesis are therefore in some cases significantly longer than those submitted for separate publication in journals.

A complete reference list for all the papers and the introduction is included at the end of the thesis. The reference style is in accordance with the American Psychological Association's standards, as two of the papers have been submitted to a journal which practices that style. All italics within quotes are from the quoted authors, unless otherwise indicated. Comments or substitutions in brackets ("[…]") within quotes are mine.

For one source to the first page, supervenience-like claim attributed to Ralph Gerard, see Elliot S. Valenstein, 2005, p. 161.

The thesis was supervised by Professor Bjørn Torgrim Ramberg at the Philosophy Department, Faculty of Humanities at the University of Oslo. Funding was provided by the Faculty of Humanities and the U.S.-Norway Fulbright Foundation for Educational Exchange.

2. Introduction

In this introduction I describe and motivate the problems discussed in the papers that constitute the thesis' argumentative bulk. (2.1-2.2) These papers all concern the so-called "Exclusion Argument" which is described in (2.3). Aims and methods are set forth in (2.4-2.5). The two first papers discuss problems with the Exclusion Argument's "No Overdetermination" principle, whereas the final two are dedicated to problems with the "Causal Closure" principle. These problems and my conclusions with respect to them are described and summarized in (2.6) and (2.7), respectively. The introduction ends with some concluding remarks about the prospects for the Exclusion Argument. (2.8)

There are three reasons why this introduction is relatively lengthy. *First*, the Exclusion Argument is but one part of a larger set of problems involving "downward causation," that are to a certain extent also discussed in the papers. In the introduction I therefore spend some time situating the Exclusion Argument within a larger argumentative and historical context. *Second*, part of the thesis' impact is metaphilosophical, and concerns the nature of problems of downward causation and how these should be approached. In this respect, the introduction serves to describe and motivate my methodology. *Finally*, while the papers display a high degree of thematic unity, they were written with different arguments and aims in mind. With respect to these aims, they speak for themselves, and the thesis' main arguments are to be found therein. But in this introduction I also endeavor to spell out their interrelations and, somewhat more tentatively, to describe the implications they appear to have for the Exclusion Argument.

I have allowed myself a liberal use of quotes in the introduction; in the hope that they will add color to the problems and help the reader appreciate the varieties of voices and views that can be found in the debate. Tyler Burge figures prominently in my introduction and more so than the other philosophers I discuss later on, because the ways in which I agree and disagree with him are central to my approach.

2.1. Downward Causation from Descartes to Jaegwon Kim

The subject matter of this thesis is sometimes described by the rather gloomy-sounding phrase "downward causation." This expression appears to have been coined by Donald T. Campbell (1974) in an attempt to understand complex biological systems. The idea has since been invoked for similar purposes by scientists and philosophers. (Andersen et al., 2000; Sperry, 1986) It is however, primarily downward causation from the mental to the physical that will concern me here. How can mental events exercise a downward causal influence on underlying physical processes? How can, for instance the onset of beliefs and desires cause bodily movements when I act? The idea of downward causation is closely related to antireductionism about the mental. If mental events *just are* physical events, as reductionists would have it, the claim that they exert their influence from *above* the physical level seems less natural, except perhaps in accordance with a purely descriptive notion of levels. And whether antireductionism about the mental is a viable position was in fact my chief motive for delving into problems of downward causation in the first place.

Talk of downward causation presupposes some way of imposing an upward-downward direction on causal processes. The frequently invoked picture of the world as stratified into different levels of complexity, ranging from the fundamentally physical, via the chemical and the biological to the mental and the social does just this. (Kim, 2002b; Oppenheim & Putnam, 1958) The ideas of levels and downward causation were also invoked by classical emergentists, like C.D. Broad (1925), who were arguably historical predecessors of today's antireductive physicalists. (Kim, 1992) Furthermore, talk of levels is widespread in neurobiology, a science to which we shall have occasion to return. Here relevant levels include *inter alia* cognitive, systems and cellular/molecular neuroscience. (Bear et al., 2001, pp. 13-14) Whether levels-talk in science and philosophy should be taken with ontological seriousness or rather treated as useful heuristics is, however, very much debatable (Kim, 2002b) Paul Oppenheim and Hilary Putnam's levels of complexity may, for example, turn out to be more like David Marr's (1982, sect. 1.2) famous levels for computational psychology, that is the levels of (i) which function is computed, (ii) which algorithm is used to compute it, and finally (iii), how the algorithm is implemented physically. These are arguably mere levels of analysis

or description. It is at least not clear whether they correspond to levels of existence or anything of that sort. Indeed, "downward causation" might turn out to be just a figure of speech.

Nevertheless, situating "downward" mental causation within a hierarchy of levels lends vivid sense to the antireductionist idea that the mental can somehow be something "over and above" the physical, and yet affect the physical causally. But even setting antireductionism aside, macro levels are often pictured as being "above" micro levels. As we shall see, it may be that mental-to-physical causation involves events at the macro level causing effects at the micro-level. Indeed, historically scientists and emergentist philosophers appear to have postulated what Brian McLaughlin (1992) calls "configurational forces." These are special (say, mental or vital) but fundamental forces that are only exercised by objects of some complexity and exert a downward causal influence on objects at the level of parts. For these reasons I shall stick to the phrase "downward causation" in this introduction.

As Jaegwon Kim likes to point out, problems of downward causation from the mental to the physical are not new to philosophers. (Kim, 1998, ch. 2 and forthcoming.) Contemporary problems are in important ways similar to problems that plagued Descartes. After arguing for substance dualism – the view that Body and Soul are distinct substances – Descartes found himself hard-pressed to explain how the Soul can cause bodily movements in actions. This question was raised *inter alia* in a famous letter princess Elisabeth of Bohemia wrote to Descartes.

> How can the soul of a man determine the spirits of his body so as to produce voluntary actions (given that the soul is only a thinking substance)?[2]

In that letter, Elisabeth worried that being non-extended the Soul could not affect the Body, as the causal mechanisms in Cartesian physics – in particular the mechanism of *pushing* – require both cause and effect to be extended. (In paper (#4) we shall see that

---

[2] Elisabeth's letter to Descartes of May 6/16, 1643. Translated in Nye, 1999, p. 9 / AT III 661. All references marked "AT" are to the Adam & Tannery (1964-1976) edition.

analogous ideas about the nature of *neural* mechanisms put constraints on what causes are relevant in cellular/molecular neuroscience.) This worry appears to have made her contemplate localizing the Soul within the spatial, physical domain after all.

> I confess that it is easier for me to concede the matter and the extension of the soul than to concede that a being that is immaterial has the capacity to move a body and to be moved by it.[3]

> Although physical extension may not be necessary to thought, it isn't repugnant to it either, and could derive from some other function of the soul, one not less essential to it.[4]

As we shall see throughout the introduction, Elisabeth's arguments provide a striking parallel to the modern problems we will be considering.[5] Anyway, conventional philosophical wisdom has it that objections like that raised by Elisabeth lead to the demise of substance dualism, and the rise of physical monism. Today, most philosophers of mind believe there is only one type of substances or things, and that these are *physical* substances. All things, save perhaps abstract entities like numbers, are physical insofar as they have physical properties and are located within the spatio-temporal domain.

Physical monism may be viewed as a first step in the direction of *physicalism*, a doctrine to which we shall return repeatedly. While physicalism is a broad church, most of its followers, including myself, will agree upon crediting the physical domain with a certain ontological primacy *vis-à-vis* the mental and other non-physical domains. In crude outline, this means that every non-physical property *depends* wholly on physical properties for its instantiation, but not the other way around. (Kim, 1984a) Put slightly differently, all non-physical facts or phenomena etc. obtain or occur *in virtue of* physical facts or phenomena etc. (Witmer, 2001) Typically, physicalists attempt to cash in such claims by way of some notion of supervenience. They say, for instance, that worlds (or

---

[3] Elisabeth's letter to Descartes of June 10/20, 1643. Translated in Nye, 1999, p. 22 / AT III 685.

[4] Elisabeth's letter to Descartes of June 10/20, 1643. Translated in Nye, 1999, p. 26 / AT III 685.

[5] In particular, the arguments invoked in the Descartes-Elisabeth correspondence strongly resemble those in Burge's disputes with contemporary physicalists. I discuss this more fully in a Norwegian publication of mine. (Galaaen, 2006)

perhaps things, or perhaps regions) that are indiscernible in physical respects are indiscernible in *all* respects. (Horgan, 1982; Kim, 1984a; Lewis, 1983) In a frequently invoked theological metaphor supervenience would have been very convenient for God during Genesis. Once He had fixed the physical laws and the distribution of physical facts, the rest of Creation, including mentality, would fall into place automatically. There are a number of modal twists and turns to this story, and it is not entirely clear whether supervenience really captures the idea of psychophysical dependencies. (Kim, 1998, pp. 9-15) These difficulties need not concern us for present purposes. Supervenience amounts at least to what database theorists call a *functional dependency* between the physical and the non-physical, and this is a strong claim in its own right. (Garcia-Molina et al., 2002, ch. 3.4) Since physical indiscernibility guarantees non-physical indiscernibility, a physical description of worlds (or objects or regions) could work as a primary key in a database containing *all* information about the worlds (or objects or regions). By looking up the physical key for a world (or object or region), we could read all there is to know about that world (or object or region) from the database.

By assenting to the in-principle possibility of such a database, physicalists go a long way towards crowning physics as the queen of the sciences. Nevertheless, physicalism, thus understood, does not necessarily amount to reductionism about the mental. For instance, if one thinks of the mental in terms of properties, physicalism appears *prima facie* compatible with some things – say, human beings – having irreducible, but supervenient, mental properties in addition to their physical, subvenient properties. Similarly, physicalism is compatible with the essential need for non-physical methods and concepts in the non-physical sciences.

In fact, as Kim emphasizes, some form of *antireductive* physicalism about mentality has emerged as the mainstream post-Cartesian view. (Kim, 1998, p. 2) Accordingly, antireductive physicalists typically contend that supervenience is physicalism enough. Reductive physicalists disagree. Kim (1998; 2005), for instance, thinks problems of downward causation have returned to haunt antireductive physicalists; driving them in the direction of outright reductionism.

Now, there are in fact many problems of downward causation. Elisabeth's was the one of finding a psychophysical mechanism for downward causation that is compatible

with the nature of physical causal mechanisms. Leibniz, on the other hand, argued that downward causation is incompatible with the laws of physics, as it would violate his conservation laws. (See, e.g., McLaughlin (1993) or Papineau (2001) for discussion.) Varieties of these and other problems of downward causation will recur later on in this thesis. They all threaten to render the mental epiphenomenal – that is, causally inert – with respect to the physical.[6] I shall be focusing primarily on the incompatibility which Kim perceives, however. *How can downward causation be compatible with the so-called "Causal Closure" of physics?* In outline this physicalist principle contends that all physical effects like bodily movements have sufficient *physical* causes. Given such physical causes, and the additional assumption that physical effects are not generally causally overdetermined, Kim thinks there is no *room* for irreducibly mental causes of bodily movement. Unless mental events can be *identified* with physical events, they are excluded as epiphenomena. This, in a nutshell is the "Exclusion Argument." It goes by other names as well, but is perhaps most befittingly described, by Kim, as "Descartes's revenge against the physicalists." (Kim, 1998, p. 28) If Kim is right, replacing substance dualism with antireductive physicalism will not save the causal efficacy of the mental. The antireductionist's irreducibly mental events will be condemned to the very epiphenomenalist fate Elisabeth predicted for non-spatial Souls. Kim may well be right that such a turn of events would have amused Descartes. (Kim, 1998, p. 39)

2.2. The Significance of Downward Causation

Problems of downward causation, then, involve heavy-weight philosophical questions related to the time-worn, but arguably still unresolved (and un-*dissolved*), mind-body problem. (1) Can epiphenomenalism about the mental be ruled out? If so – how? (2) Does the mental reduce to the physical? If so – how? It is not hard to come up with reasons for

---

[6] "Epiphenomenalism with respect to physical events" is not the same as "epiphenomenalism *period*." The former kind of epiphenomenalism appears at first blush compatible with mental events causing other, non-physical events. For simplicity's sake I shall nevertheless sometimes use "epiphenomenalism" to refer to the more restricted claim that mental events do not cause physical events. The context will make clear which sense is intended.

caring about these issues. First, and quite independently of the question of antireductionism, epiphenomenalism appears to be an immensely unattractive position. Consider just the notion of agency and that of its companion, moral responsibility. Arguably, responsible agency is in most cases dependent on bodily movements being caused by purported mental entities like intentions. Hence, agency would seem to presuppose the reality of mental causation. In epiphenomenalism, on the other hand, there is no room for agents that change the world through their actions. Our limbs move, but the impression that they sometimes move because we *want* them to turns out to be a grand illusion. The way of the world is strictly under the control of physical, non-mental causes. An oft-quoted passage from Jerry Fodor adds considerable drama to this, and makes epiphenomenalism seem even less like an option to be seriously considered.

> […] if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying …, if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world. (Fodor, 1990, p. 156)

In my view, Fodor clinches the case. Any doctrine that is committed to epiphenomenalism has to be dismissed, one way or the other. *Antireductionists* – for whom mental causation becomes *downward* causation – should therefore be very much interested in defending the possibility of this kind of causation. They need, among other things, to find a loophole in the Exclusion Argument. Most *reductionists*, of course, are equally eager to salvage agency and mental causation. Kim, for one, makes this very clear. (Kim, 2005, p. 9) However, for reductionists there is no antecedent commitment to the causation being *downward.* For them, mental causation is just a species of physical causation. Nevertheless, as we shall see, some reductionists use problems of downward causation to motivate and/or argue for their reductionism. Accordingly, reductionists, too, should find problems of downward causation interesting. If antireductionism really leads to epiphenomenalism, that certainly counts strongly in favor of the reductionists' position.

That is not to say that reductionism is an attractive option. In the eyes of many, the second question I posed above, about the viability of reductionism, is intimately

related to psychology's autonomy *vis-à-vis* the physical sciences. (Van Gulick, 1992; Fodor, 1974) Suppose problems of mental causation require us to reduce the mental to the physical. Then the physical sciences achieve a kind of hegemony that, at least in principle, if not in practice may seem incompatible with the autonomy of psychology. Indeed, if the Exclusion Argument generalizes to other non-physical sciences, it would reduce even the geological and the biological to the physical. (Fodor, 1990; Block, 2003)[7] Many appear to find the idea of this much concilience both unpalatable and unrealistic, and argue instead for a picture of a more "dappled world," that is studied by a plurality of relatively autonomous sciences. (Cartwright, 1999)

I admit that it is not clear to me what weight the autonomy worries about reductionism have, nor what kind of autonomy we can reasonably hope for.[8] The autonomy worry may be related to a slightly different set of worries. Kim portrays the modern mind-body problem as one of:

> […] accommodating the mental within a principled physicalist scheme, while at the same time preserving it as something distinctive – that is, without losing what we value, or find special, in our nature as creatures with minds. (Kim, 1998, p. 2)

He probably has in mind features like qualitative consciousness and intentionality or "aboutness." These features, which are frequently attributed to the mental, appear both special and – at least at first blush – non-physical. Remarking on the place of intentionality in a physical world, Fodor may incidentally also capture one reason for

---

[7] This is the so-called "generalization argument." Its proponents, like Fodor and Block, use it to argue that there must be something wrong with the Exclusion Argument, as the mainstream view has it that special sciences like biology are *not* reducible to physics. For the purposes of the present thesis I set this response to the Exclusion Argument aside.

[8] For one thing, a lot of research in cognitive psychology takes neuroscientific evidence into explicit account when choosing between theories about (say) the structure of working memory. (Baddeley, 2003) Hence, at least one branch of psychology does not appear to be autonomous in a very strict sense. On the other hand, it is highly likely that psychological and behavioral methods will remain essential at least as heuristic tools in the foreseeable future. So psychology is certainly autonomous in *some* sense.

taking mind-body reductionism to be threatening: "If aboutness is real, it must be really something else." (Fodor, 1987, p. 97) Then there are putative normative features of the mental. John McDowell (1994), for instance, spends much time defending "the logical space of reasons" – in which he situates *inter alia* intentional states like beliefs – against perceived physicalist attacks. He does so because he worries that intentional states cannot be justified, or justify each other, if they belong wholly within "the logical space of nature," where we find causal, but no *reason-giving*, relations.

Cherished features of the mental, then, appear to be at stake. Reductionism might be taken to deprive the mental of its special character, or perhaps eliminate that character altogether. In Nathaniel Hawthorne's *The Scarlet Letter*, Hester Prynne is forced to wear an "ignominious mark," a scarlet "A," as a punishment for her sin of adultery. Perhaps the mark "P," for physical, would be a similar disgrace to the mental. Not unlike the prospect of neurobiological and evolutionary explanations of why we have certain moral intuitions,[9] problems of downward causation and reductionism *do* appear disturbing from certain points of view.

I mention these worries because they make the reductionism question more engaging and exciting. They constitute possible motivations for defending antireductionism and downward causation. But I want to note that what weight they carry is highly sensitive to how one thinks of reductionism, and what kind of reductions may be forthcoming. If feasible, a *conservative reduction*, where the mental is actually identified with something physical, would not eliminate features like intentionality. Intentionality would be real, but physical. The mental would also be special in one sense, since only physical objects of some complexity would have it. Eliminative reductions, on the other hand, where features like intentionality are thrown away, *would* be threatening. And judging from historical cases of scientific reductions, there is arguably a spectrum of partly revisionary reductions in between these extremes. (Bickle, 1998, ch 2; 2003, ch. 1; Churchland & Churchland, 1991; Schaffner, 1993, ch. 9) In the case of a revisionary reduction, then, psychology – the science of the mental – would to some extent be

[9] Cp., for instance, Greene (2003).

corrected by the reducing physical theory. The mental might not be *quite* what we took it to be, but even so, it might be real and physical. *If* the mental is reduced, then, it is an open – and I take it; empirical – question to what extents its distinctive features will be conserved. Personally, I find great comfort in a remark made by neuroscientist Eric Kandel, who makes reductionism appear less disagreeable.

> For biologists working on the brain, mind loses none of its power or beauty when experimental methods are applied to human behavior. Likewise, biologists do not fear that mind will be trivialized by a reductionist analysis, which delineates the component parts and activities of the brain. On the contrary, most scientists believe that biological analysis is likely to increase our respect for the power and complexity of mind.[10]

Be that as it may, reductionism is a hotly debated issue both in the philosophy of mind and science. Attempts to motivate or arrive at reductionism by reference to problems with downward causation should therefore have a bearing on these debates, quite independently of whether we take reductionism to be a "bad thing" or not.

2.3. Exclusion Arguments and Their Presuppositions

I turn now to a more detailed diagnosis of the Exclusion Argument. But strictly speaking, there is no such thing as *the* Exclusion Argument. At least two different arguments with two different conclusions figure in the literature, the basic premises of which are formulated in different ways by different authors. Different versions of the argument are also discussed in the papers constituting this thesis.[11] When the differences do not matter, or when the context makes clear which argument is intended, I shall nevertheless speak of the Exclusion Argument in the singular. It will be useful to view the problem of exclusion as arising from the apparent inconsistency of four *prima facie* plausible

---

[10] Kandel, 2006, p. 9. Admittedly, Kandel's neuroscientific notion of reduction may differ significantly from Kim's identities. This quote is nevertheless expressive of a rather friendlier attitude that one might take towards reductionism in some of its guises.

[11] Which argument is discussed in which paper will not matter for the summaries of the papers later in this introduction.

assumptions. (Cp., e.g., Hansen, 2000; Loewer, 2002, Sturgeon, 1998) We shall later have occasion to consider *some* of the many subtleties involved in formulating these assumptions precisely. For now, loose outlines will suffice.

> (Impact) Some mental events cause physical events[12]
> (Antireductionism) Mental events are not identical with physical events
> (No Overdetermination) Physical effects like bodily movements are not generally causally overdetermined
> (Causal Closure) Any physical event that has a sufficient cause has a sufficient physical cause

Let us briefly return to the rationale for (Impact) first. As we saw the prime example of the mental's causal impact on the physical is that of actions. Setting aside subtleties from the philosophy of action, actions are bodily movements that are caused (in the appropriate way) by mental posits like intentions or belief-desire pairs (or something of that sort). My desire to finish my thesis in time and my belief that writing would be an efficient way of achieving that aim, cause me to write right now. (Impact), then, follows from a general causal view of actions, and I shall assume that *some* such view is viable throughout the thesis.[13]

Second, (Antireductionism) about the mental can be motivated *inter alia* by reference to the special, and *prima facie non-physical*, features of the mental mentioned

---

[12] The terms "impact" or "impact of the mental" are due to Sturgeon (1998). I might have called this premise "(Downward Causation)" to make clear that this is at stake for antireductionists. But as mentioned above, if the reductionist conclusion is brought about by an Exclusion Argument, the causal impact of the mental on the physical might not happily be described as "downward."

[13] Possible examples of mental causation that do not appear to involve actions might be "embarrassment causes blushing," "psychological stress causes gastric ulcer" etc. Note that the Exclusion Argument will not apply to mental events that lack physical effects. Kim, however, argues that even mental events that cause other mental events in cognitive-*cum*-causal processes must do so by causing some physical events. (Kim, 1998, ch. 2; 2005, ch. 2) According to Kim, then, mental-to-mental causation too presupposes the viability of downward causation.

above. Standard arguments *against* reductionism, like the "multiple realization argument" invoked by Fodor (1974) and others add to the evidence for (Antireductionism). I will not enter into detailed discussions of such antireductionist arguments in this thesis, except to note that reductionists have responded to them in various ways. (See Bickle (2001) for a summary.) Also, as mentioned above, reductionism, and by implication; (Antireductionism), can be cashed out in many ways. Since Kim in many ways is the main exclusionist I consider in this thesis, I shall for present purposes follow him in understanding mind-body reductionism in terms of mind-body identities. Note, though, that Exclusion Arguments could be reformulated in terms of a revisionist model for reduction. In such a case, the mental event would be replaced by a physical event, rather than conservatively identified with it. But as emphasized above, a revisionary reduction need not involve the elimination of all features of the mental. It is my impression that the possibility of a partly conservative, partly revisionary reduction of the mental has not been given the attention it deserves in the mental causation debate.

My focus will be primarily on the status of the two final premises, namely (No Overdetermination), and (Causal Closure). The basic idea behind (No Overdetermination) can be brought out by considering standard examples of overdetermination. The death of a condemned soldier is caused by the shots of two members of the firing squad, each of which would alone be sufficient to cause his death. Bodily movements are not supposed to be overdetermined like that, at least not typically. We shall return to the rationale for dismissing overdetermination. For now, it suffices to note that philosophers have found overdetermination objectionable because it is odd, or because it appears to make the mental cause dispensable. (Kim, 1998, pp. 44-45)

The formulation of (Causal Closure) – henceforth "(Closure)" – also involves many subtleties, but again the idea is simple. To assent to (Closure) is to credit the physical domain with a radical causal self-sufficiency. It will never be necessary to look outside the physical domain to find sufficient causes of physical effects. (Kim, 2005, p. 16) In contrast, there are mental events, like perceptions, that lack sufficient mental causes, so the mental is not causally self-sufficient. But note that some physical events may have mental causes *in addition* to their sufficient physical causes. In other words, (Closure) does not render (No Overdetermination) redundant.

Such are the building blocks of Exclusion Arguments. We can now see in outline how one Exclusion Argument generates a conflict between (Antireductionism) and the remaining three premises. Start out by picking an arbitrary mental event that has a physical effect (say, a bodily movement) in accordance with (Impact). By (Closure) this physical effect must also have a sufficient physical cause. But by (No Overdetermination) there cannot be any additional causes that are distinct from the physical cause. Accordingly, the mental cause must be *reduced to* the physical cause, contrary to (Antireductionism). This type of Exclusion Argument is discussed by, *inter alia*, Papineau (2001) and Sturgeon (1998). For reasons that will become apparent shortly, I dub this the "Simple Argument."

Kim (1998, ch. 2; 2005, ch. 2), however, does not assume that reductionism is possible. He only argues that *if* (Antireductionism) is true, then mental events are excluded as causes of physical events. His contention, then is that the conjunction of (Closure), (No Overdetermination) and (Antireductionism) yields (Epiphenomenalism).[14] Here (Epiphenomenalism) should be read as the negation of (Impact). What Kim attempts to show is that if we assume (Antireductionism), we end up with (Epiphenomenalism). If we do not, reductionism becomes our only option as in the above argument. All in all, Kim's argument can be viewed as posing a stark dilemma for antireductionists. Either mental events like the onset of beliefs and desires do not cause physical events like bodily movements, *or* they are reduced to physical events. For this reason I call Kim's argument the "Disjunctive Argument."

The difference between the two arguments should interest us presently, as it will contextualize what I am trying to do in this thesis. The "Simple" in "the Simple Argument" is not intended pejoratively. If sound, the argument would be powerful and

---

[14] Strictly speaking, Kim sometimes formulates the Exclusion Argument in an idiosyncratic manner. He uses the assumption of mind-body supervenience, rather than (Closure) to come up with a physical cause of effects like bodily movements. (Kim, 1998, ch. 2; 2005, ch. 2) But the (Closure) premise arguably plays a role later on in the argument. See Hansen (2000) for discussion of the role of supervenience versus (Closure) in Kim (1998). I shall, however, focus on the more conventional way of raising the exclusion problem from the four assumptions described above.

convenient indeed. As my high school mathematics teacher repeatedly pointed out, "a good mathematician knows to be indolent at the right times." Not improbably the same applies to good philosophers. My point, then, is that the argument is simple in that it arrives at mind-body reductionism without us having to go through the laborious business of actually carrying out the reduction. This would be very convenient, as one might have expected that even a plausibility argument for mind-body reductionism would require detailed investigations of the physical sciences. The Simple Argument, however, would allow us to skip such extra-philosophical excursions. Like an existence proof, it tells us that mental events must reduce to *some* physical events. (Notably without telling us *which*. That, however, might be construed as an advantage, as the mainstream view is that we are a far leap from actually being able to carry out any mind-body reductions.)

It is perhaps no surprise then, that the exclusion debate is typically – but as we shall see, not always – conducted in relative isolation from considerations of science. The premises are only rarely assessed on empirical grounds. True, (Closure) sounds like a highly empirical claim. But it is widely accepted, and if it could be accepted on the basis of the well-known explanatory successes of the physical sciences, we might not have to consider detailed evidence for it. Furthermore, we shall see that some philosophers offer relatively straight-forward arguments for the principle without appealing much to actual evidence.

On the other hand, the Disjunctive Argument strongly motivates reductionism, without guaranteeing it. It leaves reductionists with work to do. Kim notably, goes on to provide a functionalistic model for reductionism which he hopes will be sufficient to reduce all mental events, save those involving *qualia*.[15] (Kim, 1998, ch. 4; 2005, ch. 4) Other contemporary reductionists contend, e.g., that mind-body identities provide the best explanations of mind-body correlations. (McLaughlin, 2001) Others again attempt to face what has been called the "Put Up or Shut Up Challenge" from antireductionists, by arguing that reductions are actually forthcoming in contemporary neuroscience. (Bickle, 1998; 2003) Needless to say, these arguments are all controversial, and the question

[15] I explicitly set aside problems of *qualia* and consciousness in this thesis.

whether they are successful falls outside the scope of this thesis. Note though, that in contrast with a reductionism that relies on the Simple Argument *alone*, all of these arguments would have the advantage of providing at least some general information about what the mental is reduced *to*.[16] The Simple Argument can, of course, be supplemented with such considerations, but these are not required to arrive at its reductive conclusion.

I have reviewed these differences not only to show that the Exclusion Argument takes different forms, but also because a central claim of this thesis will turn out to be that there *does not appear to be any simple way to mind-body reductionism*. Even if one wants to rely solely on what I have called the Simple Argument to arrive at reductionism, more detailed investigations of explanatory practice – which is arguably our best source to questions about causation – are needed to ground the *premises* of that argument. I shall focus on problems relating to (Closure) and (No Overdetermination). Since these premises figure in the Disjunctive Argument as well, this also means that there is no simple exclusion based route to the *motivation* of reductionism.

2.4. Aims and Methods: How Should We Approach Downward Causation?

Before moving on to the details of this thesis, we should pause to appreciate some important methodological points. We have now ample reasons for *caring* about mental causation, and if we are antecedently inclined towards antireductionism we should care also about *downward* causation. But do we really have reasons to *worry*?

It might appear that the problems are easily solved or dissolved. That is, already based on this preliminary sketch, the problems may be perceived as pseudo-problems,

---

[16] Kim's model contends that functional mental properties are to be identified with the physical properties that realize or implement the function. McLaughlin's identities would identify the mental with its physical or functional correlates. These are general stories, to be sure, but they do tell us where to look for the identities. The "looking for" will be a matter of empirical investigation. Bickle's model for reductionism does not require identities, but since it draws on actual research it points to potential or actual reduction bases for mentality.

unworthy of our serious attention. Does not the success of psychological causal explanations of bodily movements speak irresistibly in favor of the causal efficacy of the mental? Does not the apparent paucity of actual psychophysical reductions – let alone the stock arguments *against* reductionism – recommend antireductionism as the default position for the nonce? Accordingly, what metaphysical reasons could possibly convince us into believing that mental events are epiphenomenal or else really physical, when our explanatory practices suggest that they are neither? Perhaps I would be wise, then, to stop writing at this point and set worries about downward causation, epiphenomenalism and reductionism aside.

This is, in fact, the way some philosophers like Tyler Burge (1989; 1993) view contemporary physicalist debates about mental causation quite generally. He thinks worries about epiphenomenalism "have an air of make-believe" to them (1993, p. 102), and contends that:

> Materialist [by which I take it he means "physicalist"] metaphysics has been given more weight than it deserves. Reflection on explanatory practice has been given too little. The metaphysical grounds that support the worries are vastly less strong than the more ordinary grounds we have for rejecting them. (Burge, 1993, p. 97)

As will become clear, I think Burge's point of view has a lot going for it, at least insofar it suggests that the premises of a sound Exclusion Argument need to be grounded in explanatory practice. This is an overall theme of my thesis. Nevertheless, I argue that his dismissal of our worries is a bit premature. We shall see that Burge's and other philosophers' practice- or science-based dismissals are highly sensitive to which practices we look at. I will hold that Burge relies on a problematic and somewhat stipulative account of what practices are relevant to understanding mental causation. In particular, he fails to appreciate the way in which neuroscience blends an explanatory interest in mentality with one in physical mechanisms. Thus, I shall argue that we can follow Burge in paying heed to explanatory practice without dismissing problems of downward causation altogether.

It is also important to notice that Burge appears to misconstrue the worries that plague physicalists. For most physicalists, problems about mental causation are questions

about *how* mental causation takes place, rather than problems about *whether* it does. (Dretske, 2003; Kim, 1998, p. 61; McLaughlin, forthcoming) The problem raised by the Exclusion Argument, for example, is best viewed as an "how"-question: Given (Closure), how is mental causation possible without overdetermination? Explanatory practice arguably tells us *that* mental causation occurs, but the "how"-question still appears to remain. Burge's position is of little help to those of us who take an interest in this question.

In the spirit of localizing our problems within a wider historical context, it is interesting to note that the distinction between the "how" and the "that" of downward causation, is also mirrored in the correspondence between Descartes and Elisabeth. In his response to Elisabeth's question, Descartes argued that causal interaction between Soul and Body could only be grasped by reference to the primitive idea of the Soul's union with the Body.[17] He suggested that Elisabeth abstain from metaphysical meditations and turn instead to sensory experience in everyday life and ordinary conversation in order to make this idea clear to herself. Substituting "explanatory practice" for "sensory experience" in Descartes' response to Elisabeth, we have in effect Burge's dismissal of problems of downward causation. Descartes may well have been right that:

> […] people who never philosophize and use only their senses have no doubt that the soul moves the body and that the body acts on the soul.[18]

But of course, this is no answer to Elisabeth's "how"-question. In fact, the princess went on to complain, much like Burge's physicalist opponents, that:

> I see also that the senses show me *that* the soul moves the body, but that they do not show me really (any more than the Understanding or the Imagination does) *the way in which it does*.[19]

---

[17] Descartes' letter to Elisabeth of June 28, 1643. Translated in Kenny, 1970, p. 141 / AT III 690.

[18] Descartes' letter to Elisabeth of June 28, 1643. Translated in Kenny, 1970, p. 141 / AT III 690.

[19] Elisabeth's letter to Descartes of July 1, 1643. Translated in Nye, 1999, p. 26, my italics. (Not reprinted in AT-edition.)

This dispute about the "how"-question, does not appear to have been resolved in the subsequent correspondence between Elisabeth and Descartes.[20]

I have emphasized this historical parallel not only to contextualize contemporary problems, but also because the metaphilosophical questions it raises are central to the aims of the present thesis. My approach in what follows is guided by three assumptions:

> (1) I agree with Burge that explanatory practice trumps metaphysical worries about epiphenomenalism. Epiphenomenalism *must* be dismissed.
> (2) I also agree with the Elisabethan objection that this still leaves the question of *how* epiphenomenalism is to be avoided open.
> (3) Provided, that is, that the "how"-problems themselves, and the premises giving rise to them, fit well with explanatory practice.

It is important to appreciate that (3) is no trifle proviso. First, reactions to the Exclusion Argument – and to a certain extent; other problems of downward causation as well – from Burge (1993), Barry Loewer (2002), E.J. Lowe (2000) and Scott Sturgeon (1998) have one thing in common. These critics argue in various ways either that the Exclusion Argument lacks support in explanatory practice or that it is in fact incompatible with explanatory practice. Accordingly, each of the four papers constituting the present thesis concerns one of these responses.

Second, *compatibility* with explanatory practice may not be enough to counter all of the responses of the above-mentioned critics. For Burge, at least, who appears to be the critic who has the least patience with the Exclusion Argument, also questions the *motivation* of its "how"-problem. At this point, then, there emerges what may be a largely overlooked and substantial disagreement between Burge and physicalists like Kim. As we shall see, Burge appears to be thinking that many of the "how"-questions of mental causation in general could turn out to be *bad* or "inappropriate" questions insofar as they are not *supported* by explanatory practice. If so, the disagreement begins before the

[20] The remainder of their correspondence is discussed by Nye (1999).

"how"-questions are even asked. For instance, with respect to the call for a mechanism in mental causation, pressed by Elisabeth, and in a modern guise by *inter alia* Fodor (1990; 1991b) and others, Burge contends that:

> I have no satisfying response to the problem of explaining a mechanism. […] What is unclear is whether the question is an *appropriate* one in the first place. Demanding that there be an account of mechanism in mind-body causation is tantamount to demanding a *physical* model for understanding such causation. It is far from obvious that such a model is *appropriate*. It is not even obvious why *any model is needed*. (Burge, 1993, p. 114, my italics)

It will become clear that Burge is skeptical to "how"-questions largely because he takes them to be attempts to understand mental causation in terms of physical causation. He is hostile to such attempts. Given the current state of science, he thinks it is far from clear that physical causation – e.g., in physical mechanisms – is relevant to understanding mental causation. In fact, he strongly suggests that *mental* causation should be understood on its own terms, that is, by reference to *psychological* explanatory practices.

Very interestingly, Descartes appears to have been of a similar mind. *Physical* causation, he contended, should be understood in terms of extended bodies and mechanisms like pushing. But *mental* causation cannot be understood in terms of such physical causation. Mental causation is, however, understandable in itself, with the aid of the idea of the Soul-Body union. This reasoning lead him to suggest that Elisabeth's call for a mechanism was misguided, because she had "[…] confounded the notion of the soul's power to act on the body with the power one body has to act on another."[21] Burge and Descartes, then, both appear to question the relevance of physical causation as a source to understanding mental causation.

What is ultimately at stake – in the Burge-physicalist debates, in the Descartes-Elisabeth correspondence and in this thesis – is therefore not only the "how"-questions of

---

[21] Descartes' letter to Elisabeth of May 21, 1643. Translated in Kenny, 1970, p. 139 / AT III 690. Note, though, that this dismissal of Elisabeth's question seems to fit rather poorly with Descartes' apparent attempt to offer a mechanism for mind-body causation involving the pineal gland. (See for instance Descartes, 1985, p. 340 / AT XI 352.)

mental causation, but whether these questions are *well motivated*. If we are to meet Burge on his own turf, as I shall attempt, then a strong focus on explanatory practice recommends itself. Not only do we need to show that the premises giving rise to problems like the Exclusion Argument are *compatible* with explanatory practice, they should also be motivated by *reference* to explanatory practice.

Put briefly, my overall aim *is to investigate whether a sound Exclusion Argument can be formulated from within explanatory practice, as it where*. As I explain in the following sections, my primary focus in this endeavor will be on (Closure) and (No Overdetermination). Returning to the dialectical situation in which we left Descartes and Elisabeth, Descartes' advice was basically: "Turn away from metaphysics to everyday experience, through which you will grasp a primitive idea which in turn will solve (or perhaps dissolve) your problem." But apparently this did not help Elisabeth much. Another way of putting my overall aim, on the other hand, is this. *Can* we *solve or defuse* our *problems of downward causation by following Burge's advice and turn, not to a primitive idea, but to explanatory practice?*

In order to avoid provoking disappointment in the reader later on I should warn her right now, that this question will *not* be fully answered by the end of the thesis. I *will* suggest that the papers in the thesis jointly lend additional support to the idea that some model of "supervenient" mental causation could be a viable solution to the exclusion problem. (Cp., e.g., Fodor (1990); Jackson & Pettit (1988)[22]; Kim (1984b)) According to supervenient causation, mental events are causally efficacious only indirectly, in virtue of the physical causes on which they supervene. But more work no doubt needs to be done on this question. On the positive side, I will, however, use considerations of explanatory practice to show that some of the important problems raised by the above-mentioned critics can be bypassed. Furthermore, my discussions will have a bearing on the nature of

[22] Jackson & Pettit's model of "programming explanation" ascribes only causal-explanatory *relevance*, and not causal *efficacy* to mental events. I have nevertheless included it under the rubric "supervenient causation" as it is in many ways similar to that of Fodor (1990) and Kim (1984b), and might usefully be *reinterpreted* as attributing causal *efficacy* to the mental.

mechanistic explanations, how the Exclusion Argument is sensitive to theories of causation and other matters of contemporary interest.

2.5. "The Physical" in Downward Causation

With all these references to explanatory practice, it is high time I said something about which practices are relevant to downward causation. A central contention of this thesis is that this should not be a matter of stipulation. Which practices are relevant to mental causation is itself an empirical question.

Notice first that folk-psychology and academic intentional psychology quickly spring to mind in this context, as these are practices that take an interest in the mental's causal impact on the physical in actions. But problems of downward causation are just as much about "the physical." Now physicalists' use of the term "physical" quickly produces in the minds of many readers a question about its definition. In many ways this philosophical reflex is entirely legitimate, and may prove as adaptive as its physiological counterparts. The question about "the physical" will in fact recur throughout the thesis. In one guise it takes the form of the infamous "Hempel's Dilemma." This dilemma contends that physicalism is either trivial or false, depending on whether "the physical" is defined by reference to an idealized future theory or by reference to contemporary theories. (See, e.g., Crane & Mellor (1990) for a statement and Melnyk (1997), Papineau (2001), Smart (1978) or Stoljar (2005) for rejoinders.) This dilemma poses a challenge to physicalism in general.

The problems I have in mind, however, are for the most part specific to the Exclusion Argument and mental causation. As we shall see, "the physical" in the mental causation debate is frequently understood as the subject matter of some branch of *physics*. However, neuroscience is arguably much more closely related to mental causation than is physics, and yet it has received relatively little attention in the philosophy of mind, and in particular in the exclusion debate. Of course, hardly anyone would claim that neuroscience – or empirical evidence in general, for that matter – is *irrelevant* to the mental causation debate. But, setting some exceptions aside, there has nevertheless been something of a paucity of explicit discussions of neuroscientific data in the philosophy of mind. In part the rationale for this disregard has been that neuroscience is allegedly yet in

its infancy. To the extent that neural terms like "C fibers" are invoked in the philosophy of mind at all, they are often intended as place holders for concepts that will be provided in some future, perhaps idealized neuroscience. (See Bickle et al. (2006) for discussion.)

At least for the purposes of debating the Exclusion Argument, I think a supplementary approach which explicitly focuses on current neuroscience rather than physics or idealized future neuroscience recommends itself. I shall argue that the problems raised by Burge (1993), Loewer (2002), Lowe (2000) and Sturgeon (1998) are naturally addressed with the aid of neuroscience rather than physics. A secondary aim of this thesis is therefore *to investigate how current neuroscience can be brought to bear on the premises of the Exclusion Argument*. To this end I will in fact formulate and discuss a "Neural Exclusion Argument" in paper (#4). This argument has the advantage of being about *current scientific attempts to relate the mental to the physical*. If neuroscience is an attempt "to link molecules to mind" (Bickle, 2003, p. 3; Kandel et al., 2000, p. 3), then a sound Neural Exclusion Argument might imply that the links currently being investigated by neuroscientists should be viewed as potential mind-body reductions. In contrast, there does not at this point appear to be any developed science like "behavioral quantum mechanics" or "psychological quantum mechanics." While interesting in its own right, an Exclusion Argument formulated in terms of physics – or at least in terms of *micro*physics – would tell us little about how we should interpret current scientific attempts to find the mind's place in nature. So if neurophilosophy is the attempt to address philosophical questions by drawing on neuroscientific theory, then this thesis is partly – but not wholly – an attempt to apply neurophilosophy to the Exclusion Argument.

There are of course other ways to approach the Exclusion Argument. The argument has predominantly been discussed as a part of the philosophy of mind, and the debate has often focused on questions about the nature of properties, events, causal relata, the modal status of its premises and so on. While I sympathize with Burge's emphasis on explanatory practice as a more reliable source to questions about mental causation than metaphysics, I should in no way be taken to dismiss metaphysics as irrelevant. Indeed there are many unresolved questions about the Exclusion Argument's metaphysical underpinnings that I might legitimately have discussed instead. I shall, nevertheless adopt a different, in some ways Burgean, approach and will as far as possible set such questions

aside. I hope that some readers will find this approach attractive, and I hope that my tentative explorations into neuroscience are sufficient to make my claims plausible.

The approach requires some quite delicate balancing. Some readers will perhaps think my discussion of scientific examples too superficial. I hope they are wrong, but I contend at least that I have described philosophically interesting features of neuroscience that are relevant to the Exclusion Argument. Others may have little patience with the discussion of empirical evidence, thinking perhaps that traditional metaphysical approaches to the Exclusion Argument will eventually prove more fruitful. To these readers I can only say that: given the lack of consensus on the relevant metaphysical matters, the seemingly inescapable appeal to more or less brute intuitions, and the extreme controversy surrounding the Exclusion Argument in general, I thought it worthwhile to try out a different – and I believe, in many ways novel – approach.

With these methodological assumptions in hand, I turn to the specific problems with (No Overdetermination) and (Closure) that will concern me in the papers.

2.6. Overdetermination and Causal Competition

The very idea of causal exclusion presupposes some way of generating a *causal competition* between mental and physical causes of events like bodily movements. Ultimately there can only be one sufficient cause. If exclusionists are right, the physical cause wins the contest, thus turning any irreducibly mental events into epiphenomena. The source of this competition is, of course, (No Overdetermination). I shall not be concerned with giving the principle a completely uncontroversial formulation, nor with providing a general answer to when (if ever) overdetermination may be acceptable. My question is rather this. Why should mental and physical events compete in the first place? Why cannot bodily movements have irreducibly mental causes in addition to their physical causes? Interestingly, the idea of such causal competition has drawn fire both from antireductive physicalists like Loewer (2002) and from antiphysicalists like Burge (1993). In a sense, both appear to endorse overdetermination. Their arguments are discussed in papers (#1) and (#2), respectively.

There are several interesting similarities between Burge's and Loewer's arguments. As we shall see, both think the idea of causal competition presupposes what is

sometimes called a "productive" view of mental causation, that is, roughly, the view that mental causes must literally do causal *work* to produce their effects. Both take this productive view to be a piece of objectionable metaphysics that fits ill with explanatory practice. Both, then, give explanatory practice primacy over metaphysical assumptions. And both use considerations of explanatory practice and the nature of causation to dismiss the threat of exclusion as a metaphysical fiction. In spite of this, in my view, praiseworthy methodology, I argue that Burge's and Loewer's arguments are inconclusive. I do so by questioning their interpretations of explanatory practice. Even though papers (#1) and (#2) were not primarily written with this aim in mind, they may when considered together be taken to support some model of supervenient causation as a solution to the exclusion problem. I sketch this possible consequence in the conclusions following the synopsis of the two papers.

2.6.1. Synopsis of Paper (#1)

Loewer contends that "Kim is thinking of causation as a relation in which the cause generates or produces the effect." (2002, p. 658) He correctly points out that it is hard to see what Kim's talk of productive causation amounts to, but nevertheless grants that overdetermination appears objectionable if causation is thought of in this way. However, he dismisses causation as production on the grounds that it: (a) is incompatible with modern physics, insofar as nothing short of a cross-section of an event's past light cone will be sufficient to produce it. (b) Involves a commitment to "indigestible metaphysics" (2002, p. 661), insofar as productive causal relations would fail to supervene on the fundamental physical facts and laws of the world. Finally, he proposes an alternative, counterfactual theory of causation, which he thinks makes overdetermination innocuous.

In this paper I offer an account of productive causation that is compatible with science, does not involve metaphysical commitments of the sort Loewer finds objectionable, but that nevertheless can be used to rule out at least some kinds of overdetermination. What, then, is productive causation? Metaphorically it is sometimes said that causes must be "biffy" or have a kind of "oomph" to them. Talk of causal "powers" and of causes doing "work" to produce their effects also looms large in the philosophy of mind. Presumably this kind of talk is part of what makes philosophers like

Loewer associate production with heavy-duty metaphysics and medieval notions of causation. Ted Sider, for instance, makes a nice parody of one production-like way of dismissing overdetermination:

> Causation is a kind of fluid divided among the potential causes of an effect. If one potential cause acts to produce an effect, that fluid is used up, and no other potential cause can act. (Sider, 2003, p. 721)

However, there might be another, more deflationist, way of interpreting the production-talk. A central assumption of cognitive linguistics is that the way people talk about a phenomenon like causation can reveal how they think about or construe that phenomenon. When Kim and others talk of causes in terms of work, powers or forces, it suggests that they also *reason* about causes as they reason about work, powers or forces. Given their way of conceptualizing causation, then, it is perhaps small wonder that they take the idea of more than one sufficient cause to be odd at best. For implicit in work-talk, for instance, is the idea that acts of work "add up" to yield a product. Once that work is done, there is nothing left to do. Indeed, doing more work might yield a different product. If causes are analogous to acts of work, overdetermination appears almost incoherent. This interpretation, then, makes transparent at least one reason for thinking overdetermination is objectionable.

The big question is of course whether causation can *legitimately* be conceptualized in terms of work and the like, without committal to causation *being* work in any robust metaphysical sense. I think it can. That is, productive causation can be dissociated from heavy-duty metaphysics. The idea that causes "add up" like acts of work can be captured by ontologically innocent "productive constraints" or "principles of causal combination." Such principles play important roles in causal reasoning, and can be given sober, even mathematical forms. I suggest a theory of causation be counted as productive if it puts such productive constraints on its causes.

I use this understanding of production to counter Loewer's arguments. First, it is far from clear whether this notion of production involves any failure of supervenience. Second, I argue that so-called "empirical" theories of causation like the "transference theory" count as productive in the present sense. And yet these theories are *designed* to

be compatible with physics. Third, whatever the status of such empirical theories, productive causation is an important part of many mechanistic explanations. Such explanations are widespread in the special sciences, in particular in neuroscience. So even if physics turns out not to involve productive causation, an important part of modern science does. Put briefly, I think of mechanisms as assemblies of causal influences that combine to yield an effect in accordance with productive constraints. I illustrate this idea with examples of neural mechanisms and show how they can give rise to exclusion based reasoning. In a slogan; "mechanisms don't overdetermine their effects."

But this, of course, takes us right back to the Elisabeth-Descartes dispute. Why should *mental* causes be understood as mechanistic-*cum*-productive causes? Even though some physical causes are productive, mental causes may not be. For all I have said some model of supervenient causation where mental causes supervene on physical causes, that presumably do the producing, may still be viable.

It might therefore appear that this paper does not show much. Kim will insist that mental causes must be productive, and Loewer will deny this. Deadlock. But this description is misleading for two reasons. (1) Loewer's arguments against production were entirely general. Mental causes are not productive because *no* causes are productive. But if I am right, there is nothing wrong with productive causation *per se*. The argumentative resources invoked by Loewer (considerations of physics and worries about the digestibility of metaphysics) therefore appear insufficient to answer the following, more specific question. Is there anything special about *mental* causes that require them to be productive? As we shall see in paper (#2), Burge suggests there is not. Accordingly, this "negative" result of paper (#1) points in the direction of what may be a fruitful approach to the question of productive mental causation.

(2) Then there are positive results. Various notions of productive causation have attracted interest in the philosophy of causation recently. My account of production as compatible with modern science is therefore interesting as a contribution to ongoing debates. Similarly, mechanisms are a hotly debated subject in contemporary philosophy of science and neuroscience. My discussions of causal combination in mechanisms therefore concerns the nature of a highly important scientific posit. Finally, Sider (2003) contends that there has been something of a paucity of convincing arguments for (No

Overdetermination) principles. In this respect, the idea of productive constraints provides at least one more explicit reason for finding some forms of overdetermination objectionable.

2.6.2. Synopsis Paper (#2)

As discussed above, Burge (1993) has little patience with physicalist worries about mental and downward causation. He takes these worries to arise because philosophers are misguided by physicalist metaphysics and fail to appreciate the essential role of explanatory practice in determining which events are causally efficacious. While Burge's paper has been quoted often enough, I know of no systematic account of his reasons for dismissing these worries. This paper's primary contribution is to offer such an account, by tracing the disagreement between Burge and physicalists to Burge's own notion of causal powers. I describe how Burge uses his notion of causal powers to dismiss physicalist constraints on causal efficacy in psychology – e.g., the call for a mechanism in mental causation – on the grounds that it is unclear whether these constraints can be motivated from *within* psychological explanatory practice. Such constraints may, Burge seems to acknowledge, be appropriate in the physical sciences. But philosophers cannot take constraints from these sciences and apply them, at their own discretion, to the very different explanatory practice of intentional-psychology. Much like Descartes, Burge thinks than mental causation can and should be understood on its own without reference to physical causation. Problems for *mental* causation, then, must be raised from within *psychological* explanatory practice. This is important, because it shows that Burge may not be utterly insensitive to the "how"-"that" distinction in problems of mental causation. He is in fact, as I urged above, apparently skeptical about the *motivation* of at least some of the "how"-questions.

This paper covers a lot of material, and to the extent that I agree with Burge, it constitutes the methodological backbone of my thesis. In this introduction I limit my focus to the paper's implications about (No Overdetermination), productive causation and mechanisms. Though he finds the description "overdetermination" misleading, Burge happily endorses the idea of bodily movements being the outcome of two "patterns of events" – one physiological and one mental. He finds the idea that these patterns would

exclude or compete with one another "perverse" (1993, p. 116). Like Loewer, Burge thinks the idea of exclusion depends on what I call a productive view, or on "thinking of mental causes on a physical model–as providing an extra 'bump' on the effect." (1993, p. 115) However, unlike Loewer, he does not object to the productive view *per se*, but contends that it is highly unclear whether a *physical* view of causation should apply to *mental* causes. We have seen that he is also skeptical about the demand for a mechanism in mental causation for the very same reason, that is, because demanding a mechanism

> […] is tantamount to demanding a physical model for understanding such causation. It is far from obvious that such a model is appropriate. It is not even obvious why any model is needed. (Burge, 1993, p. 114)

The demand that mental causes be productive and the related call for a mechanism in mental causation are, then, both dismissed on the grounds that their motivation stems from practices that are *external* to psychology.

This brings us to my disagreement with Burge. Burge's dismissal depends on his way of individuating the relevant explanatory practices, in particular on his treating psychology as highly independent of and autonomous in relation to neuroscience. He tells us that there is no causal competition, because psychology and physical sciences like neuroscience explain

> […] the same physical effect [i.e., a bodily movement] as the outcome of two *very different* patterns of events. The explanations of these patterns answer two *very different* types of enquiry. Neither type of explanation makes essential, specific assumptions about the other. (Burge, 1993, p. 116, my italics)

Burge may be thinking primarily of *folk*-psychology – rather than academic or cognitive psychology – as the practice that describes the mental pattern of events leading up to bodily movements. If so, he may be right that the need for mechanisms and productive mental causation cannot be motivated from within (that kind of) psychology. Folk-psychology arguably takes an interest in describing bodily movements as *behaviors*, or in bodily movements as the more or less rational outcome of a pattern of event

involving beliefs, desires, deliberations etc. But it does not appear to involve any commitment to production or mechanisms. What I want to urge, however, is that identifying which practices are relevant to understanding mental causation is itself an *empirical* question. Folk-psychology – or academic intentional psychology for that matter – need not be the only practice that can serve our explanatory interest in seeing bodily movements as behaviors.

We can appreciate this by considering what Burge's "neural" pattern of events leading up to bodily movements may look like. I argue that various branches of neuroscience is characterized by (i) an explanatory interest in the productive mechanisms that lead up to bodily movements. Nevertheless, this interest is combined with (ii) an interest in seeing the movements as *behaviors*, for instance as controlled by representations and deliberations. By also taking an interest in mechanisms, neuroscience does not *abandon* its psychological or representational perspective any more than cognitive science in general does. Burge's attempt to mutually insulate the explanatory aims of psychology and mechanistic neuroscience depends on his individuating these practices as strongly independent and autonomous. But this individuation fits ill with the nature of the neuroscientific enterprise. I support this claim by adopting something of a Burgean strategy. I sketch what kinds of causal explanation various branches of neuroscience appear to give of bodily movements. Furthermore, to emphasize the explanatory interest in mentality and rationality I also consider Patricia Goldman-Rakic' theory of prefrontal cortex in some detail, as this is a region thought to be involved in many higher cognitive functions. This excursion into neuroscience also supports the claim that, *pace* Burge, some tight relation between mental and neural causes is needed to ensure mental causation.

2.6.3. Conclusions

Must mental causes be *productive* causes? Loewer and Burge both argue plausibly that (No Overdetermination) depends on this question being answered in the positive. But they both suggest that it should in fact be answered in the negative. I argue that Loewer's general argument against production fails. There is productive causation in science. On the other hand, Burge's specific argument that mental causes need not be productive

poses a stronger threat to the (No Overdetermination) and the Exclusion Argument. It is indeed hard to see what features of the mental should require mental causes *themselves* to be productive. However, Burge goes too far in dismissing the relevance of productive mechanisms to mental causation. Neuroscience combines an interest in mental causation with one in productive mechanisms. The question "how do mental events cause physical events?" arises from considerations of neuroscience and requires an answer involving productive mechanisms for its answer. This motivates a "synthesis" of the productive view and its non-productive "anti-thesis." *Pace* Loewer, there is nothing wrong with productive causation, *pace* Kim, mental causation is not productive.

Now, if there is little reason to believe that mental causes must be productive, but there is reason to believe that mental causation at the very least *depends* on productive neural mechanisms, some model of supervenient causation appears to recommend itself quite naturally. The idea that mental causes are only causally efficacious in virtue of underlying neural mechanisms might be the most plausible option available to antireductionists. But perhaps "supervenient causation" is nothing but a fancy word for epiphenomenalism? This, of course, is a worrisome question. But it appears to lose much of its force if Burge is right that not all causation must be productive causation. Part of the reason for thinking that supervenient causes are really epiphenomena is arguably the idea that the underlying physical causes do all the causal work in producing the effect. Accordingly, the problem is at least partly that supervenient causes are rendered non-productive. But as we saw Burge's position can potentially be used to argue that the call for productive mental causes is not well motivated. The intuition that mental causes *must* be productive will perhaps remain, but it is not clear that productive mental causation is something that can be rationally wanted. If mental events are freed from the demand that they be productive, some counterfactual theories of causation like Loewer's may be sufficient to ensure their causal, albeit non-productive, efficacy.

All in all, considerations of the production debate suggests that something like supervenient causation might be the best available antireductionist account of how mental causation is possible in a physical world.

2.7. Causal Closure and Physical Causes of Bodily Movements

Judging from papers (#1) and (#2) it is far from clear that mental and physical causes compete, as demanded by (No Overdetermination). But there is also another premise of Exclusion Arguments with which one might take issue. Such arguments require not only the idea of causal competition, but also the presence of sufficient physical causes with which mental causes may compete. For exclusion to take place, it must be established that non-mental causes alone could bring about effects like bodily movements. Most Exclusion Arguments include some variety of (Closure) for this purpose. In one guise or another, (Closure) is widely accepted by physicalists. Much like (Impact), this principle is rarely questioned in debates about the Exclusion Argument. Indeed, (Closure) is often set forth without argument, or with brief arguments that do not appeal much to empirical evidence. For instance, Kim (1992) contends that unless (Closure) holds, the laws of physics will be violated, a consequence which he takes to be intolerable. However, as McLaughlin (1992) and David Papineau (2001) argue, such simple cases for (Closure) appear to fail. These authors also suggest that what they take to be convincing empirical evidence for the principle emerged only recently. (Closure), then, is presumably a deeply empirical truth, if a truth at all. My two papers on (Closure) underscore this message, which is in accordance with the overall aims of the present project, viz. to investigate the empirical plausibility of the Exclusion Argument.

In spite of widespread physicalist agreement about (Closure), the principle and its use in the Exclusion Argument may not be as unproblematic as it seems, however. First, the claim that (Closure) is empirically plausible has been questioned. (See, e.g., Hendry (2005); Sturgeon (1998)) Second, some authors, like Nancy Cartwright (1999), think of scientific theories as models with limited scope, and are skeptical to claims about in-principle causal-explanatory completeness of theories. My papers will have a bearing on these problems, but my primary concern is with a third kind of problem raised by Lowe (2000) and Sturgeon (1998). They argue in different ways that (Closure) cannot figure in a sound Exclusion Argument, even *if* there is sufficient empirical evidence in favor of (Closure). Lowe and Sturgeon, then, question the *use* of (Closure) in Exclusion Arguments, so their arguments do not depend on actually dismissing (Closure). This

makes the arguments potentially very powerful. However, in papers (#3) and (#4), I show how the problems they raise can be bypassed.

2.7.1. Synopsis Paper (#3)

Lowe's arguments (2000) are interesting in their own right, but also because they raise important metaphilosophical questions about the role of empirical evidence in the mental causation debate in general, and the debate between dualists and physicalists in particular. He considers various versions of (Closure) and argues that they are all compatible with outright *dualist* scenarios involving what I call "invisible mental causation." The simplest scenario involves a physical effect, E, say, a bodily movement, which has a sufficient physical cause, P, in accordance with (Closure). However, P brings about E, at least in part, by causing an intermediate, but irreducibly mental event M which in turn *helps* bring about the effect. This is not a case of overdetermination, as M is a necessary part of the causal processes leading from P to E. Importantly, Lowe contends that the causal contribution of the mental event would be *invisible* from the physical point of view. Accordingly, we would have the *semblance* of physical causal processes that account completely for physical effects, but the processes would in fact be partly and irreducibly mental. Hence, Exclusion Arguments would have no bite on invisible mental causes.

Lowe argues first that scenarios of invisible mental causation are *possible*, and second, that there are theoretical or metaphysical reasons for believing in their reality. Very interestingly, he contends that this kind of dualism cannot be dismissed on *empirical* grounds. I argue that he is wrong, and offer empirical reasons for dismissing invisible mental causes.

First, I argue that the mere possibility of invisible mental causation does not cut any philosophical ice. Neuroscientists enjoy an *a posteriori* entitlement to believing that effects like bodily movements can in *normal circumstances* be fully accounted for by causal processes that are neural through and through. While the "normal circumstances" include many non-neurobiological factors, what we know about them does not suggest that they include invisible mental causes. So as far as the possibility argument is concerned, invisible mental causes can be dismissed *inter alia* on the grounds that their

postulation is *ad hoc*. (Notably, this dismissal does not depend on a general (Closure) principle of the physical domain, though it is compatible with such a principle.)

But the *ad hoc* postulation could be perfectly acceptable if there were theoretical reasons for believing in mental causation. Lowe actually suggests that there is an intriguing binding problem of behavior, analogous to the binding problem in perception studies. He suggests that this problem might require invisible mental causes for its solution. It is, unfortunately, a little hard to see what this binding problem consists in, but it suggests that neural processes leading up to voluntary bodily movements must somehow be "integrated" or "organized." Briefly, Lowe worries that in the absence of invisible mental causation neural causes would render bodily movements "coincidental." Lowe takes an effect to be coincidental if its "immediate causes are the ultimate effects of independent causal chains." (Lowe, 2000, p. 579) He seems to imply that chains of neural causes leading up to bodily movements are somehow too chaotic and independent, and thus render bodily movements coincidental. Thus, he argues that the chains could be bound by invisible mental causes that link the otherwise independent causal chains. I accept this theoretical motivation for the sake of argument, though it is in my view not sufficiently developed to motivate invisible mental causation.

However, when Lowe says that invisible mental causation cannot be dismissed on empirical grounds he fails to appreciate that his argument depends on an *empirical* premise: namely, that neural causal chains do in fact render bodily movements coincidental. Lowe does not really argue for this, but suggests that neural causation takes place in what he calls a "neural maze." (Lowe, 2000, p. 581) This argumentative failure is doubly instructive. First, it shows that empirical evidence *can* in principle be used to answer what Lowe takes to be a purely philosophical question. There is nothing in Lowe's definition of non-coincidental effects that suggests that they could not be brought about by neurophysiological causes. Second, Lowe would have to undertake more detailed investigations of neural causation than he actually does to argue for his claim. So Lowe's argument is inconclusive, even if we accept his theoretical motivation for invisible mental causation. This is the main conclusion of the paper.

Finally, and somewhat more speculatively, I offer reasons for thinking that Lowe's empirical assumption may be false. I suggest that Lowe's characterization of

neural causation involves a serious, but characteristic, misinterpretation of the neuroscientific enterprise. Thus it is far from clear that neural causes render bodily movements coincidental in Lowe's technical sense. A variety of experimental techniques have allowed neuroscientists to discover considerable structure in the nervous system, and to reveal how bodily movements are the outcome of a carefully orchestrated interplay between neural subsystems. Neural causes, then, do not necessarily appear to render bodily movements coincidental. At the very least, given my arguments Lowe needs to make more explicit the job he pictures for invisible mental causes, and argue that neural causes are disqualified for that job. That, presumably, will only be possible by considering neuroscience in more detail.

2.7.2. Synopsis Paper (#4)

Sturgeon's (1998) aim is to show that the Exclusion Argument is not supported by current scientific knowledge. His line of attack is to consider what the physical domain referred to in (Closure) and (Impact) – i.e., the premise stating that mental events have physical effects – is. He suggests that "the physical" could mean the broadly physical domain consisting of macroevents like cars colliding and arms moving. Or it could mean the narrowly physical or microphysical domain, which he equates with the domain of quantum mechanics. First, Sturgeon argues that the Exclusion Argument equivocates between these senses of the physical. He thinks that (Closure) is plausible, if at all, only for the quantum mechanical domain. On the other hand, he takes it that (Impact) is plausible only for the broadly physical domain, since the claim that mental events cause quantum mechanical events is not part of explanatory practice. Hence, the only plausible (Closure) and (Impact) premises involve different readings of the physical, making the Exclusion Argument invalid. Second, Sturgeon uses various interpretations of the so-called "measurement problem" in quantum mechanics to argue that (Closure) may fail even for the quantum mechanical domain. So there may be *no* (Closure) principle to be used in the Exclusion Argument.

Sturgeon considers ways of amending this equivocation problem under the assumption that physical events are composed of – i.e., have as their parts – quantum mechanical events. Consider, for instance, a bodily movement which is composed of

narrowly physical events. If (Closure) holds for the narrowly physical domain these events would have sufficient narrowly physical causes. In addition, it might seem that if you cause the parts constituting an event, you *eo ipso* cause that event. But then, narrowly physical causes would also be sufficient causes of the bodily movements. If so, bodily movements would have sufficient (narrowly) physical causes with which mental causes could compete, after all.

Sturgeon, however, denies that causal influence can flow from the micro to the macro so easily. He does so by arguing for a (Cause & Essence) principle, according to which to cause an event is to bring about the essence of that event. Furthermore, he appeals to special features of quantum mechanical reality like "superposition" and "projection" to argue that it is far from clear that quantum mechanical events are essential to broadly physical events. We cannot tell whether they are, he contends, because there is a huge conceptual gap between the quantum mechanical and the broadly physical. But if the quantum mechanical is not essential to the broadly physical, then, according to the (Cause & Essence) principle, causes of quantum mechanical events are not causes of broadly physical events. So, once again, there would be no sufficient physical causes with which mental causes might compete.

Sturgeon's challenge is interesting for several reasons. It raises important questions about which, if any, physical domain is closed. As it turns out, some physicalists do appeal to a (Closure) principle for the broadly physical domain, whereas others appeal to the narrowly physical. The confusing variety of (Closure) principles at play in the literature will have to be sorted out by considering the empirical plausibility of (Closure) for various physical domains. Second, Sturgeon illustrates that formulating the Exclusion Argument in terms of *physics* can lead to difficult metaphysical questions about part-whole relations and causation, as well as equally difficult questions from the philosophy of physics. These problems are interesting and may be resolvable. But Sturgeon is surely right to argue that microphysical causation is conceptually remote from mental causation. This motivates a shift of attention to a physical domain that bears a closer relation to mental causation. I therefore show how Sturgeon's challenge can be bypassed by formulating exclusion in terms of neuroscience rather than physics properly speaking.

This may appear ill-advised. The neurobiological domain is rather obviously *not* closed in the absolute sense of (Closure). It suffers causal input from the outside in phenomena like perception, and, furthermore, neural causation depends on a variety of non-neurobiological background conditions. However, the Exclusion Argument does not, as Sturgeon and many others appear to be implicitly assuming, depend on a general (Closure) principle, which provides us with absolutely sufficient causes from any particular physical domain. Rather than appeal to (Closure) principles of this sort, I expand on the idea I invoked against Lowe, and argue for a neurobiological "(quasi-Closure)." Setting aside causal input to the organism, and letting circumstances be "normal" for neural causation, neural events have neural causes that are sufficient in the circumstances. Given the additional assumption that there are no irreducibly mental factors lurking in a minimal characterization of the "normal" circumstances, this principle is apt for figuring in the Neural Exclusion Argument I formulate.

This may sound trivial, but it is not. That is, the argument is not an off-hand appeal to in-principle neural explanations of bodily movements, because its plausibility derives from actual and detailed models of neural causation. These provide us with a good, and partly quantitative, theoretical grasp of: (i) the "connectionist" structure within which neural causation is situated, (ii) what kinds of events are causally relevant within that structure and (iii) the cellular/molecular mechanisms through which such causes must work. The theoretical picture provides us with a good model of neural causation at the cellular/molecular level in general, and of bodily movements in particular. Many background conditions are needed for the picture to do its explanatory work, of course. And as is always the case in the special sciences, these conditions are not fully specifiable. Furthermore, they are of a theoretically quite *heterogeneous* nature. Hence, the appeal to (quasi-Closure) does not depend on the idea of a particular, privileged theory being causally complete or closed. Nevertheless, I argue that the success of the theoretical picture, and what we do know about the background conditions, entitles us to believe that no irreducibly mental causes figure in the background conditions.

However, that irreducibly mental causes are not needed has not always been clear. We have not always had good neural models of neural causation. Until quite recently, the theoretical picture of neural causation was rather crude and detailed knowledge of the

relevant causes and mechanisms were absent. Indeed, early versions of the picture included things like irreducibly mental or vital forces as putative causes. To show that the present picture is something of a historical contingency I adopt a historical perspective. I sketch how evidence for something like "quasi-Closure" emerged gradually along with the acceptance of central theoretical assumptions like Ramón y Cajal's "Neuron Doctrine" and discoveries about the nature of neural signaling. I tentatively suggest that these discoveries can, in retrospect, at least, be viewed as the gradual causal exclusion of putatively necessary vital or mental causes of neural events.

The idea that we can appeal to neural causes in an Exclusion Argument is therefore not trivial or obvious. But perhaps it amounts to explanatory *hubris*? It would be naïve to assume that the theoretical picture of neural causation may not be revolutionized in certain ways, and indeed I mention ways in which it may *currently* be changing. However, some theoretical doors do shut as science moves forward, and I argue that forthcoming changes to the picture are not likely to include the reentry of irreducibly mental or vital causes.

Anyway, if sound, my argument shows that the problems Sturgeon raises can be bypassed. (1) Whatever the status of general (Closure) principles for the broadly or the narrowly physical domain, exclusionists can use neuroscience to point to causes with which mental causes will – provided (No Overdetermination) is also accepted – compete. (2) Sturgeon's considerations of part-whole relations and the (Cause & Essence) principle have no bite on the Neural Exclusion Argument. For neural events cause contractions of muscle fibers and – perhaps unlike quantum mechanical events – such contractions compose into bodily movements in a well-understood way. There is no problem with saying that a minimal number of fiber contractions are essential to a bodily movement. Finally, (3) my argument has the advantage of appealing to *detailed* and *extant* scientific models. Thus it avoids worries about in-principle explanations harbored by philosophers of science like Cartwright (1999).

2.7.3. Conclusions

I think the conclusions with respect to (Closure) are more clear-cut than those involving (No Overdetermination). In accordance with the basic aims of the thesis I have

emphasized the deeply empirical nature of (Closure) principles. I have argued that the problems raised by Lowe and Sturgeon can be addressed empirically. In particular, by appealing to neuroscience rather than physics, I have argued that a Neural Exclusion Argument can avoid these problems. This argument does not depend on a general (Closure) principle, as most Exclusion Arguments do. While the idea of formulating exclusion without (Closure) and with reference to neural causes is not entirely novel – see, e.g., Kim (2005, ch. 6)[23] – it has received relatively little attention. My attempt to work out the details of such an argument and present what evidence there is for it should therefore be of value to the exclusion debate. Whatever problems there are with Exclusion Arguments, they do not appear to involve our inability to come up with physical causes with which mental causes may compete. This, of course, is what most physicalists think anyway, but notably, sound reasons for thinking so are far more empirical in nature than one might have expected, given the above-mentioned, relatively simple arguments for the principle.

2.8. Concluding Remarks on the Prospects of Exclusion Arguments

The four papers constituting this thesis all concern, among other things, Exclusion Arguments. I will not pretend to have settled the status of such arguments once and for all. But I think their prospects are rather bleak. What I have called the Simple Argument promises an easy route from just three premises to mind-body reductionism. It solves the mind-body problem with a pen stroke. The Disjunctive Argument would provide an equally convenient and simple motivation for mind-body reductionism, as the alternative of epiphenomenalism is almost too horrible to be contemplated. A central contention of this thesis is that this appearance of simplicity is misleading. *There is neither a simple exclusion-based route to reductionism nor to its motivation.*

This will hardly surprise anyone who has followed the exclusion debate, for judging from that debate the exclusionist's path appears to be cluttered with metaphysical

[23] Note that this argument of Kim's differs from the version of the Exclusion Argument in which he starts from supervenience, rather than (Closure). In the latter argument, mentioned in a footnote above, (Closure) arguably plays an important role, in the former it does not.

obstacles. But the apparent simplicity I have in mind is of a different kind. As I urged earlier, the arguments would also be convenient because their premises seem to be assessable without much explicit consideration of explanatory practice and empirical evidence. It is this appearance of simplicity that I hope to have shown to be misleading. The plausibility of (Closure) and (No Overdetermination) does in fact hinge crucially on features of explanatory practice and empirical evidence. In this respect I have urged *that a focus on neuroscience in connection with Exclusion Arguments is both natural and fruitful*.

With respect to (Closure), the problems seem avoidable. One way of avoiding them is to appeal to neuroscience. This is possible because *Exclusion Arguments do not depend on a general (Closure) principle*. Focusing on neuroscience has the further advantage of *relating causal exclusion to* current *scientific attempts at understanding the mind's place in nature*.

With respect to (No Overdetermination) the situation is less clear-cut. We have little reason to believe that mental causes must be productive as that premise arguably requires. *The idea that mental and physical causes compete does indeed appear to be deeply problematic*. Furthermore, considerations of productive causation may be taken to lend support to *supervenient causation as a viable antireductionist response to Exclusion Arguments*. But importantly, that is not to say that productive causation and the idea of causal exclusion are inherently problematic. I have argued that *we do find productive causation in neuroscientific mechanisms*, and that the demise of vital and mental causes in the theory of neural signaling may perhaps be viewed as *a case of causal exclusion in scientific practice*.

Finally, the thesis tells us something metaphilosophically interesting about the nature of problems of downward causation and epiphenomenalism. Practically everybody will take for granted that mental causation takes place. Some of us are interested in finding out "how." Burgean "that" answers are of no interest to us. Nevertheless, there is a lesson to be learnt from Burge. *We should take care to formulate our "how"-questions so that they are well motivated given explanatory practice*.

3. Paper (#1): Mechanisms Do Not Overdetermine Their Effects

ABSTRACT: The upshot of Jaegwon Kim's Exclusion Argument is that antireductionism about the mental leads to epiphenomenalism about the mental. Roughly, since physical events like bodily movements must have physical causes, and physical events are not causally overdetermined, there is no room for additional, non-reducible mental causes. In response, Barry Loewer has claimed that a "productive" view of causation is required to rule out overdetermination and that such causation is both metaphysically objectionable and incompatible with modern science. I show that Loewer's arguments are inconclusive. Furthermore, I offer an account of productive causation which only commits us to ontologically innocent "principles of causal combination." Productive causation in this sense is invoked in mechanistic explanations which are of central importance in modern science, especially in neuroscience. However, it is still an open question whether *mental* causation must be productive or whether some antireductionist model of supervenient causation is viable.

Force is the causal principle of motion and rest
– Isaac Newton[24]

3.1. Introduction

If sound, Jaegwon Kim's Exclusion Argument should strongly motivate the philosophy of mind community to join forces with scientists in an attempt to reduce the mental. For Kim's take-home message is stark indeed. Antireductionism entails epiphenomenalism. Either mental events do not cause anything – which is to say that they are not very real at all – or they are identical with physical events, which some may find equally disturbing. The following is a condensed version of Kim's argument (see, e.g., his (1998) or (2005) for details):

[24] Newton (1962, p. 148), quoted in Creary (1981, p. 156n1)

(E1) *Causal Closure*: Any physical event E that has a sufficient cause occurring at t has a sufficient physical cause P occurring at t

(E2) *Antireductionism*: Mental events are not identical with physical events

(E3) *No Overdetermination*: If a physical event E has a sufficient cause P occurring at t, then it has no other distinct causes occurring at t

---

(C) *Epiphenomenalism:* Mental events do not cause physical events

Physicalism on the one hand, and intentional-psychological practices on the other, compel us to postulate physical[25] and mental causes of bodily movements, respectively. But (E3) generates a competition between these causes, and, alas, it only allows for one winner. (E1) guarantees the presence of a physical cause, and (E2) rules out identifying a possible mental cause with this cause. Hence, the physical cause wins the contest, turning mental events into epiphenomena; which is well nigh intolerable.[26]

The precise formulation and justification of these premises has been debated extensively in the literature. Here I focus on the No Overdetermination Principle, or the Exclusion Principle as it is sometimes called. My concern is not with the possibility of a completely uncontroversial formulation, nor with providing conditions that determine when (if ever) overdetermination is acceptable.[27] I will for the most part set these general

---

[25] Unless otherwise specified the physical domain should presently be understood widely as encompassing the biological domain.

[26] Kim (1998, p. 42; 2005, p. 40) thinks that antireductionism rules out even mental-to-mental causation (e.g. in cognitive-*cum*-causal processes), because that kind of causation allegedly presupposes mental-to-physical causation. Anyway, as is often remarked, the weaker conclusion that mental events cannot cause physical events would be devastating enough for antireductionism.

[27] A variety of nonequivalent formulations of such principles exist, cp. Kim (1989); Lowe (2003); Menzies (2003). Sometimes what counts as overdetermination or objectionable overdetermination is left more or less implicit. (Sturgeon (1998); Papineau (2001)) My (E3) is consonant with one of Kim's recent formulations (2005, p. 17). But see Kim (2005, p. 42) for a different formulation. There is no consensus about the conditions under which overdetermination is objectionable, see Bennett (2003) for discussion and a suggested test. Sometimes overdetermination is considered from a more metaphysical angle. Does, for example, a baseball and its parts objectionably overdetermine the breaking of a window? (Sider, 2003) I do not know whether such cases amount to overdetermination, or whether such overdetermination might be [Footnote continued on next page]

and abstract questions aside, and rely instead on more concrete considerations of actual causal explanations, in assessing overdetermination. My worry about principles like (E3) is rather this. Why should mental and physical events *compete* for the status of being causes of bodily movements? To be sure, proposed causes *sometimes* do compete. That happens, for instance, when the explanatory success of invoking one kind of cause, threatens to make the other kind of cause disreputable. Thus, the bacterial (or more recently; the viral) causes proposed to explain the Black Plague competes with, and ultimately excludes, God the punisher, who was once supposed to have brought the pandemic down upon the sinners of the time. The example is not meant to lessen religion. The point is simply that medicine has driven the explanatory practice which invoked God as a direct cause of diseases out of business. But God contrasts sharply with mental and physical events, both of which are explanatorily potent *vis-à-vis* bodily movements, and firmly engrained in healthy and successful explanatory practices. Neither of them will be excluded easily. As Burge (1989; 1993) has persuasively argued, if you want to know which events or properties are causally efficacious, your best guide is going to be which events or properties figure in healthy explanatory practices. For reasons like these people have found the idea that mental and physical causes compete misguided, or even, as Burge puts it, "perverse" (1993, p. 116). Summing up, it might seem like Kim's insistence with (E3) that we ultimately only get to keep *one* cause is over-restrictive. Accordingly, antireductionists might find it natural to respond by dismissing the culprit, i.e., the instigator behind the competition – (E3). Here is a generic way in which they might do just that: (R1) Overdetermination is only problematic under a "productive" view of causation, where causes literally do causal work to bring about their effects. (R2) But mental causes are not productive causes, hence overdetermining mental causes are, contrary to (E3), unproblematic. (C) Hence, the Exclusion Argument fails.

Interestingly, this line of reasoning has been pressed in different ways by Tyler Burge (1993), who is an avid antiphysicalist, and Barry Loewer (2001b; 2002 and forthcoming), who is an equally avid, but antireductive, physicalist. Burge's argument for

---

acceptable. But even if it is acceptable, arguments are needed to show that *mental* causes can overdetermine their effects in this way.

(R2) is roughly that *mental* causation should not be construed along the lines of physical production, hence mental causes are not productive. (See, e.g., Burge, 1993, p. 115) Loewer on the other hand argues quite generally that: (1) the productive view is incompatible with modern physics. (2) Its metaphysics are objectionable, because productive causal relations might fail to supervene on basic physical facts. In contrast, he takes overdetermination to be innocuous given a counterfactual theory of causation. Now, I think Loewer misconstrues the situation and that we need to divorce productive causation from theories of causation with major metaphysical commitments to non-supervenient causal relations and the like. When properly understood such causation only commits us to ontologically innocent principles of "causal combination" that constrain how causes add up to yield their effects. Given this understanding Loewer's arguments are inconclusive. Furthermore, we need productive causation in causal mechanisms. Since such mechanisms are important parts of the special sciences – neuroscience will be our case in point – a significant portion of modern science *does* presuppose productive causation. *Pace* Loewer, productive causation *per se* is not a problem. However, the question posed by Burge – why should *mental* causes be productive? – still remains.

3.2. Loewer's Case against Production

Loewer claims that causes must somehow "generate" or "produce" their effects if overdetermination is to be ruled out:

> Kim is thinking of causation as a relation in which the cause generates or produces the effect. I am not sure what these metaphors come to, but they suggest that in some way, E [i.e., the effect] grows out of C [i.e., the cause]. In any case, if we think of causation in this way then each of the reasons that Kim gives against overdetermination appears more convincing. (Loewer, 2001b, p. 320)

Whatever the meaning of the productive metaphors, Kim used to think that causal exclusion did not depend on heavy-duty assumptions about causation (1989; 1998, p. 67), but recently he seems to rest more of his case against overdetermination on a productive

view. (2002a; 2005, p. 17-18, p. 38n6 and forthcoming) Loewer appears to offer two arguments[28] against this view, the first of which I dub the "Argument from Science." He tells us that "causation as production fits ill with contemporary physics." (Loewer, 2002, p. 661) Now there certainly are many problems with causation in physics, see Field (2003) for a review. Loewer first appeals to Russell's claim that causal concepts are not explicitly mentioned in the fundamental laws of physics. (Loewer 2002, p. 612; Russell 1912) So presumably we do not need causal notions to do physics. Second, he argues that fundamental laws like the Schrödinger equation in quantum mechanics, (Loewer, 2001b, p. 323; 2002, p. 661), relate the total state of a system at one time to the total state at a later time. Accordingly, nothing short of the total state at some earlier time will be sufficient to produce an effect.[29] He presumably takes this to be a specific problem with the productive view, since that view allegedly cannot pick out parts of states as causes. In contrast, counterfactual theories of causation are supposed to fare better, because they can single out the events that "make a difference to E's [i.e. the effect's] occurrence." (Loewer, 2002, p. 661)

Loewer's second argument, which I call the "Metaphysical Argument," is based on worries of a broadly Lewisean kind. He worries that productive causal relations might fail to supervene on the fundamental physical facts. That is, two worlds could be physically indiscernible and yet differ in what produces what. (Loewer, 2002, p. 661) The failure of supervenience would indeed be an extra burden for the productive view, perhaps it would even amount to what Loewer calls "indigestible metaphysics" (2002, p. 661). But it is hard to see why proponents of the productive view must be saddled with this burden. Rather than argue for this, Loewer refers to Michael Tooley's theory as an

---

[28] I am not sure whether he takes the arguments to be independent.

[29] According to Loewer it would be more accurate to say that nothing short of a cross-section of an event's past light cone will be sufficient to produce it. (Loewer, 2002, p. 661n12) Events outside an event's past light cone are after all supposed to be incapable of affecting the event causally, because causal influence allegedly cannot travel faster than the speed of light. (Whether this claim about causal influence is compatible with instantaneous causation, and hence with L.G. Creary's (1981) *causal* interpretation of Newton's second law – see below – is another matter.)

example of a productive theory where supervenience fails. (Loewer, 2002, p. 661n13) The real argument, then, is presumably to be found in the examples of remote possible worlds that Tooley (1990) takes to violate supervenience. Tooley uses these examples to motivate a kind of non-supervenient necessitation relations. But while his view may be *a* productive view, it is, as we shall see, not clear that *all* productive views are committed to the failure of supervenience.

To replace the productive view Loewer proposes a revised version of David Lewis' (1986) counterfactual theory of causation, which he thinks makes overdetermination "innocuous" (Loewer, 2002, p. 661). *Prima facie* this may seem right: Why should a "physical" counterfactual dependency $\neg P \mathrel{\Box\!\!\rightarrow} \neg E$ rule out a "mental" dependency $\neg M \mathrel{\Box\!\!\rightarrow} \neg E$? Would not Kim's problem vanish if only he revised his theory of causation? But things may not be quite that simple. In addition to their familiar problems with preemption etc.[30] it is not clear that counterfactual theories are compatible with symmetrical overdetermination,[31] nor with causal closure.[32] However, these may be largely technical problems, and I shall set them aside to consider issues I take to be more pressing.

3.3. Characterizing the Productive View

I think Loewer's arguments against the productive view are inconclusive and that we need productive causation to do science. Showing this requires me to get clearer about the elusive notion of production, however. It is sometimes said that productive causation is an "intrinsic" relation, or informally that causes must have "oomph" or be "biffy" (See Hall (2004) on production; Lewis (2004) on "biff.") Kim distinguishes between

---

[30] Such problems are discussed in, e.g., Collins et al. (2004a).

[31] A standard case of symmetrical overdetermination is that of two soldiers firing their weapons at the victim at the same time where each of the bullets alone would suffice to kill the victim. In this case a counterfactual theory might imply that neither of the soldiers' shots causes the death, because the death does not depend counterfactually on either of them. (See, e.g., Collins et al. 2004b, pp. 32-33)

[32] Kim (1998, p. 45; 2005, pp. 46-50) argues that overdetermination would violate causal closure in at least some possible worlds.

"productive and generative causal processes" and "noncausal regularities" that are "parasitic on real causal processes." (Kim, 1998, p. 45) But apart from a brief reference to Salmon's (1984) process theory, Kim (1998, p. 45, p. 45n28) offers no account of what the distinction amounts to. Talk of *causal work* and *causal powers* also looms large in the debate. Here is Kim again: "Given that [the physical event] *p* has a physical cause *p**, what causal work is left for [the mental event] *m* to contribute? The physical cause therefore threatens to exclude, and preempt, the mental cause." (Kim, 1998, p. 37) In a similar vein, Jackson contends that dispositions are not causes, because "there is no causing left to be done by the relevant dispositional properties" (1996, p. 394) that has not already been done by the dispositions' categorical bases. *Prima facie*, this kind of talk *does* appear to involve major metaphysical commitments that are apparently absent in Loewer's own counterfactual theory of causation. It is therefore easy to see why Loewer associates production with "indigestible metaphysics."

But there is a better way of understanding production. In general, the way people conceptualize or talk about a domain like causation can reveal how they reason or think about that domain. (Lee, 2001) When Kim and Jackson talk about causes as forces, as powers, or in terms of work, they also *reason* about causes as they reason about forces or work. There is, as it were, a transfer of ways of reasoning from one cognitive domain (that of forces/work) to another (that of causes). (Lakoff & Johnson, 1999, ch. 11) This explains why overdetermination *does* – as Loewer (2001b, p. 320 and forthcoming) apparently acknowledges – seem objectionable when causation is construed as production. For in the domain of forces or work, talk of there-being-no-work-left-to-do has a legitimate application. As an initial, trivial and everyday illustration, imagine that you are a factory owner, paying workers to the extent to which they contribute to the daily production. If the claims made by your workers at the end of the day add up to more than the measured production, you will engage in exclusion-style reasoning and look for the wretched proletarian whose alleged toil is fictitious or "epiphenomenal." A second example involves forces in Newton's second law, $\sum\mathbf{F}=m\mathbf{a}$. For any object x, and time t, the vector sum of all forces $\mathbf{F}_1$, $\mathbf{F}_2$,…, $\mathbf{F}_i$ acting on x at t equals the product of x's mass, m, and acceleration, **a** at t. This law, then, constrains how the forces combine to yield accelerations. The $\mathbf{F}_i$'s must add up to a specific sum to yield any particular acceleration

of x, lest we get a different measurable **a**. If we add more forces we must take care not to thereby change the net force $\sum \mathbf{F}$.[33]

This illustrates how forces and acts of work are subject to principles that constrain how they add up to yield their results. I call such principles "productive constraints" or "principles of causal combination." These are strikingly similar to No Overdetermination Principles. Given a particular measure of acceleration or a particular effect there are non-trivial limits on what forces or acts of work we can postulate to explain that acceleration or effect. Kim's exclusion-based reasoning consists in applying analogous constraints to mental causation. It is almost as if Kim "measured" the effect – a bodily movement – and found that given its known or hypothesized neural causes, adding an additional mental cause would yield an effect different from the one he measured. Not only are irreducibly mental causes not needed, they would distort the effect. This should be compared to another, and spatial, metaphor sometimes used to describe exclusion. There just is no "room" for such causes. Interestingly, neuroscientist Eric Kandel implicitly portrays the history of theorizing about neural signaling from Luigi Galvani to Alan Hodgkin and Andrew Huxley as involving the gradual *exclusion* of vitalistic causes in favor of physical causes. He concludes using just these metaphors. "There was [eventually] no *need* or *room* for 'vital' forces or other phenomena that could not be explained in terms of physics and chemistry." (2006, p. 83, my italics) There being "no room left" is in fact crucial to distinguish productive causation from *joint* causation. Practically any theory will want to allow for joint causation where several causes combine to yield a sufficient cause. But a productive theory will rule out adding overdetermining causes to this set of causes, even if we were so inclined for theoretical reasons.[34]

---

[33] Obviously, this law should not be read as excluding earlier causes of the acceleration, i.e., forces that are exercised at earlier times. A similar lesson applies to other productive constraints. In fact Kim (1989) incorporates this in his Exclusion Principle, according to which only complete and independent causes compete. In causal chains later causes depend causally on earlier causes, so there is no competition.

[34] For instance, John L. Mackie's (1993) account of causes as "INUS conditions" allows for joint causation without being productive. If A, B, and C all occur and are insufficient (but necessary) parts of a sufficient (but unnecessary) condition for E, they jointly cause E. Suppose, however, that C correlates with D, as neural events may correlate with mental events. Then D is also an INUS condition for – and hence; a cause [Footnote continued on next page]

If one thinks of mental causation in this productive manner, then, causal exclusion is a natural consequence. The big question is of course whether applying productive constraints like those from the force/work domains to mental causation is *legitimate*. Setting the big question aside until the final section, I think principles of causal combination are important in many areas of scientific causal reasoning. Knowledge of what causes are at play in a given context is important, but must often be supplemented with knowledge of how the causes act together, whether they oppose each other and so on. L.G. Creary's (1981) distinction between "laws of causal influence" and "laws of causal action," (my "principles of causal combination") illustrates this. In mechanics Creary's laws of causal influence become force laws like that of Coulomb or Hook, which determine the causal influences from charged objects and springs respectively. But in typical mechanical problems objects are influenced by several forces. That is why we need a law of causal action like Newton's second law to relate these influences to resulting accelerations. As we shall see similar causal principles (whether they are "laws" or not) are needed in mechanisms more generally. I therefore suggest that a theory of causation counts as productive *if* it includes principles of causal combination that constrain how causes add up to yield their effects.[35] This has three virtues. First, it fits well with the way Kim and others think and talk about causation. Second, it explains why exclusion-type reasoning seems natural to them. Third, it dissociates productive causation from major ontological commitments. Productive constraints are simply rules of causal combination that can take sober, even mathematical forms. That such constraints apply to causation is at least *prima facie* compatible with a deflationist reading of talk of causal work, causal powers and the like. That some productive theories may ground the

---

of – E. Under a productive theory, on the other hand, D – the mental event – could be ruled out because its addition would yield a different effect E

[35] Clearly, not all principles governing how quantities add up are causal principles. For instance, the mass of two objects equals the sum of their individual masses, and yet this is not a matter of causation. I have no general way of distinguishing causal from non-causal combination to offer. However, my arguments below will only depend on principles we are antecedently inclined to label "causal."

productive constraints in, e.g., non-supervenient necessitation relations is no argument against productive constraints as such.[36]

3.4. First Response to Loewer: Empirical Theories of Causation

With this understanding in hand I turn to Loewer's arguments. The first thing to note is that a number of "empirical" theories of causation threaten to undermine Loewer's Argument from Science. These projects set conceptual analysis of causation aside and investigate the physical nature of causation in *our* possible world. Here are some rough outlines: (1) *Transference theories:* causation is the transference of some quantity like kinetic energy or momentum from the cause-object to the effect-object. (Aronson, 1971; Fair, 1979; Kistler, 1998) (2) *Process theories:* cause and effect are linked by spatio-temporal causal processes, where a process is causal if and only if it is capable of transmitting changes to its structure that arise from a single interaction. (Salmon, 1984) (3) *Conserved Quantity theories*: a causal process is a world line of an object that persistently possesses (or perhaps transmits) a conserved quantity like charge. (Dowe, 2000; Salmon, 1994) We have already seen Kim's appeal to Wesley Salmon's theory, and recently he has also expressed sympathy with the transference theory, which will be my focus here: "[…] it might be that efficacious/productive causality is 'implemented' or 'realized', in this and nomologically similar worlds, by the flow or transfer of a certain physical quantity." (Kim, 2002a, p. 677, see also 2005, p. 47n12 and forthcoming) As a response to Loewer this is perfectly to the point. While there are all sorts of problems

---

[36] Two analogue cases come to mind: First, it is a nice question in the philosophy of physics whether we should reify forces as ontological entities, or treat them merely as handy middle terms mediating between ontologically "innocent" claims about charges, positions etc. and claims about accelerations. (Jammer, 1957, ch. 11-12) But whatever we decide, worries about reifying forces obviously do not rule out that force-laws may legitimately be applied. Second, Ayer (1954) contended that the alleged conflict between freedom and determinism depends on a naïve animistic and anti-Humean conception of causation. (Note the similarity with Loewer's criticism of exclusion.) But then van Inwagen (1975) raised a problem for freedom using a formulation of determinism that did not invoke *any* conception of causation. These cases illustrate that sometimes philosophical problems can be raised independently of whether our principles or constraints are grounded in heavy-duty metaphysical notions.

with the transference theory, it is at least a theory that is (a) productive and (b) *designed* to be compatible with science.

Take (a) and its productivity first. Kim actually says that overdetermination "makes little sense" (2005, p. 47n12) when causation is understood in this way. A classical example from mechanics brings out why he is right. Train carriage A collides with carriage B, which is of equal mass, but at rest. In the collision the carriages connect and move on as one. According to the transference theory A causes B to move *by* transferring some of its momentum to B. The explanation of B's acceleration appeals to the conservation of linear momentum for the system: $m_A\mathbf{v}_{A,\text{ before}} + m_B\mathbf{v}_{B\text{ before}} = m_A\mathbf{v}_{A,\text{ after}} + m_B\mathbf{v}_{B\text{ after}}$. Accordingly, A's velocity is reduced by a half. $\mathbf{v}_{B,\text{ after}} = \mathbf{v}_{A,\text{ after}} = 0.5\mathbf{v}_{A,\text{ before}}$. Again, we have causal influences (momenta) combining to yield an effect. The principle of combination is the conservation of momentum, which puts clear limits on what causes – notably only causes *qua* momentum-instantiating objects – we can add to explain a given acceleration. It is no accident, then, that there is a tradition dating back to Leibniz that appeals to conservation laws to rule out mind-body interactionism. (McLaughlin, 1993; Papineau, 2001) So at least under Max Kistler's (1998) assumption that the quantities transferred must be conserved quantities, the transference theory counts as productive.

What about (b)? Are empirical theories like the transference theory incompatible with modern physics? Though this question is perhaps best left for philosophers of physics like Loewer, I want to make two points about it. First, the theories are based on, and *designed*, to be compatible with physics. So at the very least they pose a challenge that Loewer must face. Second, I suspect that Loewer is conflating "sufficient to produce an effect" with "nomologically sufficient for an effect" when he says that nothing short of a cross-section of an event's past light cone will be sufficient to produce it. (Loewer, 2002, p. 661n12) We certainly need something like the light cone story to rule out possible interfering causes and get a cause that is nomologically sufficient for the acceleration of carriage B above. But many of the empirical theories of causation are, or can be construed as, singularist theories of causation, according to which causation is not constituted by law-like regularities. And in a singularist sense of production (Anscombe, 1993) all that matters is that *this* time there were no interferers, so carriage A did indeed

make carriage B move. This is a general point from the philosophy of causation and is independent of specific questions about physics. I hasten to emphasize that my aim is not to defend the transference theory. My message is simply that Loewer appears to disregard important productive theories and the possibility of a singularist notion of production.

As we saw, Loewer also offered a Metaphysical Argument according to which productive causation might fail to supervene on fundamental physical facts. But it turned out that this argument rested on an appeal to the failure of supervenience in Tooley's (1990) theory of causation. Does the point generalize to *all* productive theories? As Kim (2002a) points out it is hard to see why the problem should apply to production understood as, e.g., transference. In fact, Kistler (1998) explicitly takes his transference theory to be opposed to that of Tooley. More generally, if production is understood in terms of causal combination as I suggest, worries about supervenience seem misplaced, and it is hard to see wherein the indigestible metaphysics lie. Arguments are needed to show that productive causation understood in this way is metaphysically objectionable. All in all, I suspect that Loewer has over-focused on the problems with heavy-duty metaphysics. This metaphysical focus prevents a more neutral understanding of productive causation. For these reasons I take Loewer's arguments to be, at best, inconclusive.

3.5. Second Response to Loewer: Mechanisms

Whatever the case may be in physics I think productive causation is needed in mechanistic explanations. Such explanations are of central importance in special sciences in general, and in neuroscience in particular. The case of neuroscience is especially important, since its claim to relevance for mental causation is at the very least as strong as that of physics. Here, then, are six quick reasons for the importance of mechanisms, see the papers cited for details.

(M1) *Mechanistic explanations are widespread in special sciences like biology, geology and perhaps even in the social sciences*. (Elster, 1989; Fodor, 1990, 1991b); Glennan, 1996, 2002; Hedström & Swedberg, 1998; Machamer et al., 2000; Woodward, 2002) Assuming we want our account of causation to be consonant with scientific practice this alone should make us want to account for mechanisms.

(M2) *Mechanisms are important for distinguishing spurious from causal generalizations*. Whether or not we infer a causal relationship from a correlation in science depends on whether we know, or are inclined to believe that the proposed causes and effects are linked by a mechanism. (See, e.g., Glennan, 1996)

(M3) *Mechanisms play a similar psychological role in ordinary causal inference*. Subjects tend not to infer causation from mere probabilities like P(sun-rising/rooster-crowing)>P(sun-rising/¬rooster-crowing), because they do not believe there are mechanisms connecting ordinary sounds and astronomical objects. (Cheng, 1997)

(M4) *Mechanisms are needed for manipulation purposes*. Knowledge of mechanisms can add to our manipulative capacities, e.g., by allowing us to increase or decrease the causal outcome of a process or to repair mechanisms that are broken. It is no accident that medicine is a science of mechanisms and not just correlations.

(M5) *Mechanisms are needed to explain correlations*. More controversially, we need mechanisms to explain why regularities hold in the special sciences.

(M6) *Mechanisms are needed because it is their components that do the causing*. Many find attractive the idea that regularities are symptoms of causation rather than constitutive of it. An example is "methodological individualism" in the social sciences according to which population-based regularities do not causally explain unless they are traceable to the *acts* of individuals in social mechanisms. (Elster, 1998; Hedström & Swedberg, 1998)

All in all, I take it to be an undeniable descriptive point that mechanisms do in fact figure prominently in special sciences, and there are additional reasons for regarding them as more or less the heart and soul of those sciences. But then it is surely a plausible constraint on *anyone's* theory of causation that at the very least (C1) *it should enable us to offer a workable account of mechanisms*. What is more, (C2) *that account should be conservative of scientific notions of mechanisms*. It is perhaps conceivable that philosophers can discover metaphysical reasons for reforming scientific mechanism talk, but pending knock-down arguments for this, we should set scientific practice first. Since Loewer intends his case against productive causation to be partly grounded in science, this is a constraint that he too should accept. Here, then, is a potential problem for Loewer. If mechanistic explanations invoke productive causation, then our philosophical

account should too. If mechanisms do not overdetermine their effects, then that is something we will just have to accept. To turn this into a real challenge I must show that mechanisms are productive in the sense that they are governed by productive constraints.

But what is a mechanism? As many have remarked, mechanism is *prima facie* an anti-Humean notion. Mechanisms are often supposed to be a link or "secret connexion" between cause and effect (Glennan, 1996; Psillos 2004), the existence of which Hume denied according to standard interpretations. Salmon (1984, p. 155) explicitly took his mechanistic process theory to answer Hume's challenge and actually "shew us" (Hume, 1978, p. 159) such a link.[37] Mechanisms are also frequently introduced to avoid the shortcomings of Humean deductive-nomological theories of explanation. (Elster, 1989, ch. 1) Now, the notion of a mechanism is (at least) ambiguous between how causes bring about their effects and how computational and other functions are implemented. In the first case we may wonder how pressing a button causes the doorbell to ring, and find out by directly investigating the causal chains linking the two events. In the second case we proceed indirectly, for instance by first specifying a function to be computed. (Say, to compute depth from 2D retinal images). Then we descend through David Marr's famous levels by finding out *how*, i.e., by which algorithm, that function is computed, and finally *how* that algorithm is implemented by neural causes in the visual system.[38] But in both cases I contend that we can view knowledge of mechanisms, at least in part, as knowledge of how causes bring about their effects.

Given this characterization a mechanism description is what you get when you ask "How did X cause Y?" And one need only be mildly sympathetic to mechanisms to agree that (perhaps setting aside the fundamental level of causation if there is one) it must always be possible to fill out a claim "X causes Y" with possibly nested "by ψ"-clauses describing *how* X caused Y. How did Peter cause the kettle to explode? Answer: by

---

[37] Though he accepted Hume's ban on causal language in that account. (Hume, 1978, p. 157); Salmon (1984, p. 155) An account of mechanisms need not be anti-Humean in the sense of being conceptually non-reductive.

[38] Marr, 1982, sect. 1.2. Note, though, that actual research arguably goes on at several of these levels simultaneously. (Churchland & Sejnowski, 1992, pp. 18-19)

clogging it and *expanding* the water inside. How did he cause the water to expand? Answer: by *heating* it. How does death of dopaminergic neurons in the substantia nigra cause hypokinesia or diminished movements in patients with Parkinson's? Possible short answer: by *decreasing* the rate of *firing* of neurons in motor cortex which normally *stimulate* muscle fibers to *contract*.

This reveals a very interesting connection between mechanisms and causal verbs. Mechanisms are naturally expressed by causal verbs, because these verbs can tell us how causes bring about their effects. In fact, the applicability of different kinds of causal verbs (say, lexical versus periphrastic causatives) varies with the way in which the cause is brought about. (See, e.g., Talmy, 1988; Wolff, 2002.) This is a lesson from cognitive linguistics, but the main idea goes back to G.E.M. Anscombe who took causal verbs to encode ways in which effects "derive from, arise out of, come of, their causes." (Anscombe, 1993, p. 92) The connection is all the more interesting since Kim has recently appealed to Anscombean derivativeness in characterizing his own productive view. (Kim, 2005, p. 18) Furthermore, Peter Machamer et al. take Anscombe and causal verbs as a starting point when characterizing mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer et al., 2000, p. 3) Their examples illustrate how neuroscientists actually use causal verbs like *phosphorelate*, *depolarize*, *inhibit* and *activate*. Given that even scientific mechanisms are naturally expressed by verbs encoding work and production in science, it should perhaps come as no surprise that productive causation is invoked in mechanistic explanations. In fact, verbs of creation like *write* appear to have productive constraints as part of their application conditions. Suppose we are ignorant of Wittgenstein's family relations and are being told first, that "Ludwig wrote the blue book," and, then, that "Margarete's brother wrote the blue book." We would be puzzled much in the same way that Kim intends us to be puzzled by the presence of mental and physical causes of bodily movements. And, much in the spirit of Kim's (1989) Exclusion Principle, we would not accept the two writing-claims without a story about how Ludwig and Margarete's brother are related with respect to writing. Did they co-write the book? Was Ludwig an "epiphenomenon" who used Margarete's brother as a ghost writer to do the real writing work for him? Are the two writers identical? A

positive answer to any of these questions would remove the tension between the two claims. This is revealing with respect to how we conceptualize and think about work.

But I do not think we have to consider causal verbs to appreciate the productivity of mechanisms. For constraints on causal combination are important parts of mechanistic explanations. I base this claim on the following general idea. To describe a mechanism for a causal process is to show how various causal influences are combined so that the process comes about. Similarly, to build a mechanism by which X causes Y from scratch, we would start by collecting causal influences, and connecting them together so that X does in fact lead to Y. To achieve this we must harness and control the causal influences, and this requires that we exploit principles of causal combination. This abstract idea certainly needs elucidation, but I hope that can be achieved by considering concrete examples of neural mechanisms. I note in advance, however, that my description appears consonant with three recent accounts of mechanisms, which, despite their differences, all stress the *complexity* of mechanisms. We have already seen Machamer et al. (2000, p. 3) on the "organization" of acts and entities in mechanisms. Stuart Glennan thinks a mechanism for a behavior is "a system that produces that behavior by the interaction of a number of parts." (2002, p. S344) Similarly, James Woodward's account invokes "an organized or structured set of parts or components." (2002, p. S375) I think the role of complexity in mechanisms is a key to understanding their productivity.

I start by briefly considering a recent argument for "ectopic neurotransmission." It is well known that neurons communicate via specialized synapses, where neurotransmitter is released from presynaptic "active zones" onto "postsynaptic densities." But it has also been suggested that in some cases "ectopic" release of neurotransmitter at sites distinct from the above-mentioned specialized regions might play a role in neurotransmission. ("Ectopic" means out of place.) To investigate this Coggan et al. (2005) developed a biologically realistic computational model of a specific type of excitatory synapse in chick embryos. (See Lučić & Baumeister (2005) for general discussion.) They found that the simulated postsynaptic effect did not conform to the actually measured effect, *unless* ectopic transmission was included in the model. What interests me presently is not the mathematical and biological details of the model – many of which escape me – nor its correctness, but rather the nature of their argument. They

argue for the presence of an additional mechanism (involving ectopic release) in certain cases of neurotransmission, on the basis that unless this mechanism is included we cannot explain measured effects (excitatory postsynaptic currents). This way of reasoning appears similar to Kim's idea that causes must add up to produce their effects. The causal influences, then, are synaptic and ectopic release of neurotransmitter respectively. How these influences combine depends on many factors including membrane topology, how transmitter molecules diffuse, the number and distribution of various receptor subtypes and so on. These constraints on combination are given a mathematical form in the model, and play an essential role in its predictions. Knowledge of causal combination is as important as knowledge of causal influences, and can give rise to exclusion-type reasoning.

Consider next the mechanisms responsible for maintaining the resting membrane potential in neurons. This is a paradigmatic example of a neural mechanism and is laid out in varying degrees of detail in text books such as Bear et al. (2001, ch. 3) or Kandel et al. (2000, ch. 7). The explanandum is simply that the outside of a neural membrane is normally (e.g., when no action potential is taking place) positively charged with respect to the inside. The difference is due to uneven concentrations of ions (most importantly Na^+, K^+, Ca^{2+}, Cl^- and organic anions) in intra and extra cellular space. In fact, there is a potential difference of –65 mV across a typical neural membrane. As in many neural mechanisms the explanation is in terms of ion flow across the membrane. Two factors, or causal influences, affect ion behavior and determine whether there will be a net influx or efflux of a given ion through its ion channels. (I disregard the influence of active transport mechanisms in ion pumps.) First, ions are subject to a diffusion force pushing them along their diffusion gradient from areas of high concentration to areas of low concentration. Second, being charged, ions are also subject to an electrical driving force, since opposite (equal) charges attract (repel) one another. Again, to get the explanation going, we must consider how the influences combine. Imagine that the concentrations of K^+ and some anion A^- are high on the inside of the cell, and low on the outside. Making the membrane selectively permeable to K^+, but not to A^-, by inserting potassium selective channels will have the following result. First, K^+ will flow out of the cell along its diffusion gradient. But as the inside becomes more negative than the outside the

electronic driving force comes into play and begins to pull K^+ ions back inside. When these two forces *balance each other* an equilibrium state is reached where there is no net flow across the membrane and K^+ ions have an equal probability of flowing in or out. The equilibrium potential (EP) for K^+, which would result if neurons were (like glial cells) only permeable to potassium, is typically –80 mV. We can formulate the influences *à la* Machamer et al. in terms of acts of entities, of charges *attracting* each other and so forth. And we can describe the principle of combination as *balancing*. These descriptions make intuitive sense, because they are borrowed from everyday language and experience. But the constraints can also be written into a quantitative description of the equilibrium potential known as the Nernst equation:

$$EP_{ion} = (RT/zF)\ln([ion]_{outside}/[ion]_{inside})$$

Here R is the ideal gas constant, T is the absolute temperature, z is the ion's valence, F is Faraday's constant and the subscripted brackets represent ion concentrations. At body temperature, then, EP is determined by the valence (z) and concentration gradient ($[ion]_{outside}/[ion]_{inside}$) for an ion. Importantly, the combination is productive. Suppose we measure a particular EP to, say, –80 mV, but do not know the valence of the ion or its concentrations. The equation constrains the causal influences we can postulate to explain that EP. We can of course change the postulated causal influences, e.g. by switching to an ion with a different valence z or by swapping the values of $[ion]_{outside}$ and $[ion]_{inside}$ – but that will typically change the EP as given by the equation.

However, actual neurons are more complex, since they are permeable to several ions. So the principle of combination underlying the actual resting potential (RP) must also be more complex and interesting. In fact RP is approximated by the Goldman equation, which I provide for potassium, sodium and chlorine ions. (Bear in mind that I include technical details like these because they illustrate a philosophical point about the productivity of mechanisms.)

$$RP = (RT/F)\ln((P_K[K^+]_{outside} + P_{Na}[Na^+]_{outside} + P_{Cl}[Cl^-]_{Inside})/ (P_K[K^+]_{inside} + P_{Na}[Na^+]_{inside} + P_{Cl}[Cl^-]_{outside}))$$

Here, P_{ion} is a measure of the membrane's relative permeability for an ion at resting potential, and depends on the number of open ion channels for that ion. RP therefore results from constraints on causal influences that are partly determined by valences and concentration gradients, and partly by the structural make-up of the membrane. Using everyday language we can metaphorically view the ions as *competing*, each ion *striving* to reach its EP. The result of the *battle* will depend on the degree to which the membrane *lets* the various ions pass through. In this particular case few would take the distinctively anti-Humean causal metaphor seriously. Even so, it may be cognitively useful, by allowing us to think of the mechanism in terms of concrete experiences we are familiar with. In other cases, e.g. in the *opening* of ion channels, everyday causal verbs may be literally applicable. But nothing hinges on this way of expressing the mechanism. For the causal combination is written into the ontologically sober Goldman equation. The value of P_{ion} determines the degree to which an ion contributes to the actual RP. At resting potential P_{Na} and P_{Cl} are much smaller than P_K, i.e., the membrane is more permeable to potassium. Therefore, in a typical neuron the actual RP = –65mV, which is close to EP_K, (–75mV). But things might have been otherwise (and does actually vary across neurons). If we want to postulate increased permeability for an ion, or add the influence of more ion types, we must take care not to get a different RP-value than the one we actually measure, for given that value there may not be any work or influencing "left to do" for that ion. To appreciate how influences combine to yield RP it is instructive to vary concentrations and permeabilities in online computer simulations of the Goldman equation.

The study of neural plasticity provides a third, less quantitative example of a productive mechanism, which is intimately related to cognition and mentality. It is nowadays almost a commonplace that many processes involving learning or memory formation require changes in synaptic strength. (Synaptic strength is the efficiency with which one neuron excites or inhibits another.) One important type of such change is long-term potentiation (LTP), and is believed to be crucial in *inter alia* hippocampus-involving memory processes. LTP is particularly interesting for our purposes because proposed models of its mechanisms are to some extent still controversial, so it makes sense to say

that mechanism hypotheses compete. LTP is subject to an enormous amount of empirical research, and has received attention from philosophers of neuroscience like John Bickle (2003, ch. 2), Patricia Churchland and Terrt Sejnowski (1992, ch. 5), Carl F. Craver (2002; 2003) and Craver and Lindley Darden (2001). In comparison, my treatment here must needs be dramatically simplifying and tuned to a very specific philosophical interest in causal combination. I will argue that different proposed mechanisms must either combine to yield LTP or they must exclude one another. In other words, LTP research is subject to what I call productive constraints. But what is LTP? Interestingly, it seems to be involved in some physical implementations of Hebb's famous rule. Donald Hebb proposed on theoretical grounds already in 1949 that learning might depend on activity-dependent changes in synaptic strength. In a slogan, his rule states that "neurons that fire together wire together." More precisely, if neuron A stimulates neuron B when B is already firing (due, for example, to other excitatory inputs from neurons C, D,…), then the A-B synapse should be strengthened. Strengthening an excitatory synapse A-B means to increase A's excitatory influence on B, making it more likely that B will fire when A fires. One of several quantitative measures of LTP is therefore an increase in the amplitude of the excitatory postsynaptic potentials (EPSPs) A causes in B. (Neurons compute whether they should fire or not based on spatial and temporal summation of EPSPs – caused by excitatory inputs – and inhibitory postsynaptic potentials (IPSPs) – caused by inhibitory inputs.) Tim Bliss and Terje Lømo, working in Per Andersen's lab in Oslo, famously brought about this effect at synapses in the rabbit hippocampus experimentally and published their results in 1973. They delivered a high-frequency electrical stimulus (a tetanus) to presynaptic neurons, and then measured the postsynaptic neuron's response to subsequent normal stimulations from the presynaptic neuron. The resulting EPSPs in the postsynaptic neuron displayed an increased amplitude when compared to measurements prior to the LTP inducing tetanus. (See fig. 1) The effect can last for hours and days, thus making LTP a potential player in, *inter alia*, hippocampal memory mechanisms.

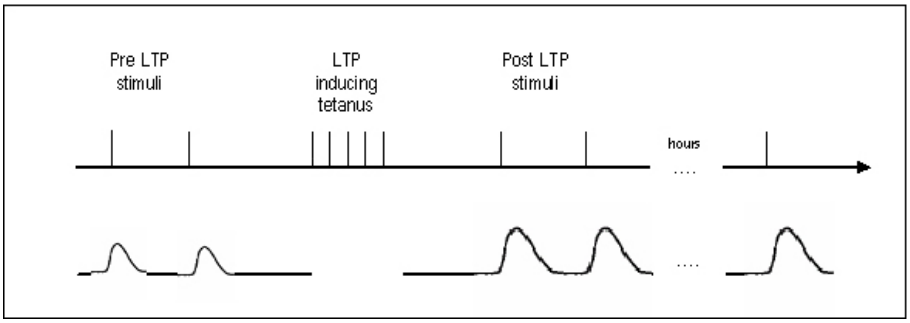

Fig. 1. Top: Test stimuli and LTP inducing tetanus delivered to postsynaptic neuron. Bottom: Measurements of EPSPs reveal a long-lasting increase in EPSP amplitude (a measure of LTP) subsequent to tetanus. (Adapted from Churchland & Sejnowski, 1992, fig 5.10)

How might one build a mechanism producing this effect? That is, how might causal influences be combined so as to yield LTP? One way would be to insert a "co-firing detector" in the postsynaptic neuron that is sensitive to whether the membrane of this neuron is strongly depolarized (which it would be *inter alia* when it is firing) while simultaneously being stimulated by the presynaptic neuron. Such an indicator would "light up" if the pre and postsynaptic neurons fire together, as in Hebb's rule. Then one could hook that detector's influence up with a process leading to the strengthening of the synapse. In the hippocampus, and in many other places in the central nervous system, nature appears to have chosen NMDA receptors for the detector job. These are actually Ca^{2+} channels that can open and let Ca^{2+} stream into the cell. However, they will only open if two conditions are fulfilled. First, glutamate – an excitatory neurotransmitter released by the presynaptic hippocampal neuron – must bind to the receptor. Second, since the channel is normally clogged by large Mg^{2+} ions, the receptor must also change its spatial configuration to let the magnesium clog pop out. (Mg^{2+}, then, is a *preventing* influence.) Now the NMDA receptors are in fact also voltage gated, so they will only change their configuration (and let out the clog) as a result of the nearby membrane being depolarized. Given this set up, then, an increased value of $[Ca^{2+}]_{inside}$, due to the calcium influx, signals co-firing of the pre and postsynaptic neurons. So from the point of view of information flow we have a story about how synapses can come to "know" when they

should initiate LTP. From a causal point of view, we can say that the structural make-up of Hebbian synapses is such that calcium can causally *influence* the LTP process when allowed to do so. Now synapses can be strengthened in numerous ways. Here are some hypothetical mechanisms that might occur downstream from Ca^{2+} influx in the case of LTP.[39] (H1) *Postsynaptic change: AMPA-receptors are made more sensitive*. AMPA receptors respond to released glutamate from the presynaptic neuron by opening and letting sodium ions flow into the cell, thus depolarizing the membrane. Making AMPA receptors more sensitive by adding a phosphate (PO_3) group to them would thus increase glutamate's depolarizing influence. (H2) *Postsynaptic change: increasing the number of AMPA receptors*. More AMPA receptors are inserted into the membrane. (H3) *Postsynaptic change: creating more synapses*. In a process involving the activation of genes and subsequent protein synthesis the structure of the dendritic spines is changed so that more synapses are formed. (H4) *Presynaptic change: increasing the amount of glutamate released*. A retrograde messenger, perhaps gaseous nitrous oxide (NO), diffuses backwards across the synaptic cleft and increases the amount of glutamate subsequently released. LTP mechanisms like these have been a matter of some controversy (Craver, 2002, p. S86), and several mechanisms may plausibly combine. My claim is that they are productive, insofar as they must add up to explain LTP effects like increased amplitude of EPSPs. If we had a good computational grip on LTP, then, we could imagine, say, (H4) being excluded on the grounds that, say, (H1) and (H2) are well-confirmed and account for the LTP effect alone. Adding (H4) would presumably give an even stronger synapse. I hasten to make two provisos about this. First, different mechanisms may be involved in different types of LTP, e.g., in short-lasting "early LTP" versus long-lasting "late LTP." Second, I do not mean to imply that neuroscientists' choice between the hypothetical mechanisms is always governed by the constraint that actual mechanisms must add up. (Though the reasoning in Coggan et al. (2005) appears to be partly of this type, and controls in some neuroscientific experiments may perhaps

[39] Discussions of LTP mechanisms can be found in neuroscience text books like Bear et al. (2001, ch. 22), reviews like Malenka & Bear (2004) or in the philosophy of neuroscience literature mentioned above. Bickle (2003, ch. 2), in particular, considers the biochemical processes in considerable detail.

play a similar role.) Computational models are only one of several methods invoked in neuroscience's search for mechanisms. Rather, I take the productive constraint to be a reasonable background assumption. For the mechanisms culminate in changes in synapses (e.g., in increased neurotransmitter release or the insertion of more receptors) that affect the presynaptic neuron's causal influence on the postsynaptic neuron.

Summing up, I have argued that principles of causal combination play an important role in mechanistic explanations. These constrain what causal influences we can postulate to explain the outcome of mechanisms, so productive causation is in fact invoked in mechanisms. This is something that our account of mechanisms should respect whether we opt for an account like that of Machamer et al. (2000) – which does appear to involve major metaphysical commitments – or (say) some variety of a regularity account like Glennan's (1996; 2002) or of a counterfactual account like Woodward's (2002). The relevant sense of production does not involve absolute nomological sufficiency or cross-sections of an event's past light cone. In fact, only causes of a specific kind are excluded, e.g., only additional *ions* are excluded as causes of membrane potential. I do not think this is a problem for the Exclusion Argument, however, because I do not think that argument must presuppose that bodily movements have absolutely sufficient physical causes. I take it to be a background assumption of neuroscience that *ceteris paribus* neural events have neural causes of the broad kind that we have been considering. The fact that neural causes are in the circumstances sufficient for bodily movements is sufficient to generate an exclusion problem, even without invoking a general causal closure principle according to which bodily movements have absolutely sufficient causes. (See paper (#4) and Bickle, 2003, p. 60; Kim, 2005, p. 155) So even though many neural mechanisms add up to produce bodily movements, bodily movements are not causally overdetermined by mechanisms.[40]

---

[40] A natural objection to the claim that mechanisms do not overdetermine their effects is the following. Current models of neural mechanisms are multi-level insofar as they involve entities from several levels of decomposition. (See, e.g., Craver, 2002; Schaffner, 1993, ch. 6) That is, the relevant mechanism descriptions include vocabulary from, e.g., the organismal, cellular and molecular levels. Does this mean that there actually are several overdetermining mechanisms at work corresponding to each of these levels? [Footnote continued on next page]

3.6. Conclusions

Let us recapitulate. If Loewer is right, as I think he is, Kim's Exclusion Argument requires that there be productive causation. In particular, it requires that *mental* causes be productive. Lower's anti-productive arguments are entirely general and do not rely on any specific features of mental causation. Mental causes are not productive because *no* causes are productive. But this is too quick. If productive causation is understood in terms of principles of causal combination, we can see that productive causation, and with it exclusion-based reasoning, can indeed be based on science. Furthermore, productive causation thus understood need not be cashed in terms of heavy-duty metaphysical notions. There appears therefore to be nothing wrong with productive causation *per se*. This is interesting in its own right. But it does not help much with respect to the Exclusion Argument, for Burge's above-mentioned challenge still remains. Why should *mental* causes be productive?[41] Mental causes, presumably, must rationalize, but it is not clear that they must produce. Why should beliefs and desires produce muscle contractions as action potentials in motor cortex do? Now, Kim does appear to offer a kind of Argument from Agency in favor of productive mental causation:

> Why do we care about mental causation? Because, first and foremost, we care about human agency. To save agency, however, we need the productive concept of causation; we want agents, in virtue of the beliefs and desires and intentions they hold, to cause their limbs to move in appropriate ways and thereby produce changes in their physical surroundings. I don't think the

---

This does not appear to be Craver's and Kenneth Schaffner's intention, however. Rather they seem to be emphasizing that mechanism descriptions currently available in neuroscience are typically partial, and that methodologically neuroscience proceeds simultaneously at several levels. Furthermore, I take there to be a very real sense in which there is only one mechanism involved here, albeit one that can to a certain extent be described at multiple levels of composition. So it is not clear to me whether such cases really amount to overdetermination, nor whether any such "multi-level overdetermination" would be acceptable. This question is certainly worthy of further investigation.

[41] Loewer (forthcoming) actually appears to agree that there might be a non-heavy-duty account of production, perhaps something along the lines of Hall (2004). But, like Burge, he now questions the specific assumption that mental causes must be productive.

> kinds of dependencies that can be captured by counterfactuals alone would be enough for the job. (Kim, 2002a, p. 675, see also Kim forthcoming.)

His motivation for reductionism, then, boils down to this. (1) Agency requires productive mental causation. (2) Productive mental causation requires reducing the mental. (3) Hence, we had better reduce the mental. Now (2) is weaker than the claim that mental causation *period* requires reducing the mental, and seems to me to actually be supported by the Exclusion Argument. But on closer inspection (1) lacks independent support in what Kim says.[42] Philosophers like Burge and Loewer, who apparently do not share Kim's intuitions, will therefore be quick to charge him of begging the question. For all he has told us some model of "supervenient causation," where non-productive mental causation supervenes on physical productive causation might still be a viable antireductionist response to the exclusion problem. (Fodor, 1990; Jackson & Pettit, 1988; Kim, 1984b; Loewer & Lepore, 1987) Are we, then, back where we started? Loewer will deny that mental causes are productive, and Kim will go on to insist that they must be. While debating the validity of intuitions is unlikely to resolve this deadlock, I think considerations of explanatory practice might help. Consider Burge's claim that neural and psychological explanations of bodily movements describe "the same physical effect as the outcome of two *very different* patterns of events. The explanations of these patterns answer two *very different* types of enquiry. Neither type of explanation makes essential, specific assumptions about the other." (Burge, 1993, pp. 115-116, my italics) Now, I am not convinced that Burge's quick appeal to explanatory differences will make the antireductionist's day.[43] But it does point in the direction of a new, and potentially friutful, way to approach an old question. Is "supervenient causation" just epiphenomenalism in disguise? I think we should approach this question anew, by

---

[42] Kim (forthcoming) does offer additional considerations in favor of this claim. He seems to appeal partly to well-known problems with counterfactual theories in general and partly to the idea that counterfactual theories allegedly make mental causation "too easy," because it turns omissions into actions. Suffice it to say that it is unclear whether these considerations have much force. Kim admits that they do not constitute a "knock-down argument," and Loewer (forthcoming) remains unconvinced by them.

[43] Cp. paper (#2).

considering our psychological explanatory practices to see whether there are any specific features of mentality that requires mental causation to be productive. If there are no such features, we might have a before us a very attractive Burgean synthesis of the productive view and its non-productive anti-thesis. *Pace* Loewer, there is nothing wrong with productive causation, *pace* Kim, mental causation is not productive.

4. Paper (#2): Pace Burge: Some Empirical Warrant for Epiphobia

ABSTRACT: Many physicalists think that some tight relation like supervenience or identity between mental and physical events as well as the presence of neural mechanisms underlying mental causation is required to rule out epiphenomenalism about the mental. Tyler Burge has argued that worries like these arise only because physicalists are misguided by metaphysics, and fail to appreciate the essential role of explanatory practice in determining causal powers. I explicate Burge's notion of causal powers and his anti-physicalist arguments. While his notion of causal powers is valuable, his use of it to argue against physicalism presupposes a strong autonomy of psychology *vis-à-vis* neuroscience. By considering the neuroscience of voluntary behavior in general and Patricia Goldman-Rakic' theory of prefrontal cortex in particular, I show that Burge's way of individuating scientific practices is deeply problematic. It fails to appreciate that neuroscience combines an explanatory interest in cognition with an interest in neural mechanisms. My discussion serves to motivate the physicalist call for mechanisms in mental causation as well as the need for some tight relation between mental and physical causes. Importantly, this is a motivation from *within* explanatory practice, and accordingly a motivation that Burge should accept.

How can the soul of a man determine the spirits of his body so as to produce voluntary actions (given that the soul is only a thinking substance)?

– Princess Elisabeth of Bohemia[44]

4.1. Introduction. Physicalist Constraints on Causal Relevance in Psychology

An interesting feature of mainstream philosophy of mind is that many of its problems arise because physicalists hold views about mental causation like the following. (These views and the corresponding problems will be clarified below.)

---

[44] Letter to Descartes of May 6/16, 1643. Translated in Nye, 1999, p. 9 / AT III 661.

(P1) Mental causes of bodily movements must depend in some suitably strong sense on the movements' physical causes.

(P2) Mental causation must involve mechanisms.

(P3) Mental causation requires "narrow" or locally supervenient mental content.

(P4) Mental causation must be backed by laws.

An interesting feature of Burge's philosophy of mind is that he is very skeptical of these claims about mental causation. He seems to take the corresponding problems to be pseudo-problems that arise only if we are misguided by the physicalist metaphysics of (P1)-(P4) instead of paying attention to explanatory practice. Once we realize the essential role of explanatory practice in determining causal relevance it becomes clear that (a) in general, depending on their explanatory purposes, different practices can give explanations that may or may not conform to constraints like (P1)-(P4); (b) in particular, it is far from clear that psychological explanations must conform to these constraints. I call this Burge's *Pluralistic Attitude to Causal Explanation*. (PACE, for short.)

This paper has two parts. First, I offer a detailed analysis of Burge's reasons for skepticism about (P1)-(P4), by focusing on his notion of causal powers.[45] (Section 4.2-4.3) I think his framework is valuable and useful, and that his emphasis on explanatory practice as *the* reliable source to knowledge about causation constitutes sound advice. Second, I argue that, in spite of the attractions of Burge's position, his dismissal of (P1)-(P2) fails, because his individuation of practices fails to capture the multidisciplinary nature of cognitive and neuroscience. (Section 4.4) I will not take a stand on (P3) and (P4), as they fall outside the scope of the present paper. Nevertheless, I describe Burge's arguments against these constraints to show that his disagreement with physicalism is quite general.

Let me start by quickly sketching how (P1)-(P4) give rise to central problems in the mental causation debate. More details will be provided below, but note that I will to a certain extent abstract away from differences between various physicalists to emphasize

[45] Burge typically formulates his arguments in terms of "causal powers." For the purposes of this paper I use this and alternative expressions like "causal efficacy" and "causal relevance" interchangeably.

how Burge's position differs from that of physicalism in general. The problems of (P1) are clustered around *causal exclusion*: given that bodily movements have sufficient physical causes – and given that bodily movements are not to be causally overdetermined by more than one sufficient cause – is there really any room for additional mental causes? Will not putative mental causes be excluded and rendered epiphenomenal? This is the threat of the so-called Exclusion Argument. (Kim, 1998; 2005) Physicalists have suggested that exclusion can be avoided since the mental and physical causes are related by, e.g., supervenience, type-identity or token-identity. But whether any of these relations are satisfactory or viable remains controversial.

(P2) – the request for a mechanism – is motivated *inter alia* by the idea that since mental causation cannot be fundamental causation, it must be mediated by mechanisms. The problem, then, is to offer a model for mental causation that involves mechanisms. (Fodor, 1990; 1991b)

The problems with (P3) arise because the received view has it that ordinary mental content is wide, that is, individuated with reference to states that are external to the organism. Content is also supposed to play a role in causing bodily movements. But if it is wide, how can content play a causal role locally, via the brain? Since causation must run locally, mental causal explanations require a *different* sort of content, narrow content, that is locally supervenient. (Fodor 1987, ch. 2; 1991a) Whether some notion of narrow content is viable remains controversial, however.

Finally, (P4) seems problematic because there arguably are no strict laws of the mental. So if mental events are to be causally efficacious, that requires something like a redescription of them in terms of the vocabulary of physics which according to many does allow for the formulation of strict laws. (Davidson, 1980) However, it remains controversial how this solution differs from epiphenomenalism, as mental events only seem to be causally efficacious in virtue of their physical properties.

Notice that these hard problems arise because principles like (P1)-(P4) put constraints on the causal relevance of mental properties. *If* mental properties are to be causally relevant, *then* they must satisfy the constraints of being appropriately related to other kinds of properties, being mediated by a mechanism, being locally supervenient or being apt for entering into laws. Burge, on the other hand, wants us to *stop worrying*

about these constraints. For instance, with respect to (P1) and the Exclusion Argument, he claims that mental and neurophysiological explanations of the same movement can coexist happily and independently, and that there is no apparent reason why we should have to invoke some tight relations between mental and physical causes to rule out epiphenomenalism. I call this Burge's *compatibilism*, since he takes *overdetermination* to be a misleading description (Burge, 1993, p. 101n3). Compatibilism is supposed to follow from the interesting, but somewhat unclear, idea of different explanations citing different *patterns of events*.

> A man's running to the store is explained by his believing that his child would suffer without the needed medicine and by his decision not to wait on a doctor. […] It would be perverse to think that such mental events must interfere with or alter, or fill some gap in, the chain of physiological events leading up to and including the movement of his muscles in running. It would be perverse to think that the mentalistic explanation excludes or interferes with non-intentional explanations of the physical movement. I think these ideas seem perverse not because we know that the mental events are material. They seem perverse because we know that the two causal explanations are explaining the same physical effect as the outcome of two very different patterns of events. The explanations of these patterns answer two very different types of enquiry. Neither type of explanation makes essential, specific assumptions about the other. (Burge, 1993, pp. 115-116, my italics)

I shall return to this quote and the idea of patterns of events repeatedly. The quote shows that Burge and most physicalists alike take epiphenomenalism (the view that mental causes are excluded and rendered causally impotent by physical causes) and interactionist dualism (the view that mental causes interfere with or alter physical causal processes) to be *perverse* or unacceptable. They differ, however, *inter alia* when it comes to constraints like (P1) and the question whether tight relations are necessary to avoid exclusion and interference.

Burge is equally skeptical to (P2) and the question about mechanisms of mental-to-physical causation. He says that it is unclear "whether [this] question is an appropriate one in the first place" and that it is "not even obvious why any model [for mental-to-physical-causation] is needed". (Burge, 1993, p. 114) Physicalists on the other hand, typically think that a mechanism *is* needed. They feel that at least some of the questions

raised by (P1)-(P4) are perfectly *good questions*, and that it is incumbent upon us as philosophers of mind to attempt to answer them. However, this is *not* necessarily because they take epiphenomenalism to be a real possibility – most physicalists do not. It is important to distinguish between the following flavors of what Jerry Fodor (1990) called "epiphobia": (i) one can worry about epiphenomenalism and take it to be a real possibility; (ii) one can be convinced that epiphenomenalism is false, and yet think that constraints like (P1) and (P2) *must* be satisfied, despite the difficulties involved in accounting for *how* they come to be satisfied. I contend that most physicalists are type-(ii) epiphobics. Fred Dretske (2003) makes much the same point against Burge, but the latter (2003) remains unsympathetic to type-(ii) epiphobia.

Be that as it may. We should get clear about Burge's reasons for dismissing our problems as pseudo-problems. I argue that his notions of "causal relevance" and "patterns of events" are the keys to understanding these reasons. While Burge's 1993 paper has been quoted often enough I know of no systematic discussion of how these notions are supposed to ground, for instance, his compatibilism regarding mental and physical explanations. This is unfortunate, since we cannot assess his claims before we understand their foundations. I aim to amend this by presenting one systematic way of laying out a Burgean view of causal relevance and explanatory pluralism. To anticipate one of my main conclusions, Burge appears to dismiss the need for (P1)-(P4), because in his view constraints on causal relevance in a given science or explanatory practice must be motivated from *within* that practice. Burge, then, seems to be questioning the *motivation* for type-(ii) epiphobia.

Before I move on, however, some, qualifications are in order. Because Burge is not particularly revealing about the structure of his arguments, I am not sure whether my account is in all respects an accurate description of his views. I do, however, think my exposition captures the general Burgean spirit of his 1993 paper. Whether I get all the details correct is nevertheless of secondary importance, as I believe that structuring the arguments in this way illuminates the nature of Burge's dispute with the physicalists. Furthermore, I think the general Burgean position I am describing is interesting in its own right. My objection is directed solely at Burge's use of the framework discussed here (or something very much like it) to dismiss the need to relate different types of explanation

to each other (P1), as well as his skepticism about the need for mechanisms in mental causation (P2). Put briefly, my point is that Burge's arguments against (P1)-(P2) appear to fail, not because his general view of the role of explanatory practices in determining causal relevance is flawed, but because his way of individuating the relevant practices is problematic. In short, it is highly unclear whether Burge's views can be justified in accordance with the internal standards of those practices, that is, by the very standards he himself acknowledges as decisive.

4.2. Burge's Theory of Causal Explanation and Causal Relevance

We have seen that Burge is a compatibilist – i.e., he thinks that mental and physical explanations of an action can coexist happily without us having to relate them metaphysically – *because* the action is explained as the outcome of "two very different patterns of events." (Burge, 1993, p. 116) To understand this claim I will also rely on Burge (1989), where he offers a more detailed account of his notion of causal powers. Based on the above quote, in addition to others we will encounter below, it seems clear that causal explanation for Burge is a matter of describing effects as the outcome of patterns of events.

> *View of Sciences and Causal Explanations*: Different sciences such as psychology and neuroscience study and describe different patterns of events in their causal explanations.

In fact, patterns seem to determine causal relevance by putting constraints on which properties are relevant relative to the pattern.

> The causal powers of a kind of event are to be understood in terms of the patterns of causation that events of that kind enter into. Such patterns are identified as explanatory in causal explanations. And the properties that 'determine' the causal powers of an event are those that enter into causal explanations. (Burge, 1993, p. 100)

Burge's favorite example is taken from physiology.

> Not all the causal powers of an entity, considered in the abstract or from the point of view of physics, are relevant to typing it. The heart has numerous "causal powers" that are irrelevant to its being a heart. It will [for example] color a surface red if dropped from a given height; [...]. None of these powers is relevant to typing something as a heart. None are causal powers recognized by biology or physiology. What are relevant are those causal and receptive powers exercised by the heart that yield the patterns of causation studied in physiology – the powers exhibited when it carries out its basic function, pumping blood in the circulatory system of an organism. (Burge, 1989, p. 316)

Unfortunately, Burge does little to characterize the operative understanding of patterns of events. Here is one way of explicating the notion that seems to fit with his examples.

> *Pattern of Events*: A pattern of events is a causal interplay between kinds of events that are described by a science, which in turn is characterized by an explanatory interest in this particular type of interplay.

So events may be causally related to each other in many different ways (say, physiologically, psychologically or geologically). That explains why there are different sciences which study different patterns of events, or, if one prefers, different "networks" generated by causal relations among kinds of events. The take-home message of the heart example is that in order to explain an object's or event's role in a pattern, a given science only needs to refer to a *proper subset* of the object's or event's causal/receptive powers. Causal relevance of a property, then, is relative to some particular science with its characteristic explanatory aims and purposes. Switching to another science may make the same property causally irrelevant. Accordingly, "pattern-contribution" appears to be both necessary and sufficient for causal relevance in a science or explanatory practice.

> *Causal Relevance*: A property P is causally relevant to an explanatory practice if and only if citing P contributes to the description of a causal pattern of events which is of interest to that practice. Otherwise, P is causally irrelevant to the practice.

Now, given Burge's way of thinking of causal relevance, the causal efficacy of the mental appears to be *guaranteed* by psychological explanatory practices like folk-psychology. The action explanations offered by folk-psychology are typically taken to be causal explanations. These explanations arguably appeal to intentional properties expressed by phrases like "believing that p" and "desiring that q." Since such properties endow folk-psychology with tremendous predictive powers (Fodor, 1987, ch. 1), there can be little doubt that they are causally relevant, according to Burge's criterion for causal relevance. If a property figures in explanations that are widely acknowledged as successful and causal, there appears to be little or no reason for thinking it epiphenomenal. (Note that it is not always clear whether Burge has folk-psychology or academic/cognitive psychology in mind as a source to knowledge about mental causation. I shall at any rate be drawing on both in what follows.)

Psychological explanatory practices, then, provides a positive answer to the question "is the mental causally efficacious?" But as we saw above, few physicalists would dispute this, even though they are also interested in finding out *how* epiphenomenalism can be ruled out, given additional constraints on causal relevance like (P1)-(P3). (Dretske, 2003; Kim, 1998, p. 61; McLaughlin, forthcoming) In fact, physicalists need not dismiss Burge's notion of causal relevance. This notion is, after all, compatible with there being further constraints like (P1)-(P4) that must be satisfied if a property is to be causally relevant. (These constraints might, for instance, help determine the conditions *under which* a mental property contributes to pattern-description.)

Why, then, does Burge have so little patience with the "how"-questions? I will suggest that Burge is not just being insensitive to the "how"/"that"-distinction. Rather, he appears to question the motivation of the constraints and the "how"-questions they give rise to. Burge is in general hostile to using standards of one particular scientific practice in other contexts. As an example consider his ironic remark that proponents of narrow, i.e., locally supervenient, mental content think that "cognitive psychology (*unbeknownst to itself*) needs [conceptions of narrow content] as surrogates for more ordinary, non-individualistic conceptions". (Burge, 1989, p. 304, my italics) The implicature being that: psychologists do not need philosophers to lecture them about the type of content to use. We shall see several illustrations of this attitude below. While I am aware of the danger

of over-generalizing from Burge's claims, I believe the following principle captures at least a broadly Burgean attitude to explanation.

> *Principle of Responsible Practices*: Demands, criteria and standards for explanation in an explanatory practice must be justified internally to the practice, not externally, i.e., roughly, not with reference to standards that belong in other practices or in pure philosophy.

As we shall see, this principle fits well with Burge's way of criticizing constraints (P1)-(P4). He dismisses the idea that constraints like these must be satisfied to rule out epiphenomenalism, because he questions whether they can be motivated from *within* psychological explanatory practices. The principle and Burge's arguments, then, have a slightly Carnapian flavor to them. Just as there are good (internal) and bad (external) existence questions according to Rudolf Carnap's famous distinction (1950), there are good (internally justified) and bad (externally "justified") constraints on causal relevance according to Burge.

But what does independence mean here? The quotation I have been focusing on, taken from Burge (1993, p. 116), hints at the way in which Burge takes psychology and neuroscience to be independent. He dismisses the need to relate them *because* they (i) study very different patterns of events, (ii) their explanations answer very different types of enquiry and (iii) they do not make any essential assumptions about each other. If we want to move on from what Burge explicitly says here to a general account of internal justification we may say that:

> *Internal Justification* means justification by reference to standards belonging to the practice itself or to a related practice. Here two practices are related if one makes essential assumptions about the other or they express a related explanatory interest.

The Principle of Responsible Practices, then, gives rise to what I call PACE, or Burge's Pluralistic Attitude to Causal Explanation. Constraints on causal relevance may vary

across practices, and practices are free to set their own explanatory standards in relative isolation from other practices. In particular, I propose that this attitude – rather than sheer insensitivity to the "how"/"that"-distinction – is what makes Burge skeptical to the application of constraints like (P1)-(P4) to psychology.

But why should we believe in the Principle of Responsible Practices? Fortunately, Burge has at least one explicit line of argument leading to something like the principle. Burge's primary interest in his 1989 and 1993 articles is in causal explanation in psychology. In this connection he often makes remarks like the following.

> Our understanding of mental causation derives primarily from our understanding of mentalistic explanation, independently of our knowledge – or better, despite our ignorance – of the underlying processes. Materialist accounts have allowed too wide a gap between their metaphysics of mental causation and what we actually know about the nature and existence of mentalistic causation, which derives almost entirely from mentalistic explanation and observations. (Burge, 1993, p. 103)

If this argument is intended to apply more generally to other practices, as I think it is, we can view it as an instance of a general *epistemic primacy strategy*: (i) knowledge of causation and explanation in any given practice stems primarily from that practice. (ii) Therefore, explanatory demands, criteria and standards from other practices simply cannot tell us much about causation and explanation in the given practice, much less force it to reform its explanatory standards. (iii) Therefore, practices are themselves responsible for determining their standards.

The idea of patterns of events might also be used to formulate a direct argument, based on a more general Burgean worry that external constraints on explanations may distort the practice's distinctive character. We shall see that Burge's dismissal of (P3) and the call for narrow content, for instance, is based on a worry that wide content is required to capture the patterns of interest to psychology. Similarly, we saw that he thinks physiology must focus on a certain set of properties of the heart if it is to capture the pattern of events in which the heart functions to pump blood.

To explain its characteristic patterns, then, a given science *must* focus on certain properties. If this focus is pushed in other directions – for example if psychology is

forced to use narrow content – it might *loose* its pattern, and instead end up describing some other pattern. The subject matter of the science may not turn out to be the same if external constraints on causal relevance are allowed. I.e., roughly, *Threat of elimination*: (i) if practices cannot set their own explanatory standards, they cannot explain their characteristic patterns. (ii) Sciences must explain their characteristic patterns. (iii) Therefore, sciences must set their own standards.

By formulating the argument in this manner we can see why constraints (P1)-(P4) might seem threatening to someone like Burge; for one might worry that a Davidsonian "change of subject" or eliminativism is lurking behind these constraints. (See Donald Davidson, 1980) This threat is, I take it, very real. But it is not clear that changing the subject is always something to be feared. We should not forget the *positive* take-home message of thinkers like Paul K. Feyerabend (1988) and Thomas Kuhn (1964). Setting aside worries about whether scientific progress is a matter of replacement rather than accumulation, history appears to tell us that progress involves conceptual change more often than not.

4.3. Burge's Arguments Against Physicalist Constraints on Causal Relevance

We are now in a position to see how Burge argues against physicalist constraints on causal relevance in psychology. By denying these physicalist constraints Burge is exemplifying what I have called PACE. PACE potentially gives rise to a plurality of explanations that differ across the sciences depending on the nature of the patterns to be explained. Since our theory of causal relevance determines the set of properties we are allowed to play with in our causal explanations, I view Burge's theory causal relevance as the crucial element in his arguments against the need for (P1)-(P4). For Burge, the question whether these constraints must be satisfied hinges primarily on whether their satisfaction is necessary for pattern description. He argues that it is far from clear that they are needed to describe the patterns of events that psychology tracks.

4.3.1. (P1) and the Call for Tight Relations between Mental and Physical Causes

One statement of this constraint is Kim's (1989) Principle of Causal-Explanatory Exclusion. According to this principle there cannot be two complete and independent causes of the same effect. It is hard to make perfectly explicit what "complete" and "independent" should mean, but I am not convinced that we have to do so or indeed buy completely into the controversial principle, in order to appreciate the weaker kind of advice it gives us. If we have reasons to believe that voluntary bodily movements systematically have two kinds of causes, one physical and one mental, we should try to relate the causes somehow. Whether they are motivated by the Exclusion Argument or not, philosophers like Kim seem to find the presence of two explanations intuitively puzzling and even suspect a certain amount of tension between the two causes, lest they are tightly related. Suggested relations that might remove this tension include the following. *Supervenient causation*. Mental causes supervene in some strong, perhaps mereological/constitutive, sense on the neurophysiological causes that do the "causal work" of bringing about the effect. (Fodor, 1990; Kim, 1984b) *Token-identity*. We do not really have two causes here because the mental cause = the physical cause at the level of tokens. (Davidson, 1980) *Type-identity*. The same, but the mental cause = the physical cause at the level of kinds (Kim, 1998, ch. 2; 2005; ch. 2). *Overdetermination*. The bodily movement *is* overdetermined by two sufficient causes, but that is acceptable since the covariation between the two causes is systematic and non-accidental due to some dependency relation like supervenience, suggesting again a model of supervenient causation. (Loewer, 2002)

Another motivation for relating the two causes is that many of us have antecedent physicalist expectations to the effect that mental causation must occur via more fundamental physical mechanisms (see below). This also calls for an account of how mental and physical causes are related metaphysically. So, setting aside details, there are reasons for thinking it a necessary condition on the causal relevance of mental properties that they be appropriately related to physical properties. (Or, if not strictly speaking a necessary condition, then a relatively strong requirement.)

Let us look at the *mereological (part-whole) relation* since in Burge's thinking this is the "critical one" (1993, p. 113n14), and since I shall argue *contra* Burge that this one *does* seem to be relevant to bridging psychology and neuroscience. Mereology is attractive because it promises both a general account of the way in which the mental depends on the physical, and a schema for physical explanations of mental causation. As for dependency, it has been suggested that objects that are indistinguishable with respect to physical micro-structure must also be indistinguishable with respect to mental properties. (Mereological supervenience, Kim, 1998, p. 17) More importantly, we may be able to explain how mental causation is implemented physically by reference to the behavior of the parts constituting the nervous system.

With respect to models like supervenient causation, Burge admits that such "projects can be interesting," but he denies that they are *needed* to account for causal relevance of mental properties. (Burge, 1993, p. 102) And even if part-whole relations are *sometimes* relevant for explanatory purposes, that need not always be so because

> […] the relations of identity and physical composition are relations that have specific scientific uses. For example, we explain the behaviour of a molecule in terms of the behaviour of its component parts. It is far from clear that these compositional relations have a systematic scientific use in bridging psychology and neurophysiology […] They are guesses about what sort of relation might obtain. (Burge, 1993, p. 116)

In a footnote he elaborates his skepticism to the relevance of mereology in bridging psychology and neuroscience:

> There are forms of materialism that maintain that all objects are decomposable into inorganic physical particles. […] they make a claim for the relevance of physical composition to our understanding of mental entities that seems to me (so far) quite unsupported by anything we know. (Burge, 1993, p. 113n14)[46]

[46] Note that Burge in this quote is talking of decomposing mental entities into physical entities, whereas in the quote above he is talking about explaining the behavior of wholes in terms of the behavior of parts. His skepticism about the usefulness of part-whole relations to understanding the mental and mental causation appears quite general. When I defend the use of part-whole relations in understanding mental causation, [Footnote continued on next page]

Despite the relative merit of appealing to mereological relations in some scientific contexts, Burge argues that we do not, at least not at present, know whether it would be fruitful to apply them in relating psychology and neurophysiology. And we are even less entitled to make such relatedness a criterion for causal relevance. In the final section I shall argue that we *do* know enough to believe that part-whole relations are relevant to bridging psychology and neurophysiology. Some tight relation like perhaps mereological supervenience *is* required to understand how physical and mental causes are related. Part-whole relations are also needed to understand how mental causation is implemented neurally. For now, however, I shall simply yield to the call of generality and recast what Burge explicitly says in a general Burgean principle:

> PACE1 *Freedom of Metaphysical Relatedness*. In different sciences the interesting metaphysical relations between causally relevant properties may vary; in particular the part-whole and identity relations may only be relevant in certain scientific contexts.

4.3.2. (P2) and the Call for Mechanisms

According to a popular and attractive idea causation requires causal mechanisms that mediate between cause and effect. While many different accounts of mechanisms have been proposed, it is often claimed that they must involve causation at some physical level. (Cp., e.g., Fodor, 1990) Especially in the philosophy of mind there is a lively discussion about causation as production, and of causes doing causal *work* to bring about their effects, and this work is often implicitly or explicitly construed as physical work. (Cp., e.g., Jackson, 1996, p. 394; Kim, 1998, p. 37; 2002a; 2005, p. 18) If one is thinking of causation in this way, it becomes pressing to account for how mental causation is related to the physical mechanisms that bring about bodily movements. Must mental causes themselves be productive causes, as Kim appears to require? Or is it sufficient that

---

what I have in mind is the use of part-whole relations in mechanisms, to explain how mental causation is implemented physically.

they supervene on underlying physical mechanisms? (Fodor, 1990) A less committal motivation for causal mechanisms is the following. Psychology simply is not a fundamental science, so mental causation cannot be basic or fundamental causation. Hence, mental causation requires underlying, presumably neurophysiological, mechanisms.

Not unexpectedly, Burge is skeptical of making mechanisms a constraint on causal relevance, because he thinks the call for mechanisms originates in practices which he takes to be very different from psychology. Looking for a mechanism "that would make possible causal interaction between two such different things as a physical event (or substance) and a mental event (or substance)" (Burge, 1993, p. 114) amounts to demanding a physical model of causation, and he questions the applicability of the physical model to psychology. In fact he goes as far as to question the need for *any* model of mental to physical causation:

> I have no satisfying response to the problem of explaining a mechanism. […] What is unclear is whether the question is an appropriate one in the first place. Demanding that there be an account of mechanism in mind-body causation is tantamount to demanding a physical model for understanding such causation. It is far from obvious that such a model is appropriate. It is not even obvious why any model is needed. (Burge, 1993, p. 114)

From the point of view of mainstream philosophy of mind this is a rather puzzling and radical claim. How can we account for mental causation in a physical world, without invoking some notion of a mechanism? We should, however, note that Burge is not necessarily against all mechanistic explanation, as he thinks there are "underlying physical processes" on which mental processes depend. (Burge, 1993, p. 116) So presumably neuroscientists are free to discover mechanisms, but their mechanisms will not provide a model for *mental* causation. Perhaps, then, the mechanisms explain how physiological events (like retinal events, in perception) produce *other* physiological events (like muscle contractions). They do not, however, explain how *mental* events cause physiological events.

The notion of a mechanism is arguably closely related to that of productive causes that is, roughly, of causes doing causal *work* to yield their effects.[47] (Kim, 1998, p. 37; 2002a; 2005, p. 18) Burge, in fact, thinks the Exclusion Argument and its idea that irreducibly mental causes and physical causes of bodily movements exclude one another, depends on a productive view of mental causes, that is on "thinking of mental causes on a physical model–as providing an extra 'bump' on the effect." (1993, p. 115) However, he questions whether mental causation should be understood by reference to physical causation: "But whether the physical model of mental causation is appropriate is, again, part of what is at issue." (1993, p. 115) Mental causes, presumably, must *rationalize*, but the claim that they must *produce* their effects as physical causes appear to do, seems hard to motivate from within psychology. In the more general Burgean position I am formulating, the mechanism critique is based on:

> PACE2 *Freedom of Mechanism*. Causation in different sciences may involve different causal mechanisms or perhaps they need not involve any mechanisms at all.

In contrast with Burge I think we do need an account of mechanisms in mental causation to rule out epiphenomenalism, and below I shall argue so on a scientific basis that Burge should accept.

4.3.3. (P3) and the Call for Locally Supervenient Properties

As another illustration of Burge's pluralistic attitude, consider more briefly his critique of narrow content. Conventional philosophical wisdom has it that (i) content properties like the property of believing that p are causally relevant to behavior. But (ii) content properties are not intrinsic to the organism. They depend on relations between the organism and its physical environment and linguistic community, and therefore fail to supervene locally on the organism's non-relational physical properties. (Burge, 1979)

---

[47] Cp. paper (#1) for more details on the productive view of mental causation and its relation to mechanisms.

The much discussed problem of combining (i) and (ii) can be viewed as arising from a view about *causation*, claiming roughly that: (iii) causation is a local phenomenon that runs via the organism's intrinsic properties, so causally relevant properties must be intrinsic or locally supervenient as well. Since ordinary ("wide") content is causally irrelevant by (i) and (iii), psychology should appeal to some non-ordinary ("narrow") content that is locally supervenient. (Fodor, 1987, ch. 2; 1991a)

Given Burge's view of causal relevance the debate boils down to the following question. Are properties that fail to supervene locally sometimes necessary for description of, e.g., intentional psychological patterns? Burge's answer is a clear "yes." This of course rests on his famous arguments for externalism about mental content (Burge, 1979), but he also thinks that externalism[48] is required for pattern description in physiology (the heart-example), geology (continental drift) and social sciences (interactions between persons). (Burge, 1989) Generalizing, this Burgean attitude can be captured as follows:

> PACE3 *Freedom of Property Kind*. Different sciences may use different kinds of properties in their causal explanations; in particular causally relevant properties may or may not supervene locally.

4.3.4. (P4) and the Call for Psychological Laws

Finally, consider Burge's argument against the need for psychological laws, which follows the same pattern. According to the Humean orthodoxy causation requires that there be some regularity or law relating cause and effect. Some philosophers, like Davidson (1980), think this regularity must be a strict law. Since there arguably are no strict psychological laws, Davidson suggested that mental events must be redescribed in the vocabulary of physics, which he assumed *does* allow for the formulation of strict laws. But it is at best unclear whether this strategy can live up to his critics' demand that

---

[48] The reference to environmental factors may not be explicit, as Burge is eager to emphasize. What is meant is that the environment plays an essential role in individuation, i.e., in determining what the relevant entity *is*. (Burge, 1989, p. 313)

mental causation must be causation *in virtue of* the events' mental properties. (See the papers on Davidson in Heil & Mele, 1993.)

Burges' critique is a different one. He accuses Davidson of using *a priori* reasoning outside its proper domain. "I think that we do not know, and cannot know a priori, that causal statements entail the existence of laws or explanatory systems that have such specific properties." (Burge, 1993, p. 112) Elsewhere (1992) he makes it even clearer against Davidson that questions about the presence and nature of laws in causal explanations are empirical through and through:

> I do not think it *a priori* true, or even clearly a heuristic principle of science or reason, that causal relations must be backed by any particular kind of law. I think that we learn the nature and scope of laws (and the variety of sorts of "laws") that back causal relations through empirical investigation. It is not clear that psychophysical counterfactual generalizations – or nonstrict "laws" – cannot alone "back" psychophysical causal relations. (Burge, 1992, p. 35, see also his 1989, p. 318)

This open attitude to the question of laws is perhaps all for the best, given the apparent paucity of strict laws in the sciences (Woodward, 2000), and given that many scientific explanations appear to depend more on descriptions of mechanisms than they do on laws. (Machamer et al., 2000, Woodward, 2002) Anyway, it seems clear that Burge dismisses the call for psychological causal laws because it is not motivated by reference to psychology itself. In general:

> PACE4 *Freedom of Explanatory System*. Different sciences may apply causal explanatory systems of different kinds; in particular it is not necessarily the case that all causal explanations be backed by strict laws.

Summing up, Burge's attitude to causal explanation is pluralistic. Depending on the patterns which are to be described causal explanations in different practices may vary along the axes of (PACE1-4). In particular, he thinks it is far from clear that causal relevance in psychology must be constrained by (P1)-(P4). In fact he thinks that some of these constraints – like (P3) and the call for narrow mental content – are misguided. His

arguments against the necessity of constraints (P1)-(P4) all follow the same pattern. Many physicalists are concerned with finding out *how* such constraints can come to be satisfied. But Burge thinks it is not clear that the constraints can be motivated by reference to psychology and the patterns of events it takes an interest in, and accordingly it is not clear that the corresponding problems of mental causation are *real* problems. As I said above, considerations of Burge's explicit arguments against physicalists like Davidson, Fodor and Kim lend support to my interpretation of his theory of causal relevance. Given this view of causal relevance, and the Principle of Responsible Practices, it is not clear whether the "how"-questions the constraints give rise to are well-motivated. If this interpretation is right, his lack of patience with "how"-questions would appear less abrupt.

4.4. Causal Relevance in Psychology and Neuroscience

With this general diagnosis of Burge's position in hand we see that it is not an efficient response to Burge to dismiss his dismissiveness, and insist that the "how"-questions raised by (P1)-(P4) are interesting in their own right. If we are to counter Burge on his own turf, what is called for is rather a motivation of these problems from within explanatory practice. In my response to Burge I shall focus on (P1) and (P2). I argue that these constraints can be motivated from within an explanatory practice that is relevant to understanding mental causation, namely neuroscience. In this enterprise, mental causation is understood as involving underlying mechanisms, and some tight relation between mental and physical causes is needed. Burge's arguments against (P1) and (P2) depend on treating the explanatory interests in mental causation and mechanisms as strongly independent. But in neuroscience these interests are combined. In particular, given what neuroscience tells us, I contend that part-whole relations *are* relevant to bridging psychology and neurophysiology. To this end I first offer descriptions of the "neural" and "mental" patterns of events in Burge's example of the man running to the pharmacy, and extract an account of what kinds of events are causally relevant in these patterns. Second, I consider Patricia Goldman-Rakic' theory of prefrontal cortex to show that neuroscience does take an interest in rational behavior, and not just in mere bodily movements.

Since (P1) and (P2) can be motivated by reference to explanatory practice, it is at best unclear whether Burge can dismiss them as unwarranted without stipulating that neuroscience is not a relevant explanatory practice when it comes to understanding mental causation. But which practices are relevant to mental causation is, I contend, itself an *empirical* question.

4.4.1. Neural and Psychological Patterns of Events

Let us now analyze Burge's case of the neurophysiological and the psychological pattern of events both of which lead to a man's running to the pharmacy in more detail. This will provide us with a further illustration of patterns of events and causal relevance, and tell us what kinds of properties are causally relevant in psychology and neuroscience. This should interest us presently, as it will turn out that rational/cognitive properties are causally relevant in many neural explanations. Perhaps most importantly, we will be in no position to assess Burge's claims that the patterns are very different and that we should feel no obligation to relate them *unless we have some grasp of what they look like.* Accordingly, there is no way round looking at psychology and neuroscience to see what kinds of patterns they describe. I am aware that these are perilous and difficult grounds for a philosopher, but the alternative is in my view even riskier. Merely quantifying over physical properties and giving them names like "P" is hardly illuminating when what we are investigating is the nature of physical explanations. Burge would surely agree. I emphasize, though, that since he does not describe the patterns he is referring to it is impossible to tell whether the following story is the one he had in mind.

I want to start by setting aside some of the subtleties originating in the philosophy of action while offering a rough description of the *psychological pattern*, framed at the level of intentional psychology. I shall be relatively brief here, since I am drawing on conventional philosophical wisdom, and since philosophers of mind have typically devoted more time to intentional than neuroscientific explanations. Consider the following psychological pattern. The man's *perceiving that* his son is ill, and his belief *that* people who are ill normally benefit from medicine, cause him to *recognize that* he has to lay his hands on the proper type of medicine if he is to relieve his son's sufferings. In conjunction with his *wanting* to help his son, and his *believing that* running to the

pharmacy would be more efficient than waiting for the doctor, this event causes the man to *decide that* he should run. That decision in turn causes the actual running.

The philosophical significance of this description is that it makes us see the running as the *rational* outcome of a pattern of events involving *cognition* and perhaps also considerations of *normative constraints* to the effect that one *should* help ones nearest and dearest. Arguably, a great number of psychological explanations express the same interest in patterns of events that exhibit (or fail to exhibit, as the case may be) rational and normative features.

Now, which properties are causally relevant to folk-psychology and the parts of academic psychology that study this kind of deliberation and action? That is, which properties are such that citing them facilitates the description of, e.g., the running as the outcome of a rational pattern of events? We can say at least that the properties to which the psychological verbs I italicized above refer must be causally relevant. After all, these are building blocks of propositional attitudes such as believing *that* p, desiring *that* q etc. Propositional attitudes are in turn typically taken to be part of the *explanans* in action explanations. More importantly, it is relatively uncontroversial that the verbs have the desired features of rationality and normativity. (Davidson, 1980) It is hard to spell out exactly what these features are and why the verbs must exhibit them. But there is no special mystery about how these descriptions single out properties that are causally relevant in the study of rational and irrational patterns of events.

More controversially, philosophers like Davidson (1980) and John McDowell (1994) claim that the vocabulary of the physical sciences inevitably fails to capture the rationality-features, so rationality-talk is conceptually irreducible to physical-talk. Let us acquiesce in this, at least for the time being, as it provides an illustration of causally *irrelevant* properties. If we cite *only* physical properties in our explanation of the running we will not be able to describe the relevant pattern. For the pattern will be *non-rational* rather than rational or irrational. Rationality simply does not apply to it. (Or so Davidson and McDowell would probably claim.)

Secondly, we shall need at least a philosopher's sketch of the *neurophysiological pattern* of events. Here I base my exposition primarily on text book introductions to the neuroscience of voluntary movement (Bear et al., 2001, ch. 14; Kandel et al., 2000, ch.

19 & 33; Rains, 2002, ch. 9 & 12). I shall turn to the details of one part of the pattern below when considering Goldman-Rakic' view of prefrontal cortex.

The first thing to notice about our pattern is that neuroscientific analysis is much more *fine-grained* than the corresponding folk-psychological explanations. This is to be expected, insofar as neuroscience studies the control of movement at several levels, ranging from *strategy* (the goal of movement and choice of movement strategy) and *tactics* (how the goal is to be reached by way of spatial and temporal coordination of muscle contractions and relaxations) to execution (direct causing of contractions/relaxations). Psychology, or at least folk-psychology, on the other hand, is almost exclusively concerned with strategy. Within Burge's framework, however, this difference does nothing to lessen the explanatory value of folk psychology as such value must always be measured relative to the explanatory aims of a given explanatory practice. For most folk-psychological purposes, a coarse-grained analysis is sufficient to describe the relevant patterns of events. If what we are interested in is *why* people run, we can leave the explanation of *how* they run to neuroscientists.

On the other hand, one might have thought that that neuroscience is not concerned with how goals, wants and strategies control behavior, and that it is therefore really a different and independent explanatory game. On closer inspection this turns out to be a mistake. Neuroscience does take an interest in this kind of control of behavior. In this respect, I contend, neuroscience and folk-psychology are on the same explanatory ground. This warrants some skepticism toward Burge's claim that psychology and neuroscience "answer two very different types of enquiry." (Burge, 1993, p. 116) Of course, the enquiries are different in *some* sense, but whether that sense serves to mutually insulate their respective explanatory aims is just the point at issue. It cannot be assumed.

A second thing to notice about the pattern is that neuroscience aims to describe how control of movement *goes on* as the action, e.g. the running, is carried out. Folk-psychology typically appears to treats actions as simple two-step causal processes. An intention is formed, a bodily movement performed. Once again, this can be construed as a minor short-coming given folk-psychology's explanatory aims. What is interesting, however, is the overlap in the explanatory interests between neuroscience and

psychology, so that *prima facie* it seems possible that explanatory competition or exclusion may arise unless (P1) is satisfied.

With the distinctions between tactics, strategy and execution in hand we can turn to the pattern of events that neuroscience tracks. While superficial and simplified in many ways, my description of this pattern nevertheless includes some details, for two reasons. First, I want to emphasize that the idea of a "neural pattern of events" is not an off-hand appeal to future idealized neuroscience, and that there is significant evidence for what we think we know about this pattern. (On the other hand, I should not be taken to imply that we have anything like a complete neurophysiological explanation of human behavior, that current models are immune to revision, or that the neural pattern of events could replace the psychological pattern just described.) Second, I want to bring out the *nature* of the neural explanations at hand. Nevertheless, some readers may wish to read through these paragraphs relatively quickly.

(1) *Perception* of course comes first, and also plays a crucial role throughout the action. But it is not our primary interest here. So let me just say that it involves the extraction of *inter alia* visual information and plays a role in the formation of the man's belief that his son is ill.

(2) The *strategic* part of the pattern is, on the other hand, highly relevant to us. This higher-level of control involves the goal of movements, and the choice of movement strategies to be used to reach that goal. Structures that are central to this level of control include the prefrontal cortex (PFC) and the basal ganglia. PFC receives highly processed information about the world from the association areas for the five sensory modalities, and possibly also information about the organism's motivational state from limbic structures. Furthermore, it can probably access long-term memory via its connections with hippocampal regions. These information-bearing connections alone suggest that PFC is in a position to integrate knowledge and motivation in the planning of strategies to reach relevant goals. This view of its function is supported *inter alia* by lesion studies, which show that damages to PFC can lead to failure in the rational guiding of behavior by motives and plans. We shall return to this in the next section. For now, let us just say that PFC is involved in the man's decision to run.

(3) Our example of the man's running involves complex sequences of movements, so *tactical* areas of the brain must be involved in controlling *how* these movements are to be coordinated in space and time and in preparing the motor system for movement. This level of control may appear irrelevant for psychological purposes, but I have included it to emphasize that neuroscience, in apparent contrast with psychology, offers fine-grained and multilevel models of the control of bodily movements. Downstream from PFC goals and strategies seem to affect such areas in an indirect, roundabout way. A cortical-subcortical-cortical loop, running via the basal ganglia connects PFC strongly with the premotor (PMA) and the supplementary motor (SMA) areas. Several lines of evidence indicate that PMA and SMA areas play this kind of tactical role. Anatomically, they receive input from PFC via the basal ganglia, as just mentioned. They are also connected with parietal regions involved in storage of motor engrams, so they can probably access pre-programmed information about movement sequences. In their turn, PMA and SMA project to primary motor cortex (M1), which is the cortical region most directly involved in causing muscle contractions. Such anatomical considerations are supported *inter alia* by measurements of regional cerebral blood flow (rCBF) in human subjects during simple and complex movements. The measurements demonstrate increased activity in SMA in addition to M1 and primary somatosensory cortex (S1) during complex tasks, whereas the increase in activity during the simpler tasks is limited to M1 and S1. Microstimulation of PMA and SMA typically causes coordinated movements at more than one joint, whereas stimulation of M1 typically yields only simple one-joint movements. And while it is not entirely clear how the tactical function of PMA and SMA is realized by representations at a cellular level, single cell recordings shed some light on this. Individual neurons in monkey PMA display activity that (a) increases during periods where the monkey has received information about the direction of movement, but is waiting for an external "go" signal; (b) is normal or absent during the actual movement; and (c) is selective for the direction of the planned movement. In other words, some neurons are active prior to a movement to the left, but not prior to movements to the right, whereas the converse holds for other neurons. This kind of representational hypotheses formulated at the level of tactics should

interest us presently. It suggests that even at this level, neuroscience may not be completely a-cognitive.

Whatever the coding-scheme of PMA and SMA, M1 seems to some extent to take over as the most direct cortical cause of most movements. M1 is "somatotopically" organized, that is, it forms a kind of body-map. Stimulating different regions within it yields movements of different limbs. As for its coding-scheme it seems to represent *movements* and not individual muscle contractions, which means that M1-neurons affect *groups* of muscles. Specifically, M1 does not seem to code for displacement of the limb, but rather for *force* and *direction* of movement. Single cell recordings reveal that the firing rates of M1-neurons increase when we add weights in resistance of movement (making more force necessary), but decrease when weights facilitate the movement (making less force necessary). So at least to a first approximation we can say that force is proportional to firing rates.

In contrast, the coding-scheme for direction of movement appears to be an exquisite example of distributed and partially superposed representation, i.e., the relation between neurons and the movements coded for is *many-many*. First, recordings reveal that *many* neurons fire in the case of movements in a given direction. More intriguingly, the same neurons are involved in *many* directions of movements. They are, however, tuned to a preferred direction of movement – in the case of which they fire maximally – with the firing rates diminishing as the actual direction of movement deviates from the preferred direction. Each neuron thus "votes" for its preferred direction of movement, with its firing rate determining the weight of the vote. The activities of all neurons in the population thus add up to yield the actual direction of movement.

(3) The lowest level of control, execution, is concerned with the direct causal control of muscle fibers. This is achieved by the spinal cord and brain stem, which enervate the muscles. The story does not end here, however. As already mentioned, neural control of movement goes on as complex actions are carried out. For instance, at the strategic level – the level describing how a movement sequence is to be carried out – the cerebellum plays a regulatory role. Lesions to the cerebellum cause disruptions in the coordination and flow of complex movements, and sufficient amounts of liquor will yield similar effects. (The cerebellum is highly sensitive to alcohol.) Anatomical considerations

also support this view of the cerebellum. It receives input from sensory areas and from PMA, SMA and M1 and projects back to these motor areas. Planned movements can thus be compared with actual movements, and appropriate corrections can be signaled back to motor areas.

What is the philosophical significance of this sketch? What does it tell us about the nature of the "neural pattern of events"? In summary, our *simplified neuroscientific pattern involves an ongoing process of perception, strategy, tactics and execution*. In particular it is noteworthy that there are actually *several* patterns of events at various levels of descriptive granularity. The causing of the man's movement can be described as (a) the outcome of the interaction of several neural systems, the functions of which are described in a representational/information processing language. (Notice again that already now the distance between psychological and neuroscientific patterns appears to have diminished somewhat.) Parts of the pattern are also describable at (b) the neural circuit level, say, in terms of the distributed coding of M1 instructions. Finally, etiological characterizations can be produced at (c) the cellular/molecular level in terms of events that constitute the causing of individual neurons to fire. (I.e., a pattern of events involving events like ion channels opening, membrane potential reaching threshold, the releasing of neurotransmitter into the synaptic cleft etc.) Our pattern thus fits well with Kenneth Schaffner's (1993, ch. 6) view of neuroscientific models as multi-level, or as spelled out in terms of events ranging from tissue to molecular levels.

As we might have expected, the whole thing turns out be quite complicated, even when simplified for philosophical purposes, which, of course, is not to say that it does not offer a promising model of behavior. Talk of a "neural pattern" of events leading up to bodily movements is not an off-hand appeal to future or idealized neuroscience. Now, Burge is concerned with what kinds of properties are causally relevant to the explanation of bodily movements. Given this sketch of the neural pattern of events I think we can extract an account of what broad kinds of properties are causally relevant in the neuroscience of behavior. First, I take it that a central aim of many neuroscientists who take an interest in cognition and behavior is to discover biological *mechanisms* that show how the components of the neural system *produce* – in a suitably robust physical sense – behavior and mental phenomena ranging from thoughts to after-images. Furthermore,

many of them aim to do so at increasingly lower levels of complexity. Aptness for figuring in mechanistic explanations, then, is arguably a central constraint on causal relevance in neuroscience.

But for present purposes it is just as important to note that this theoretical interest in mechanisms is for many neuroscientists *combined* with an interest in cognition and rationality. In the following quote, aimed at a popular audience, neuroscientist Goldman-Rakic expresses just this combination of interests:

> Until recently, the fundamental processes involved in […] higher mental functions defied description in the mechanistic terms of science. Indeed, for the greater part of this century, neurobiologists often denied that such functions were accessible to scientific analysis or declared that they belonged strictly to the domain of psychology and philosophy. Within the past two decades, however, neuroscientists have made great advances in understanding the relation between cognitive processes and the anatomic organization of the brain. As a consequence even global mental attributes such as thought and intentionality can now be meaningfully studied in the laboratory.
>
> The ultimate goal of that research is extraordinarily ambitious. Eventually researchers such as myself hope to be able to analyze higher mental functions in terms of the coordinated activation of neurons in various structures in the brain. (Goldman-Rakic, 1992, p. 73)

As John Bickle (2003, p. 3) emphasizes, Kandel et al.'s (2000) influential text book, *Principles of Neural Science*, expresses a similar interest in cognition *and* mechanisms. Such methodological manifestos by neuroscientists appear to be radically at odds with Burge's view that neuroscientific explanations of behavior are very different from, and make no essential assumptions about, psychological explanations of that behavior. (Burge, 1993, pp. 115-116) To a certain extent – given neuroscience's interest in strategy – this criticism of Burge remains valid even when psychology is construed as *folk*-psychology.

But so much for the ambitions of neuroscientists. Are their methods up to their explanatory aims? Perhaps the methodology Kandel et al. recommend does in fact involve what Davidson (1980) aptly called a "change of subject"? Will we inevitably lose the mental/cognitive perspective by looking for cellular and chemical mechanisms? Does the attempt to "understand the mental processes by which we perceive, act, learn and

remember" (Kandel et al., 2000, p. xxxv) in terms of biological mechanisms of necessity become the explanation of something else? Again, someone like Eric Kandel would deny this. His Nobel Prize lecture tells us that while beginning his groundbreaking memory research in the late Fifties he felt dissatisfied with the psychoanalytical view of the brain as a black box.

> From the beginning, my purpose in translating questions about the psychology of learning into the empirical language of biology was *not to replace* the logic of psychology or psychoanalysis with the logic of cell and molecular biology, but to try to *join* these two disciplines and to contribute to a *new synthesis* that would *combine* the mentalistic psychology of memory storage with the biology of neuronal signaling. (Kandel, 2000, p. 393, my italics. See Kandel, 2006, for more details on his "synthesis" of psychology and biology.)

Presumably a Davidsonian explanatory "change of subject" takes place just in case a change of explanatory interests, e.g., from "rational patterns" to "productive patterns," takes place. But Kandel and, as we shall see, Goldman-Rakic, are quite explicit that their explanatory interest have not changed; despite the added interest in productive mechanisms they are seeking the same quarry as the psychologists. Most importantly, they have been quite successful. A philosopher once told me that neuroscience can only trace causes of bodily movements to the periphery of the brain (to the motor cortex?), so psychology is needed to find earlier causes of behavior. But given the above description about the neuroscience of behavior, this kind of claim seems dubious. We know a great deal about how neural systems upstream from motor cortex play a role in the strategy and tactics of movement. We shall see that Goldman-Rakic' results add to the evidence that neuroscience is relevant to matters psychological.

Apart from this kind of success in discovering cognitive mechanisms, there is another reason why even cellular/molecular neuroscience – let alone higher-level cognitive neuroscience – does not necessarily abandon the psychological perspective. It seems to me that often, e.g., in the pattern description we are considering, an interest in productive mechanisms is combined with an interest in viewing these mechanisms as ways of *processing information*. There is at any rate much (admittedly unsystematic) talk in neuroscience of "signaling," "representing," "encoding," "processing" and

"transference" of information. But this is the very perspective that has been prominent among (cognitive) psychologists ever since they bid farewell to behaviorism. Similarly, this broad perspective is shared with philosophical attempts to naturalize mental content. (E.g., Dretske, 1988)

Notably, the link between an interest in cognition and the methodology of adopting an information processing perspective is not arbitrary. If the *explanandum* in neuroscience really is to be ultimately *cognitive*, as Goldman-Rakic and Kandel suggest, notions like information and its cognates are experimentally useful, perhaps even essential. Neuroscientific hypothesises about mechanisms are frequently formulated in terms of, say, possible coding-schemes that neural systems might utilize. Whether the information talk is ultimately simply an exercise in heuristics, I do not know. But I think we can safely say that to the extent that neuroscience is concerned with events like parents running to pharmacies, it is also interested in viewing these events as the outcome of (a) *productive biological mechanisms* in the neural system that (b) involve *information processing*.

Having described neuroscientific patterns of events we must proceed to track down properties that are causally relevant to this kind of pattern. Interestingly, what we find will depend on the level of descriptive granularity at which the pattern is described. If we remain at the cellular-molecular level the relevant properties will be properties like those that affect whether a given neuron will fire, and at what frequency it will fire, or properties that affect neural plasticity ("synaptic strength"). Accordingly, causally relevant properties will include the connectivity among neurons, the presence and number of ion channels of various sorts, the amount of neurotransmitter released, the rate of neurotransmitter reuptake, the various genetic and biochemical factors that can affect synaptic strength and so on and so forth. These causally relevant properties seem increasingly to be describable purely in the language of biochemistry. (Or so Bickle (2003) argues.)

What about higher levels like cognitive neuroscience and the study of how, e.g., different neural systems interact? As we saw in our brief sketch of a pattern, talk about representations of goals and plans etc. is prominent at this level of discourse. Hence, relative to *this* level, straightforwardly intentional properties come out as causally

relevant according to Burge's criterion. This should interest us in the present context. One might have thought that the *only* concepts relevant to describing the neurophysiological pattern would be rendered in a purely chemical, non-cognitive language. But it seems clear that cognitive/psychological language *is* "Burge-relevant" to the causal explanation of the neuroscientific pattern, given the current state of neuroscience. The significance of this is that neuroscience is not as independent of psychology as Burge appears to envision. Neuroscience aims to understand cognitive processes, and how they are implemented neurally, and invoking mechanisms is a crucial part of this enterprise.

4.4.2. Patricia Goldman-Rakic' Theory of Prefrontal Cortex

Throughout this discussion I have suggested grounds on which Burge's treatment of psychology and neuroscience as being strongly independent could be challenged. Here I offer a more detailed example which demonstrates that this treatment is in fact deeply problematic, by considering the theory of prefrontal cortex (PFC) that was developed by the late neuroscientist Goldman-Rakic and her co-workers.[49] This line of research is a natural example, because it represented a major breakthrough in the study of higher cognitive functions, and because even many neuroscientists at one point thought PFC and its role in higher cognition to be beyond their explanatory powers. (Arnsten, 2003) We have already seen that PFC is critically involved in the strategic control of behavior. I aim to show that (i) Goldman-Rakic' work combines an interest in cognition and rationality with an interest in underlying productive mechanisms at the cellular and sub-cellular levels; and that (ii) part-whole relations are critical in understanding how these mechanisms give rise to cognitive processes. Thus her work is an example of an explanatory practice that is interested in rationality and yet respects constraints (P1) and (P2). This renders Burge's dismissal of (P1)-(P2) implausible, for reasons he himself should accept.

---

[49] For an exposition of much of the same work with a view to consciousness, see Bickle, 2003, ch. 4.

Lesion studies going back at least to the famous case of Phineas P. Gage in 1848 (Damasio, 1994) tell us that damage to PFC can lead to a confusing variety of cognitive and motivational disorders. Different subjects with PFC lesions display a multitude of symptoms including apparent lack of motivation, deficits in control of motivation, and inability to carry out plans to reach goals. (For an overview see Rains, 2002, ch. 12.) For instance, subjects with PFC deficits may seem apathic and unable to initiate action, or they may act uncritically and in socially unacceptable ways on whims of the moment. In their attempts to reach goals they may initiate procedures that would have been effective, only to become distracted and initiate different and unrelated procedures instead. These deficiencies may be due to a general dependence on environmental cues for determining what to do, rather than relying on plans that remain stable in spite of environmental changes. ("Environmental dependency syndrome," LHermitte (1986).) Furthermore, evidence from *inter alia* imaging studies implicates PFC in the cognitive deficits characteristic of diseases like schizophrenia and Parkinson's. What this shows is that failure of PFC functioning can cause failures in *rationality, motivation* and *agency*, which belong in Burge's rational/cognitive patterns of events. These deficits have inspired several theories of PFC-functioning, of which Goldman-Rakic' is a prominent example. (I am in no position to assess its merits relative to other theories, but there can be little doubt that it represents a mainstream theory that combines an interest in cognition with an interest in productive mechanisms, and that is strictly speaking all the present argument requires.) According to Goldman-Rakic' hypothesis the essence of PFC-function is the "regulation of behavior by representational knowledge". (1987, p. 374) In an often used metaphor, PFC keeps information "on line" in the absence of a direct information flow from the environment. She therefore appealed to what many cognitive psychologists, like Alan Baddeley (2003), call "working memory," or metaphorically the mind's "blackboard" or "sketch pad" in several later papers on PFC. Now, if we can take this representational language at face value Goldman-Rakic' explanations seem straightforwardly cognitive and rational. True, much PFC research has targeted lower-level psychological phenomena like spatial working memory. And it is by no means obvious that the most fruitful notion of representation for neuroscientific purposes must be propositional or sentential representation, though that is typically taken to be the

representational notion implicit in *folk*-psychology. So it may very well be that mature neuroscientific explanations of behavior and folk-psychological explanations answer "two very different types of enquiry" (Burge, 1993, p. 116) insofar as they differ along the sentential/non-sentential axis. But even the "eliminativist" non-sententialist view does not entail that the neuroscientific *explanandum* is *non*-representational, *non*-cognitive or *non*-rational. To reach that conclusion we would have to follow Davidson (1980) who seems to reserve terms like "rationality" and "cognition" for sentential modes of representation. Given this restricted use of rationality, adapting a non-sentential notion of representation *will* count as a change of subject. However, this is probably not an option available to someone like Burge, who must insist on individuating cognition as cognitive scientists do if he is to retain the primacy of explanatory practice.[50] Also, he should have some sympathy with the kind of argument that Patricia Churchland urged already 25 years ago. Insisting on a sentential mode of representation for philosophical reasons may actually hinder progress in neuroscience. (Churchland, 1980) Returning to the relevance for personal level psychology, we should note that Goldman-Rakic, for one, was unwilling to rest content with the simple and experimentally well-behaved case of spatial memory. Rather she speculated that "the evolution of a capacity to guide behavior by representation of stimuli rather than by stimuli themselves introduces the possibility that *concepts* and *plans* can govern behavior." (Goldman-Rakic, 1987, p. 378, my italics.) This is exactly what we want from rationalizing action explanations. In fact, it may no longer be mere speculation. Quite recently (Genovesio et al., 2005) claims have been made about PFC neurons that appear to represent which strategy an experimental animal is currently using, much in the same way as other PFC neurons appear to represent spatial locations, as we shall see shortly. Even more intriguingly with respect to neuroscientific

[50] Even for Davidson, adopting the restricted notion of cognition/rationality comes at a high price. Most, if not all, animals, and even prelinguistic children, would be deprived of their status as cognizant, leaving mature humans as the only thinking creatures. That, however, is arguably implausible from a biological and evolutionary point of view. Furthermore, the notion of *cognition* would be at odds with that generally used in *cognitive science*.

explanations of actions, it has been suggested that PFC neurons might encode the subjective value of offered and chosen options. (Wallis, 2006)

The cognitive deficits discussed so far are best measured by behavioral tasks like delayed response tasks (DRT) and the Wisconsin Card Sort Task (WCS) that require working memory. I shall only consider DRTs here, since they have been instrumental to the discovery of neural mechanisms. In a classical DRT, devised as far back as the Thirties, a monkey is first showed that a food reward is varyingly placed in one of two possible wells and covered by a lid. A screen is then lowered for a given time, thus disrupting the sensory input from the wells to the monkey. When the screen is raised after the delay the monkey must respond by choosing the correct well, based *solely* on its internal representation in spatial working memory. (E.g., "this time the food is hidden in the left well.") Monkeys with appropriately induced PFC lesions are severely impaired in this task. But they succeed in alternative tasks where they can rely on associative memory plus environmental cues present at the time of choice. (For instance if the food is *always* hidden in the *left* well, the monkey can rely on an association between left and food.) These experiments therefore demonstrate that simple associative memory is singly dissociable from working memory.

Given the simplicity of the DRT-task it might seem like the difference between the two types of memory is relatively trivial, and it may be hard to see what bearing working memory has on *rationality* and *agency*. In view of certain characterizations of PFC as "the region of the brain that is most essentially related to who we are, both as human beings and as individuals" (Rains, 2002, p. 378), one might expect its essence to be phrased in more suggestive terms than the seemingly dull "working memory." However, working memory appears to me to be an extremely potent notion in accounting for our agency and relative freedom. To see this, note that, if Goldman-Rakic' view of PFC is along the right lines, schizophrenics and subjects with PFC lesions may metaphorically be said to *be hostages of their environment*. To the extent that they are limited to using associative memory – which not improbably is evolutionarily older than working memory – they are also dependent on potentially random cues from the environment for initiating action. (After all, even simple creatures like fruit flies and sea slugs are capable of learning rudimentary associations between stimuli and response.)

Indeed, Goldman-Rakic speculated that "the brain's working memory function, i.e., the ability to bring to mind events in the absence of direct stimulation, may be its inherently most flexible mechanism and its evolutionarily most significant achievement." (Goldman-Rakic, 1995, p. 483) Perhaps, then we could view failure in working memory as one type of *failure or breakdown in freedom of the will.* Obviously, having a properly functioning PFC will not make me free in any libertarian sense, but it might still be a vital part of what provides me with the limited kind of freedom that is likely to be available to biological creatures like myself.

Goldman-Rakic' use of neuroscience to rethink the nature of schizophrenia is another interesting illustration of an attempt to illuminate cognition by neuroscientific means. Methodologically she seems to have started from the above psychological characterization of PFC function, found support for this in lower-level neuroscience (see below), and noted that symptoms of schizophrenia are *psychologically* similar to those resulting from PFC lesions and that schizophrenia is also at a *neuroscientific* level associated with certain PFC deficits. This lead her to suggest that: "If, as these [neuroscientific] findings suggest, the prefrontal cortex is centrally involved in schizophrenia, perhaps we can begin to think of this disorder as comprising a breakdown in the processes by which representational knowledge governs behavior." (Goldman-Rakic, 1987, p. 404) Thus, neuroscience and psychology *combine* in an attempt to reconceptualize at least certain aspects of a psychological disorder that has been notoriously hard to define.

So far, I have only considered correlational, higher-level evidence in the shape of lesion and imaging studies to argue that both psychology and neuroscience take an explanatory interest in rationality, and that methodologically they are interdependent. This casts some doubts on Burge's claim that they correspond to very different explanatory purposes (Burge, 1993, p. 116), and the assumption that the explanations must be somehow tightly related as demanded by (P1) does not seem that unreasonable. However, all this is in perfect accordance with Burge's claim that "there are surely some systematic, even necessary, relations between mental events and underlying physical processes." (Burge, 1993, p. 116)

To counter Burge I therefore need to show that neuroscience offers explanations that combine productive mechanisms with an interest in mental causation, and that part-whole relations are involved in offering mechanistic models for mental causation. I attempt to achieve this by examining relevant PFC research as it proceeds from behavioral studies to discovering productive mechanisms. It is interesting to note that Goldman-Rakic in a popular scientific paper took care to make the following methodological comment. "Whole-brain studies tell us only part of the story; to understand the details of how signals pass to and from the prefrontal cortex, one must scrutinize the brain on a cellular scale." (Goldman-Rakic, 1992, p. 77)

A breakthrough in this lower-level search for a mechanism came with the use of single cell recordings in PFC and the discovery in the Seventies that cells in the principal sulcus of the dorsolateral primate PFC instantiate so-called "memory fields." These fields are most conveniently studied in an oculomotor (ODR) version of the DRT. Here, the monkey is surgically prepared for single-cell recordings with its head fixated as it watches a TV screen. It is trained to maintain fixation at a spot at the center of the screen while a stimulus flashes somewhere the periphery (e.g. at 0º or 45º) and then disappears. The monkey's task is to saccade – i.e., move its gaze quickly – to the remembered location where the stimulus appeared, but only when he receives a signal to do so. This signal is the disappearing of the central spot after a delay of several seconds. If the monkey responds correctly by saccading to the location of the target after the delay, it receives a small liquid reward. (Motivation is ensured by prior dehydration of the monkey.) Importantly, the absence of the target during the delay period requires the monkey to utilize an internal, working memory representation of the target location.

Recording from PFC neurons in the behaving monkey yields intriguing results. There are neurons whose firing rates (i) *increase* during the delay-period prior to a saccade to targets located at a specific direction (say 0º) or nearby directions and (ii) *decrease* prior to saccades to the opposite of this "preferred" direction (say 180º). In other words, these neurons are *spatially tuned* to a memory field consisting of the location(s) for which they fire maximally. Strong evidence indicates that these neurons are at least part of "the cellular basis" (Williams & Goldman-Rakic, 1995, p. 572) or "cellular correlates" (Goldman-Rakic, 1995, p. 477) of spatial working memory. For instance,

failure in making the correct saccade is invariably correlated with failure in maintaining neural activity during delay. (Williams & Goldman-Rakic, 1995) In this (admittedly highly controlled) laboratory setting it is therefore possible to *predict whether the monkey has forgotten the target location or not, simply by measuring the activity of* individual *neurons*.

This is a substantial result in its own right, and strongly suggests that these neurons play a role in representing memory locations. But most importantly, PFC research is beginning to reveal the mechanisms by which neurons give rise to working memories. So our *correlational* story is currently being supplemented with an equally interesting *how*-story, of which I shall briefly consider just one aspect. The aspect I shall consider is this. What *sustains* the pattern of increased firing of neurons which prefer locations in the remembered direction, and inhibited firing of neurons which prefer the opposite direction during the delay? That is, how are temporally stable memory fields implemented neurally?

Several anatomical and physiological assumptions enter into Goldman-Rakic (1995) hypothetical model to explain this. (Cf. fig. 1.) First, it appears that pyramidal PFC neurons are organized column-wise. Pyramidal neurons with similar preferred directions are gathered in vertical cortical columns, with the preferred direction changing as one moves tangentially across cortex from column to column. Pyramidal neurons with *similar* preferred directions (e.g. 90º) in the same and different columns may be connected directly and reciprocally by horizontal excitatory connections. (The majority of pyramidal PFC neurons appear to use excitatory amino acids as their neurotransmitter.) Additionally, PFC contains non-pyramidal interneurons that also exhibit preferred directions. The preferred direction of these neurons is often the opposite of nearby pyramidal neurons. Pyramidal neurons with *opposite* preferred directions (e.g., 90º vs. 270º) appear to be connected *indirectly* and reciprocally via these inhibitory interneurons. This circuitry, where neurons are organized in what Goldman-Rakic calls "excitatory-inhibitory units" (Goldman-Rakic, 1995, p. 481), would neatly explain the opposite memory fields. A 90º pyramidal neuron forms excitatory connections with an interneuron, which in turn forms an inhibitory connection with a 270º pyramidal interneuron. For instance, increased firing of the 90º neuron during the delay after a 90º

stimulus would excite the inhibitory interneuron, which in turn reduces the firing of the 270º neuron. Simultaneously the reciprocal excitatory connections between 90º neurons would ensure increased reverberating activity of many of these neurons over time during the delay.

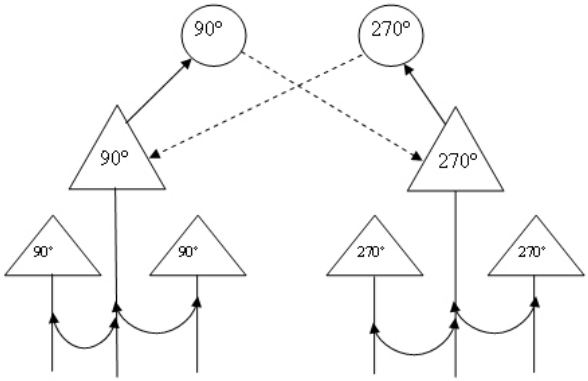

Fig. 1: Hypothetical circuitry that could maintain activity characteristic of working memory during the delay period. Only two cortical columns are included. Triangles = pyramidal neurons. Circles = non-pyramidal interneurons. Solid arrows = excitatory connections. Dashed arrows = inhibitory connections. Text indicates a neuron's memory field. (Adapted from Goldman-Rakic, 1995, fig. 5; Arnsten, 2003, fig. 5B)

This of course is no complete explanation of the mechanisms underlying spatial working memory, and the monkey's performance of the ODR-task. Another aspect, which has been the subject of much recent research, is how dopaminergic input to the pyramidal neurons contributes to memory fields. Investigations into this aspect have taken neuroscientists to the subcellular level, where various dopamine receptors are thought to play an important role. (Williams & Goldman-Rakic, 1995, see also Arnsten, 2003; Bickle, 2003, ch. 4)

What I have provided, then, is but a *sketch* of *one* aspect of *one* mechanism thought to underlie *one* cognitive function attributed to PFC. Nevertheless, it serves to underscore that current neuroscience offers promising models for how mental causation is implemented neurally. Realigning our gaze to the neural pattern of events described above, we can see how such mechanisms play a role in more complex control of

intelligent behavior, involving *many* mechanisms in *many* cerebral regions. The pattern of events leading up to the man's running takes place within a structure of interrelated neural systems, which is summarized by figure 2. (Importantly, this figure is a simplified adaptation of the figure neuropsychologist Dennis G. Rains (2002) uses to introduce the neurophysiology of voluntary behavior in his text book.)

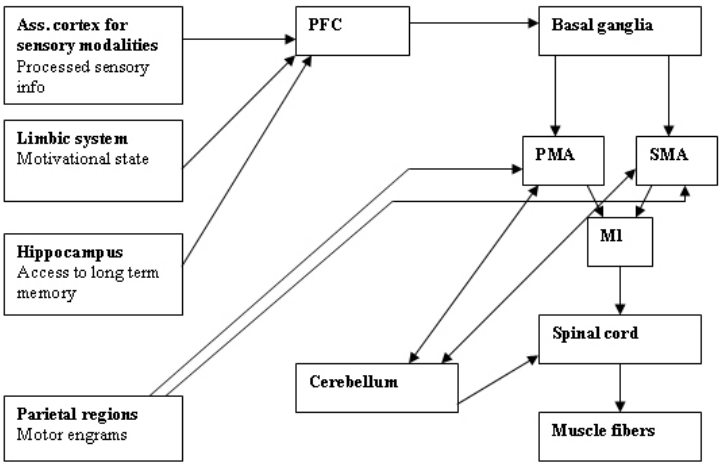

Fig. 2: Interactions of neural systems in the control of behavior. (Simplified and adapted from Rains (2002, p. 229 fig. 9.1)

4.4.3. Mental Causation, Mechanisms and Part-Whole Relations

Sketchy as they are, these descriptions of the neural pattern of events and the mechanisms thought to underlie spatial working memory have the desired features for a response to Burge. Recall that Burge is skeptical that a tight relation between mental and physical causes is needed to ground the causal efficacy of the mental. In particular he doubts that part-whole relations can be used to bridge neurophysiology and psychology. (Cp. (P1)) Furthermore, he is skeptical about the need for a mechanism in models for mental causation. Indeed, he takes it to be unclear whether *any* model for mental causation is needed. (Cp. (P2)) The rationale for this skepticism appears to be that he doubts whether (P1) and (P2) can be motivated from within relevant explanatory practices. My central

contention is that this doubt is only well-founded if we follow Burge in treating psychology and neuroscience as strongly independent. But as my examples illustrate, this treatment is in fact deeply problematic, because in neuroscience the explanatory interests in mental causation and in mechanisms are *combined*. Constraints (P1) and (P2) can arguably be motivated from within neuroscientific explanatory practice, and this practice *is* relevant to understanding mental causation.

First, with respect to the need for a mechanism, I contend that neuroscience offers putative explanations of *how* mental causation occurs in a physical world. It does so by offering models for mental causation that are mechanistic in nature. Setting Davidsonian qualms about attributing cognition to animals and worries about non-sentential modes of representation aside, we know the monkey saccades to the left because it *remembers* that the target appeared to the left and because it *wants* its juice-reward. This is mental causation. How does it happen? In neuroscience this question calls for an answer involving neural mechanisms. Goldman-Rakic' story about the mechanisms underlying working memory appears to be intended as a part of such an answer. Whether we as philosophers should interpret the answer as reductive or antireductive, I do not know. A reductionist might see a potential reduction here. She could say that by describing the neural mechanisms for mental causation we are simultaneously describing the neural process to which the mental causal process might reduce. An antireductionist would want to stop short of this, and claim that while the mental causal process involving working memory depends on, and occurs in virtue of, such mechanisms, it is not reducible to them. He would contend that the mechanisms are *underlying* mechanisms, perhaps endorsing some model of supervenient causation. But whoever is right, the call for a model of mental causation involving a mechanism arises naturally from neuroscientific explanatory practice. Indeed, to discover mechanisms in mental causation is a central part of neuroscience's explanatory aims.

Second, with respect to the call for tight relations between mental and physical causes, we know that the monkey's saccade has neural causes. The firing of PFC neurons which are presumed to have "left"-memory fields are likely to be prominent among these causes. Since we want to explain the causal role of the memory, it would seem that some tight relation between it and neural causes is required. The phrases "cellular basis"

(Williams & Goldman-Rakic, 1995, p. 572) or "cellular correlates" (Goldman-Rakic, 1995, p. 477) of working memory were presumably invoked partly for this reason. Again a reductionist might see in these neurons' firing activity a potential partial reduction base for working memory, whereas an antireductionist might see a potential partial supervenience base to be used in a model of supervenient causation. It seems, then, that mental causes must bear some tight relation like supervenience to neural causes if Goldman-Rakic' model is to help explain how working memory is possible and how it affects behavior.

Third, part-whole relations appear crucial to the mechanistic explanations we have been considering. Figure 1 offers a partial explanation of how stable memory fields could be maintained. But it is *only* explanatory because of what we know about relevant properties of the parts (e.g., that the relevant pyramidal neurons are excitatory) and how the parts are organized (e.g., how pyramidal neurons are interconnected via non-pyramidal interneurons). This does not explain how firing rates of pyramidal neurons might come to represent or *be about* spatial locations, of course. For that a story about representation, perhaps something along the lines of Dretske's teleological/functionalist account (1988, ch. 3) is required. But it does show that part-whole relations are required to explain how neural firing rates come to be reliably correlated with spatial locations. Such correlations are arguably a prerequisite for working memory representation in Goldman-Rakic' model. Furthermore, part-whole relations surface again in figure 2, which offers a model of how different neural systems interact to produce intelligent behavior. Here the properties attributed to the parts are partly *cognitive*. The figure allows us to see the role of structures like PFC in producing behavior, because we have qualified beliefs about what kinds of cognitive functions these structure serve, what kind of information they receive from other structures, and what kind of information they convey to other structures. But again, part-whole relations – and anatomical connections – are crucial to the explanation of how the structures interact in the production of behavior.

Pace Burge, mereology and mechanisms do seem to matter in low and high level neuroscience, even when the explanandum is cognitive. The importance of appeals to part-whole relations does not imply that mental events or entities are composed by physical events or entities, but it does show that the relations play an important role in

explaining how the brain serves cognitive functions. Burge can of course insist that these mereological models are "guesses as to what sort of relations may obtain." (Burge, 1993, p. 116) But they are certainly qualified guesses, guesses that are supported by empirical evidence, and that yield prospective mechanistic answers to questions like why specific lesions cause specific psychologically characterized deficiencies.[51]

The familiar questions about mental causation remain. Does, for instance, the Exclusion Argument require us to identify mental causes with physical causes? Or will some tight, but antireductive relation like supervenience suffice to rule out exclusion? Is supervenient causation just epiphenomenalism in disguise? Though it may be beneficial to consider these questions anew and more concretely from the point of view of neuroscience, it is not clear how they should be answered. I only want to maintain against Burge that they do not appear to be misguided questions. They arise naturally from reflections on explanatory practice. Considerations of the neuroscientific explanatory practice of which Goldman-Rakic' research is an instructive example appears to undermine Burge's doubts about (P1) and (P2) for reasons he himself should accept.

---

[51] Curiously, Burge may not mind these "guesses" because they come from scientists as opposed to philosophers. Georges Rey reports an interesting discussion he has had with Burge on cognitive psychology. "In correspondence, Burge has claimed to be opposed only to 'philosophical' and not to 'scientific' interpretations of physicalism and the 'computer' model [of the mind]." (Rey, 2001, p. 122n4) But if Burge were to take a parallel attitude to the questions we have been considering, his arguments would only threaten what we might call "armchair" physicalism. That is, as long as our endorsement of physicalist constraints on causal relevance is informed by science, Burge's presumed distinction between scientific and philosophical physicalism would collapse.

5. Paper (#3): Is there a Binding Problem of Behavior? E.J. Lowe on Causal Closure

ABSTRACT: The Causal Closure of Physics plays a central role in the Exclusion Argument for reductive physicalism, and even antireductive physicalists typically take it to express a strong physicalist claim to which they assent. However, E.J. Lowe argues that the principle is compatible with a dualistic view of mental causation, which few physicalists would accept. Mental events play an essential role in causing bodily movements, but this contribution is invisible from the point of view of neuroscience. He suggests that neuroscience faces a "binding problem" of behavior which likely requires such "invisible mental causation" for its solution. Finally, he thinks that invisible mental causation is immune to empirical refutation. As a response I argue first that Lowe's motivations for invisible mental causation is insufficiently developed. Second, I show that Lowe's argument for invisible mental causation does in fact depend on empirical assumptions about the nature of neural causation, which Lowe does not justify. So Lowe's argument is inconclusive, and there are no principled obstacles to using empirical evidence to dismiss invisible mental causation. Third, I argue on empirical grounds that it is far from clear that there is any "binding problem of behavior" that cannot be solved using standard neuroscientific methodology. In the absence of a problem that cannot be solved by neuroscience, the case for invisible mental causation would collapse.

Now I a fourfold vision see,
And a fourfold vision is given to me;
'Tis fourfold in my supreme delight
And threefold in soft Beulah's night
And twofold Always. May God us keep
From Single vision & Newton's sleep!
– William Blake[52]

[52] This quote, in which Blake complains about what he took to be the restricted, "single vision" of Newtonian physics is from a letter to Thomas Butts, November 22, 1802. (Blake, 1966, p. 818)

5.1. Introduction

The Causal Closure of Physics is a central tenet of physicalism. To a first approximation it states that:

> (CCP) Any physical event that has a sufficient cause has a sufficient physical cause.[53]

This is a strong and important metaphysical claim in its own right, for it is intended to ascribe a radical causal self-sufficiency to the physical domain. It also plays an important role in the much discussed Exclusion Argument. This argument concludes that unless mental events reduce to physical events, they must be epiphenomenal or causally inert *vis-à-vis* the physical. If this threat of epiphenomenalism is real, philosophers should be strongly motivated towards reducing the mental. It is part of the commonsensical and scientific understanding of mentality that mental events, like the occurrence of beliefs and desires, cause bodily movements. But being physical, bodily movements must have physical sufficient causes, by CCP. Add to this some Principle of No Overdetermination, claiming roughly that physical effects like bodily movements are not generally overdetermined by more than one sufficient cause. Then, given their physical causes, there seems to be no room for additional and non-reducibly mental causes of bodily movements. (See, e.g., Kim, 1998, ch. 2; 2005, ch. 2)

Thus stated the use of CCP in the Exclusion Argument does not beg the question against antireductionists by ruling out the possibility of bodily movements being causally overdetermined by two sufficient causes, one mental and one physical. That job is left for the No Overdetermination Principle, with which some antireductionists, like Barry Loewer (2002), take issue. Perhaps for this reason, CCP is widely accepted among reductive and antireductive physicalists alike. However, acceptance of the principle is not

---

[53] As we shall see, closure principles can be formulated in many ways. To allow for indeterminism it is sometimes said that all physical events have their *chances* determined by prior physical events. (Papineau, 2001, p. 8n2) This complication is not relevant to Lowe's position, however, and Lowe formulates closure principles independently of it.

universal among philosophers at large. In particular E.J. Lowe, whose position is dualist and antiphysicalist,[54] thinks that CCP is *false*. The reason is that he takes CCP to be incompatible with the freedom of the will, which he thinks cannot be rationally doubted. (Lowe, 2003, p. 145) More interestingly for present purposes, he argues that there is no empirically plausible version of CCP that will make the Exclusion Argument valid without begging the question against dualists. (Lowe, 2000; 2003) To demonstrate this he describes two dualistic scenarios – where physical causes really need "help" from mental causes to bring about bodily movements – that are nevertheless compatible with the principle. This argument, then, is not against CCP as such, but rather against its use in the Exclusion Argument. Since Lowe claims that the contribution of these mental causes would be invisible from a physical point of view (2000, p. 580), I call this doctrine "invisible mental causation." While the doctrine appears to be intended primarily as a response to the Exclusion Argument, and hence to reductive physicalism, I doubt that many antireductionists would subscribe to it. The dispute, then, seems to be between dualism and physicalism in general. (Though I shall argue that it is also a dispute between dualism and *science*.)

CCP can at any rate be strengthened so as to rule out invisible mental causation, but Lowe does not think that would be empirically or philosophically warranted. In fact, he suggests that we should favor invisible mental causation over physicalist alternatives. For he takes it that neuroscience faces a *binding problem of behavior* – allegedly similar to the binding problem in perception – which requires something like invisible mental causation for its solution.[55] While I think Lowe is right in demanding empirical evidence for physicalism, I shall argue that his argument for invisible mental causation is inconclusive at best. I show that this argument relies on an empirical assumption which

---

[54] For instance, he (1999) doubts mind-body supervenience, which is commonly taken to define the minimal commitments of physicalism. (Lewis, 1983; Jackson, 1998, ch. 1) It will also become clear from Lowe's suggestions that he favors an outright dualistic conception of the mental causation that few physicalists, reductive or antireductive, would accept.

[55] I emphasize that Lowe does not explicitly commit himself to invisible mental causation rather than other dualist views of mental causation. He does, however, offer reasons to prefer it over physicalist alternatives.

he does not justify. Furthermore, I argue that it is far from clear that it *can* be justified. It is not clear that there is any binding problem of behavior, or at any rate, none that is likely to require anything but standard neuroscientific methods for its solution. Consequently, I do not think Lowe has provided us with sufficient reasons for favoring invisible mental causation over physicalist alternatives. Whatever problems there are with CCP and the Exclusion Argument, invisible mental causation does not appear to be one of them.

5.2. The Case for Causal Closure

Before considering Lowe's argument I will briefly summarize the case that has been made for CCP. I do so because this is presumably the kind of arguments that Lowe thinks are insufficient to rule out invisible mental causation, but also because I am in a certain sense in agreement with Lowe. CCP is not an obvious or self-evident truth. Too often has it been accepted on the basis of very brief arguments that involve little or no explicit attention to empirical evidence. In an interview Jaegwon Kim (2000) actually claims that CCP "isn't an empirical issue for the physicalists", and suggests that CCP may be built into the nature of the *concept* of the physical. Donald Davidson (1980) makes some similar remarks. But as we shall see, the question of CCP is arguably a deeply empirical one.

It might, of course, appear that only few and relatively uncontroversial assumptions are needed to arrive at CCP. For instance, Kim (1992) contends that the failure of CCP would entail the violation of the laws of physics, so CCP must hold. This line of reasoning arguably fails. On closer inspection, and somewhat surprisingly, the truth of the laws of physics appears consistent with the failure of CCP. Briefly, the reason is that laws like Newton's second – $\mathbf{F} = m\mathbf{a}$ – are causally neutral, insofar as they do not care what causes there are.[56] $\mathbf{F} = m\mathbf{a}$ could remain true even if some accelerations were caused by (say) irreducibly vital and non-physical forces. (See McLaughlin, 1992 and Papineau, 2001 for details.)

---

[56] Cartwright (1979) uses a slightly different sense of "causal neutrality" to characterize laws like $\mathbf{F} = m\mathbf{a}$.

Similarly, Kim (1998, p. 40) argues that only CCP can guarantee that it is in principle possible to arrive at a complete physical theory, so again CCP must hold. But as David Papineau makes clear (2001), the scientific community has changed its mind about the completability of physics several times throughout history. So, even though physics likely aims at being as complete as possible, the possibility of an "absolutely" complete physics is not happily viewed as built into the nature of the enterprise of physics. Completability, then, cannot be assumed except by reference to significant empirical evidence. In fact, authors like Brian McLaughlin (1992) and Papineau (2001) suggest that what they take to be convincing evidence for CCP emerged only quite recently. (See below.)

CCP, then, is a deeply empirical claim, and one that likely requires some sort of inductive argument for its support. Lowe is therefore perfectly right to demand that our closure principle "must be one for which some measure of empirical support can plausibly be mustered". (Lowe, 2000, p. 572) This empirical constraint is important, because, as we shall see, he thinks it prohibits the use of such principles in Exclusion Arguments.

What evidence, then, is available to the physicalist? McLaughlin and Papineau argue for principles like CCP on explicitly empirical and inductive grounds. They contend that while the epistemic situation at earlier points in history made it reasonable to postulate irreducibly vital, chemical and mental forces, it no longer does. Briefly, the quantum mechanical explanation of chemical bonding undermined a prime example of emergent forces, and contributed to the fall of classical emergentism. (McLaughlin, 1992) Furthermore, special forces like friction all appear to reduce to a few fundamental physical forces. This provides inductive evidence against mental or vital forces.[57] (Papineau, 2001) Finally, Papineau offers a second argument for CCP, which he dubs

[57] I have sometimes encountered the following objection to this argument. Why should an argument against mental *forces* be an argument against mental *causes*? Lowe, for instance, makes no assumption that invisible mental causation involves mental forces. But if Papineau (2001) is right, effects like bodily movements can be fully accounted for by physical forces, or the physical circumstances that give rise to these forces. This is all that CCP requires.

"the argument from physiology," In outline, this argument contends that the molecular revolution in biology has failed to reveal any effects within the nervous system that appear not to be products of physical causes. This would seem to add to the evidence for CCP. (Papineau, 2001)

Such arguments strike me as strong, but the empirical evidence for CCP has been questioned. (Cp., e.g., Hendry, 2005; Sturgeon, 1999) It will emerge, however, that, we can address Lowe's specific problems by appealing to neuroscience, rather than physics, somewhat in the spirit of Papineau's argument from physiology.

Lowe's intriguing, but in my view ultimately ineffective, arguments are worthy of attention in their own right. But considering them is also beneficial for a second reason. It serves to underscore the thoroughly empirical nature of physicalism and physicalist claims like CCP.

5.3. The Possibility of Invisible Mental Causation

I turn now to Lowe's arguments. His first invisible mental causation scenario explicitly takes into account the times at which events occur.[58] (Lowe, 2000, pp. 575-576)

> (S1) It is true that any physical event E that has a sufficient cause must have a sufficient physical cause P at *some* earlier time. But this P causes E, at least in part, by causing an irreducibly mental event M occurring at some intermediate time. M is an ineliminable element in P's causing E.

The idea, then, is that there can be *two* causal processes by which P causes E. One is a purely physical process, which we can represent as follows: (i) P → E. The second, however, involves the causing of an intermediate and irreducibly mental event, which in turn helps bring about E: (ii) P → M → E. (S1) is a dualist scenario, but is nevertheless consistent with CCP and the absence of overdetermination. First, P *is* sufficient for E, but only because it also causes M. Second, Lowe is quite explicit that the mental cause M is

[58] Here I can only offer a brief and simplified sketch of Lowe's scenarios; see Lowe (2000; 2003) for more discussion. My objections do not depend on my simplifications, however.

not to be construed as an overdetermining cause. (Lowe, 2000, p. 576 & 577) That is, the causal contribution of M in process (ii) is not redundant given the purely physical process (i).

This, I take it, is not what most physicalists intend by CCP. Importantly, as Lowe describes the scenario, the essential contribution of the mental event would be *invisible* from a physical point of view. It would seem to a physical scientist looking only at physical processes as if he had a complete physical explanation. (Lowe, 2000, pp. 580-581) Anyway, Lowe thinks the possibility of (S1) requires exclusionists like Kim to strengthen CCP by introducing a *temporal* constraint prohibiting such intermediary mental causes. One way of doing this is as follows. (Lowe, 2000, p. 576)

> (CCP*) At every time t at which any physical event has a cause, it has a sufficient physical cause.

But Lowe's second scenario is intended to show that even this will not do. (Lowe 2000, pp. 576-577)

> (S2) It is true that every physical event E that has a cause occurring at time t must have a sufficient physical cause P occurring at t, but that physical event causes the effect E, at least in part, by *instantaneously* causing an irreducibly mental event M occurring at the *same* time t. M is an ineliminable element in the P's causing E.

Apart from the instantaneous causation, the scenario is the same as (S1). Here, then, we have a dualist scenario that is compatible even with the temporally constrained CCP*. If this possibility is to be ruled out further changes to CCP are necessary. Lowe suggests the following. (Lowe, 2000, p. 581)

> (CCP**) Every physical event contains only other physical events in its transitive causal closure.

An event's transitive causal closure includes only its immediate causes, their immediate causes, and so on. (Lowe, 2000, pp. 581-582) This principle rules out both (S1) and (S2), which Lowe, as we shall see, takes to be real and interesting possibilities.

To sum up; Lowe's challenge is this. (S1) and (S2) are dualist scenarios that are compatible with CCP and CCP* respectively. They cannot be dismissed as cases of overdetermination. To make the Exclusion Argument valid the exclusionist must strengthen the closure principle from CCP via CCP* to CCP**. The latter principle rules out (S1) and (S2). But the resulting Exclusion Argument begs the question against friends of invisible causation.[59] Indeed, it would seem to beg the question against antireductive *physicalists* as well.[60]

5.4. Lowe's Argument for the Plausibility of Invisible Mental Causation

These scenarios raise important methodological and metaphilosophical questions about physicalism and dualism. How is the burden of proof to be distributed among these two positions? What is the role of empirical evidence? Physicalists should at least grant that invisible causation (and hence; dualism) seems to be a logical possibility given the empirical evidence offered in favor of CCP. Invisible mental causation, then, can be viewed as an interesting alternative interpretation of the scientific *status quo*, and it can explain why the physical domain *seems* to be closed. So invisible mental causation is consistent with scientific knowledge. Now Lowe sometimes argues as if this consistency were sufficient to undermine the Exclusion Argument. (Lowe, 2000, p. 572) I find this

---

[59] Lowe actually goes on to present a third and different scenario, according to which irreducibly mental events do not cause physical events, but rather cause the *fact* that a specific physical event causes another physical event. (Lowe, 2000, p. 582) This antireductionist scenario is supposed to be compatible even with CCP**. An even stronger version of CCP would, then, be required to rule out this sort of invisible mental causation of physical facts. I will only discuss the plausibility of the first two scenarios here, though I think my objections apply equally to the third.

[60] Note, though, that a proponent of the Exclusion Argument might rule out invisible mental causation in a different way. She could for example stipulate that sufficient physical cause means "physical through and through" as suggested by Montero (2003, p. 174n1). The resulting argument would still beg the question against proponents of invisible mental causation, but not against most antireductive physicalists.

surprising, and think that it clearly cannot be the case. In my view, arguments for the *possibility* of dualist scenarios do not cut any philosophical ice unless the dualist scenarios have independent motivation. In the absence of additional empirical evidence or over-arching theoretical constraints favoring invisible mental causation, the natural thing to do would be to dismiss Lowe's suggestion as *ad hoc*. My view is therefore that the burden of proof is on Lowe. If he cannot offer empirical or theoretical evidence for his suggestion, I think it will be legitimate for physicalists to beg the question and rule out invisible mental causation by stipulation in their formulation of the closure principle. Indeed, I think it will be all right for them even to render the closure principle as CCP or CCP*, and grant that while the resulting Exclusion Argument is not deductively valid – because it leaves the logical possibility of invisible mental causation open – it is nevertheless a plausibility argument.

What makes Lowe's challenge so interesting, however, is that he also offers an intriguing sketch of why theoretical or philosophical constraints might actually favor invisible mental causation. Here is a lengthy, but revealing, quote:

> [If invisible mental causation is real] physical science can present us with the semblance of a complete explanation of our bodily movements, and yet it will be an explanation which leaves something out, giving our bodily movements the appearance of being coincidental events arising from independent causal chains of events in our brains and nervous systems. But isn't that precisely what current physical science does appear to do? As it traces back the physical causes of our bodily movements into the maze of antecedent neural events, it seems to lose sight of any unifying factor explaining why those apparently independent causal chains of neural events should have converged upon the bodily movements in question. In short, it leaves us with a kind of 'binding' problem, not unlike the 'binding' problem associated with conscious perceptual experience […]. (Lowe, 2000, p. 581)

Lowe, then, suggests that there may be a *binding problem of voluntary behavior*, allegedly analogous to the binding problem(s)[61] of perception. We know for example that when we perceive, say, a moving car, different perceptual features – the car's movement, shape, color etc. – are processed by different cerebral regions. How, then, do these processes combine to yield a unified percept of a moving, car-shaped and colored object? The quote attempts to raise a similar problem for behavioral neuroscience. Lowe seems to think there are *too* many, *too* messy and *too* independent processes leading up to a bodily movement, and that these processes must be bound somehow.

A hasty response might be: "So what? Messy and independent physical causes would still be physical *causes*!" The problem with this response is that it uncharitably fails to appreciate what is at stake for Lowe. And thereby it also misses out on what I presume to be a crucial background assumption he makes about mental causation. Lowe worries about binding because he worries that the independence of neurophysiological causal processes might make the bodily movement *coincidental*. (Lowe, 2000, p. 579) An event is coincidental "if its immediate causes are the ultimate effects of independent causal chains." (Lowe, 2000, p. 579) As an example he offers a version of Aristotle's story about the man walking to the well. As the man passes under a roof, the wind dislodges a slate from the roof; the slate hits the man and kills him. His death occurs by coincidence; because the causal chain leading up to the man's being at the wrong place at the wrong time is causally independent of the chain leading up to the fatal slate's ending up at that place. (Lowe, 2000, p. 579) Note that coincidental effects in this sense are not *uncaused*, neither need they be the outcome of *probabilistic* causal processes. The causal chains leading up to a coincidental effect can be as deterministic as you please.

Lowe later suggests that there is often "a strong intuition that the bodily movement in question was *not* an event which occurred 'by coincidence'." (Lowe, 2000, p. 584) Lowe's binding problem of behavior then, is that of rendering bodily movements non-coincidental. Finally, he pictures *a role for invisible mental causation in solving this*

---

[61] As Anne Treisman (1996) makes clear there are *several* binding problems. Examples are the binding of the appropriate properties to the appropriate objects, the appropriate objects to the appropriate spatial regions and so on.

binding problem, and offers the following abstract example. Suppose we have two physical causal chains leading up to the same physical event E. (i) P1 → P2 → P3 → E and (ii) P1* → P2* → P3* → E. From a physical point of view E is coincidental, because events in the two chains are not linked. But suppose invisible mental causation involving an intermediate mental event M link the two chains, thus. (iii) P1 → M → P3*. That would make E non-coincidental after all. (Lowe, 2000, p. 580) In this connection it should be noted that Lowe does not explicitly demand that all bodily movements be non-coincidental. He may think that only *voluntary* movements require binding by invisible mental causes.

Judging from Lowe's technical definition of coincidental effects, and the abstract example of binding, it is, however, is a little hard to appreciate the nature of his binding problem, and what role precisely invisible mental causes are supposed to play. The binding of neural causal chains is supposed to explain "why those apparently independent causal chains of neural events should have converged upon the bodily movements in question." (Lowe, 2000, p. 581) The problem is that it is not entirely clear what the "why"-question concerns here. But apparently a cause that links the chains, as M does in the above example, would be sufficient to answer the question.

Due to the relative unclarity of the binding problem and of its relation to the notion of non-coincidental effects I think Lowe is moving too quickly here. I am not convinced that there is such a strong intuition about bodily movements being non-coincidental in Lowe's technical sense. Neither am I sure why precisely our theory of mental causation would have to conserve this intuition if such there be. That being said, the idea that mental causes might be needed to bind neural or physical processes is interesting and worthy of further exploration. It suggests that mental causes are somehow needed to "integrate" or "organize" physical processes. There might therefore be some kind of necessary causal work to be done that physical causes do not do. So while I think Lowe's binding problem needs to be made much more concrete before it is truly

challenging, I am willing to accept the requirement that bodily movements be non-coincidental for the sake of argument.[62]

Summing up; given Lowe's view of what neuroscience tells us (or rather; does *not* tell us) and his background assumption that bodily movements must typically or often be non-coincidental, I think a charitable interpretation of Lowe's argument for invisible mental causation would be as an inference to the best explanation:

(1) From a neurophysiological point of view, bodily movements appear coincidental
(2) Bodily movements are typically non-coincidental
(3) Invisible mental causation would make bodily movements non-coincidental and explain the appearance of coincidence
---
(4) Invisible mental causation occurs

5.5. Against Invisible Mental Causation

Suppose then, that Lowe is right about premise (2) in the above argument. There really is this binding problem of behavior, insofar as bodily movements must be rendered non-coincidental. I shall argue first that the argument is inconclusive (5.5.1), because it rests on an empirical assumption he does not justify. Second, I shall offer some empirical reasons against invoking invisible mental causation. (5.5.2-5.5.3)

---

[62] Lowe (1999) offers some further considerations to motivate his binding problem, and indicates that binding is somehow necessary to make bodily movements intentional or perhaps voluntary. If the question "why" neural chains should converge on a bodily movement is intended as a call for some sort of intentional or cognitive answer it is not hard to sympathize with it. We *are* interested in explaining bodily movements as intentional behaviors. What is less clear is how the "why"-question relates to the notion of non-coincidentiality, and whether its solution requires mental causes to do extra causal work over and above that done by physical causes. But the important thing to notice is again that whatever the nature of the binding problem, Lowe takes its solution to hinge on bodily movements being non-coincidental, and his definition of such effects does not refer to intentionality.

5.5.1. Lowe's Argument is Inconclusive

It behooves me to first get clearer about the role of empirical evidence in Lowe's reasoning and in my criticism. This will also reveal that the argument is inconclusive as it stands. As we saw, Lowe portrays dualism as an empirically respectable doctrine, because it is consistent with empirical data. But as I urged above, consistency with data is not sufficient to make a doctrine empirically respectable, for it could still be dismissed as *ad hoc* on empirical grounds. About this, we simply disagree. (Lowe, 2003, p. 152n10) But I will offer an argument against invisible mental causation which does not invoke complaints about its being *ad hoc*. (5.5.2) Suppose invisible mental causes are necessary to get complete explanations of the causal processes leading up to bodily movements. That would represent a major deviation from neuroscience's explanatory *aims* and *methods*. And the success story of neuroscience has given us reasons to doubt that that kind of deviation will be necessary. I do not expect that this argument would impress Lowe directly, but, if successful, it will show that his view does after all conflict with certain scientific views. It is not easy, I maintain, to square dualism with central theoretical assumptions of neuroscience. This point is important for present purposes, as Lowe appears eager to have no quarrel with science, for instance he characterizes his dualism as "naturalistic." (Lowe, 2000, p. 572)

Anyway, my second argument against invisible mental causation is one that he should be more willing to accept. (5.5.3) I argue that it is far from clear that neuroscience must render bodily movements coincidental. So in fact, I think premise (1) above is likely to be false. This is a possibility that Lowe fails to consider, and it is instructive to see how it will, if sound, undermine his project of defending dualism on more or less *a priori* grounds, while simultaneously claiming that dualism is consistent with empirical data. Lowe makes it clear that he takes the positive reasons for believing in invisible mental causation to be theoretical or philosophical assumptions – like (2), presumably – rather than empirical evidence:

> [The reasons for believing in invisible mental causation] could not be broadly *empirical* ones. The lesson of this is that we should be prepared to acknowledge that *a priori* metaphysical

> argumentation may in the end provide the best, or indeed the only, hope for a resolution of the mind-body problem. (Lowe, 2003, pp. 153-154)

I grant, of course, that known empirical data do not count as *positive* evidence for invisible mental causation. And though I am personally skeptical about the prospects for *a priori* methods in resolving the mind-body problem, I also grant that theoretical constrains like (2) *might* make it rational to invoke invisible mental causation despite its being in my view *ad hoc*. It is a familiar point that *ad hoc* strategies can be rational provided they are needed to defend theoretical assumptions that have previously proven successful for example.

But when Lowe claims that (2003, p. 153) neuroscience cannot provide evidence *against* invisible mental causation, he arguably misses an important point. For even assuming that (2) has *a priori* or theoretical motivation, his argument does depend on premise (1). This premise contends that bodily movements must appear coincidental from the point of view of neuroscience. Importantly, it is an *empirical* premise about the nature of neural causation. Now, notice that there is nothing in Lowe's characterization of coincidental effects – as the outcome of independent causal chains – nor in his abstract example of binding by invisible mental causes that prohibits physical causes from rendering their effects *non*-coincidental. In principle, neural causal chains could be bound by neural causes. So, *pace* Lowe, neuroscience can in principle falsify a crucial premise in the argument motivating invisible mental causation. Unfortunately, Lowe does not appear to really argue for his empirical assumption about neural causation, rather he asserts that there do not appear to be any "unifying factors" in neural causal chains. (Lowe, 2000, p. 581) This is arguably not the kind of claim that should be made without explicit reference to empirical evidence. Lowe, then, needs to offer reasons – empirical reasons – for his crucial empirical premise.[63]

---

[63] Alternatively, he could revise his characterization of coincidental effects and argue that physical causes are intrinsically unsuited for the job of rendering effects non-coincidental. But such a claim about the nature of physical causes would, presumably, again be in need of empirical justification.

I am, however, skeptical that that can be done. And if (1) is – or is likely to be – false, (2) in no way counts in favor of invisible mental causation. So, unless Lowe wants to rely solely on his weaker claim that invisible mental causation is *possible* given empirical knowledge, his case for invisible mental causation would collapse. I turn now to positive reasons for skepticism about invisible mental causation.

5.5.2. Invisible Mental Causation Fits Ill with Neuroscientific Practice

In this section I make to claims about Lowe's binding problem. (i) Invoking invisible mental causation to solve the problem would involve a drastic departure from neuroscientific method as it is currently practiced. (ii) Neuroscience would also suffer a major defeat *vis-à-vis* its explanatory aims. For it is arguably part of those aims to understand *voluntary behavior*, that is, not mere bodily movements, without invoking dualistic notions like invisible causation. In both of these cases I contend that neuroscientists would be entitled to strong skepticism about the suggested deviation from their aims and methods. Their success so far entitles them to that.[64] Obviously, these claims cannot be fully justified here, but I contend that they are at least as plausible as Lowe's dualist alternative.

(i) *Deviation from Method*. It is revealing to see how neuroscience approaches the traditional binding problems of perception, which appear hard enough in their own right. I admit at the outset that my knowledge here is cursory and stems from review articles, rather than the primary experimental literature. But I think even a brief glance at the standard proposals for solutions; will confirm what I want to claim. Consider for example

---

[64] This argument sounds disturbingly similar to Kim's argument that CCP must hold (because unless it does, a complete physical theory will be in principle impossible), an argument which I questioned above. However, I only criticized Kim for assuming the possibility of a complete physical theory on more or less *a priori* grounds. Below I will briefly sketch why I think neuroscientists enjoy an *a posteriori* entitlement to their explanatory confidence. If Lowe questions this entitlement he should accuse neuroscientists, and not just physicalist philosophers, of explanatory *hubris*. Perhaps he is willing to do that, but as I have said does not appear to want his quarrel to be one with science.

the solutions discussed in an issue of the prestigious[65] journal *Neuron* that was dedicated largely to reviewing binding problems. Binding might be produced by a traditional neurocomputational processing hierarchy where neurons with less specific receptive fields feed information to neurons whose receptive fields are increasingly specific conjunctions of features. And/or it may arise from the synchronous firing of potentially distal neurons that code for the features to be bound. And/or it may depend on neural mechanisms of attention.[66] (Where "attention" is something that, e.g., parietal cortex, rather than invisible mental causes, commands.) This illustrates to me that even in the face of an extremely recalcitrant set of problems; mainstream neuroscientists do not turn to dualistic methods.

Why this confidence in neural methods? It was not always so, indeed the history of neuroscience involves the postulation of non-physical causes to explain phenomena like neural signaling. (See, for instance, Finger, 1994) Though I cannot argue it fully here I think most neuroscientists treat their domain as closed *vis-à-vis* the mental domain. This may be viewed as an expansion of Papineau's above-mentioned argument from physiology to the closure of the physical domain in *general*. (Papineau, 2001) But considerations of the more limited neurobiological domain are, as we shall see, sufficient to warrant skepticism about invisible mental causes.

My claim is not the obviously false one that neurophysiology suffers no causal input from the outside. But fixing these causal inputs, letting temperature and pressure be normal, letting there be no strong electromagnetic fields nearby and so on, *neural events*

---

[65] Here and elsewhere in this paper I appeal to sociological factors like "prestige" and "influence." Insofar as my claims are descriptive and concern the nature of neuroscientific practice this should be non-tendentious. Insofar as the claims are normative and evaluative it will perhaps be found more problematic. It is, however, my hope and belief that the sociological mechanisms that determine what research counts as prestigious also reliably – though obviously not infallibly – select for *good* research.

[66] See, e.g., Roskies (1999) and Treisman (1999), which are general reviews. Reynolds & Desimone (1999) argue for a role of attention in binding. Singer (1999) focuses on neural synchrony as a binding mechanism. von der Malsburg (1999) discusses binding from a general computational perspective. Treisman (1996), an earlier general review, also discusses binding from the point of view of non-dualistic methods. Robertson (2003) is a recent discussion of the role of attention and parietal regions.

have sufficient neural causes.[67] This claim may sound trivial, but it is not. Its plausibility derives from the detailed knowledge neuroscientists have about neural causation. Furthermore, we have little reason to believe that invisible mental causes are lurking in the *ceteris paribus* conditions under which neural causation take place, which is a non-trivial additional assumption. While much remains unknown about the nervous system, neuroscientists have a good grasp of what *kind of events* are important for motor output, for cognition and for sensation. Here, events like neurons firing with certain frequencies (or perhaps in certain more complex patterns) figure prominently. And they know a lot about *how* – that is, through which mechanisms –causes must work in order to bring about this privileged kind of effects. Certainly, electro-magnetic, chemical and other causes that affect how ions flow across neural membranes are neural causes *par excellence*. This point is forcefully made by John Bickle.

> If action potential rate is the currency of neural causation and information exchange, then the only way an event can elicit neural change is by affecting the processes that underlie action potential generation in individual neurons. That is where the rubber meets the road. (Bickle, 2003, p. 59)

This causal picture of the kinds of causes that are relevant to neural interactions and the mechanisms, through which they must work, is painted by influential text books like Kandel et al.'s *Principles of Neural Science* (2000). Such text books arguably play a sociological role in defining scientific fields. (Schaffner, 2006) As such the causal picture probably also shapes the kind of hypotheses working neuroscientists form and are willing to take seriously. It would of course be naïve to believe that the picture cannot, or will not, change or be revolutionized in certain ways. It may turn out, say, that glial cells are more than a "support team" for neurons, and perform hitherto unheard of functions. (Bullock et al., 2005) But revolutions of this kind are one thing, invisible mental causation is quite another.

Just to appreciate the kind of, I think legitimate, explanatory confidence neuroscientists appear to have, consider a recent argument for "ectopic

---

[67] For a more details about the claims I make in this section, see paper (#4).

neurotransmission." It is a well-established fact that neurons communicate via specialized synapses, where neurotransmitter is released from presynaptic "active zones" onto "postsynaptic densities." But it has also been suggested that in some cases "ectopic" release of neurotransmitter at sites distinct from the above-mentioned specialized regions might play a role in neurotransmission. To investigate this Coggan et al. (2005) developed a biologically realistic computational model of a specific type of excitatory synapse in chicks. They found that the simulated postsynaptic effect did not conform to the actually measured effect, *unless* ectopic transmission was included in the model. The interest of this for present purposes lies not in whether Coggan et al. are ultimately right, nor in their scientific details, but rather in the reasoning behind their argument. Unless we assume that *two* kinds of neurophysiological causes are at play, we cannot explain the effect we measure. Reasoning in this way requires a significant amount of confidence in one's knowledge of the causal factors that are at play in neurotransmission, and in how these causal factors must work in order to bring about the measured postsynaptic effect. It is not the kind of research that would reliably pass as good science unless it reflected a real grasp of the kinds of causes that matter in neuroscience.

(ii) *Deviation from Explanatory Aims*. I do not think it is an accidental feature of neuroscience that its explanatory toolbox includes the broad kind of causes and mechanisms I have just sketched. I think it is, or has become, an explanatory aim of neuroscience as it is conducted at the cellular and molecular levels to explain by using *that* kind of explanatory tools. So when these explanatory tools are taught to neuroscientists in text books, the aims of neuroscience are simultaneously being laid out. If I am right, having to invoke mental causation to fully explain bodily movements would represent a major failure of neuroscientists in reaching their explanatory aims. But the reader need not take my word for this, as I shall offer two brief appeals to authority to support my claim.

First, explanations by mechanisms have received much attention in philosophy of neuroscience recently. In a much-discussed paper, Peter Machamer et al. (2000), argue that such explanations are more or less the heart and soul of neuroscience. They view mechanisms as organized complexes of "entities" like ion-channels, neurons and neural circuits, and "activities" of these entities. Activities are things that entities *do*. Ion-

channels, for instance, can *open*, and neurons can *fire*. Machamer et al. illustrate how such complexes give rise to causal processes, e.g., in the mechanisms underlying neural signaling. Most importantly for present purposes, they invoke an interesting notion of "bottoming out:"

> Nested hierarchical descriptions of mechanisms typically *bottom out* in lowest level mechanisms. These are the components that are accepted as relatively fundamental or taken to be unproblematic for the purposes of a given scientist, research group, or field. (Machamer et al., 2000, p. 13)

Molecular neuroscientists, for instance, currently want their mechanisms to bottom out in the activities of molecules and ions. (Machamer et al., 2000, p. 14)[68] Whatever the level of bottoming out, I do think this has implications for invisible mental causation. If *ascending* to the level of acts of invisible mental causes is necessary to fully explain the causal chains leading up to bodily movements, that would seem to conflict with the explanatory aim of *descending* to the level of parts and explaining neural causation in terms of complex neural mechanisms.

The reader may have been worrying for a while that this is all irrelevant to my case against Lowe. Would not the contribution of invisible mental causes be a matter of cognition and mentality, and thus fall outside the explanatory interests of neuroscience? Could not neuroscientists complete the neuroscientific explanatory task of discovering mechanisms and leave mentalistic explanations for psychologists and philosophers? I

---

[68] Notably, Machamer et al. (2000) claim that the level of bottoming out varies across different branches of neuroscience, and may change over time. For the explanatory purposes of studying how neural systems interact for instance, descending to molecular mechanisms may not always be taken to be necessary. However, I take it to be a background assumption of neuroscience as a whole that there is some story to be told at this level. (Cp. the discussion of the neural causal picture above.) In fact, Machamer et al. do say that the activity of potassium channels for instance is ultimately a component of "most higher-level mechanisms in the nervous system." (2000, p. 13) Anyway, I do not think much hinges on which level is taken to be the bottom level, since invisible mental causation is not likely to appeal to neuroscientists working at higher levels either. (It should also be noted that Machamer et al. are not claiming that complex mechanisms *reduce* to what they call "lowest level mechanisms.")

think this objection rests on a misunderstanding of the neuroscientific enterprise. Many neuroscientists are interested in understanding mentality and cognition and furthermore in doing so *in terms of* mechanisms. Neuroscience takes an interest in behavior, not just in mere bodily movements. I shall return to this below. For now consider my second appeal to authority. Here is the late neuroscientist Patricia Goldman-Rakic, famous for her work on the role of prefrontal cortex in voluntary behavior, explaining her aims to a popular audience.

> Until recently, the fundamental processes involved in […] higher mental functions defied description in the mechanistic terms of science. Indeed, for the greater part of this century, neurobiologists often denied that such functions were accessible to scientific analysis or declared that they belonged strictly to the domain of psychology and philosophy. Within the past two decades, however, neuroscientists have made great advances in understanding the relation between cognitive processes and the anatomic organization of the brain. As a consequence even global mental attributes such as thought and intentionality can now be meaningfully studied in the laboratory.
>
> The ultimate goal of that research is extraordinarily ambitious. Eventually researchers such as myself hope to be able to analyze higher mental functions in terms of the coordinated activation of neurons in various structures in the brain. (Goldman-Rakic, 1992, p. 73)

I do not think my appeal to explanatory successes and aims of neuroscience would impress Lowe. What we know about the nervous system does, after all, appear consistent with invisible mental causation, and Lowe thinks there are independent grounds for believing in such causation. I will question those grounds in the next section, but even if my arguments there fail, I think I have arrived at something interesting. Lowe's quarrel appears to be as much a quarrel with neuroscience as it is a quarrel with philosophical physicalism. Science and physicalism are perhaps not easily divorced.

5.5.3. Bodily Movements Do Not Appear to Be Rendered Coincidental.

It is important to note that Lowe might be perfectly right to quarrel with neuroscience. Indeed, he would be, if there is a real binding problem of behavior that neuroscience cannot solve nor legitimately set aside. But as we saw, whether neuroscience can explain the binding of neural processes leading up to bodily movements is an empirical question.

We can look at neuroscientific models and see whether they appear to render movements coincidental. In this connection I would like to note that there might be a disanalogy between Lowe's problem of coincidentiality and the traditional binding problems. For in the case of behavior, there is a kind of causal convergence of causal processes on structures like motor cortex that cause muscle contractions. In contrast, it is part of the binding problem in perception that there apparently are no "Grandmother Neurons" on which different perceptual pathways converge.[69]

More importantly, Lowe thinks that physical science fails to find any "unifying factor" as it tracks causes of bodily movements into the "maze of antecedent neural events". (Lowe, 2000, p. 581) We have seen that he takes a unifying factor to be one which connects the causal chains leading up to bodily movements and thus explains "why those apparently independent causal chains of neural events should have converged upon the bodily movements in question." (Lowe, 2000, p. 581) This linking or connecting is supposed to be along the line of Lowe's abstract example. (Lowe, 2000, p. 580) According to one estimate there are about 100 billion neurons in the brain, so it is easy to see how one might get lost if one were to trace the causes of individual neurons' firing backwards from, say, primary motor cortex. This is, presumably, how Lowe views neural causation. From this point of view it would perhaps be hard to tell whether there any "linking" causes that ensure proper binding.

But tracing causes in this way, is arguably but one part of the neuroscientific enterprise. Neuroscience is an interdisciplinary practice which studies the nervous system using a wide variety of techniques like: imaging studies, lesion studies, behavioral and psychological tests, anatomical studies, computer simulations, pharmacological studies, single-cell recordings, microstimulation and so on and so forth. This is philosophically significant. In the neuroscience of voluntary behavior these kinds of techniques have allowed scientists to impose considerable causal structure on Lowe's neural maze. In fact, the operative understanding of how the central nervous system controls movement is quite fine-grained and includes: *strategy* (the goal of movement and choice of movement

[69] Though interestingly a proposed "Jennifer Aniston Neuron" in one human subject has recently received a lot of attention. See, e.g., Quiroga et al. (2005)

strategy in order to reach that goal), *tactics* (implementing movement strategies by way of spatial and temporal coordination of muscle contractions and relaxations) and *execution* (direct causing of contractions/relaxations). (Rains, 2002, p. 228) See, e.g., Kandel et al. (2000, ch. 33) for a similar description of how the nervous system organizes muscle-contractions to give rise to *goal*-related behavior.

Given this kind of characterization, I take it that the following are among the most central concerns of neuroscience. To *identify* structures or systems involved in the various levels of control, to investigate how – that is, by way of which *mechanisms* – the systems perform these functions, and to see how the systems *act together and communicate* to produce overall intelligent behavior. These are obviously not trifling tasks. The conclusions so far are partial and may be subject to revisions. But significant progress has been made, especially with respect to identifying structures which play different roles in the production of behavior, and with tracing out the anatomical connections through which these systems communicate. The figure G. Dennis Rains uses to introduce the neurophysiology of behavior is revealing with respect to what kind of knowledge neuroscientists are beginning to gather in this area:

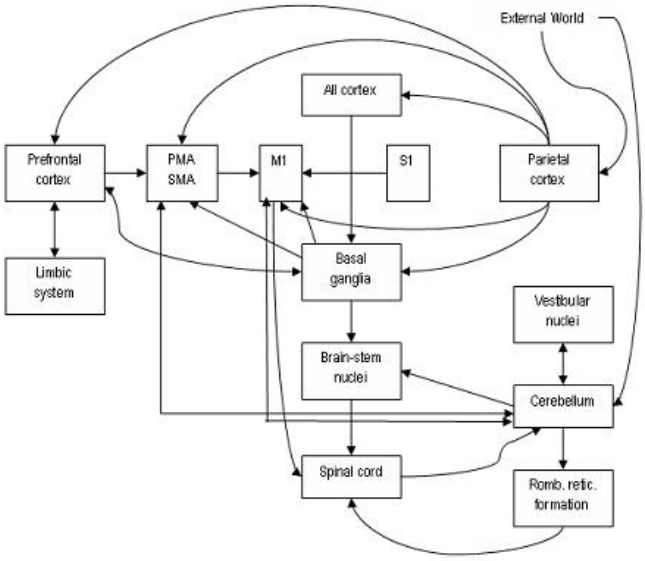

Fig. 1: Adapted from Rains (2000, p. 229, fig. 9.1)

This figure is accompanied by an overview of what the different structures are believed to do, and what they are thought to communicate to other structures. It is important to notice that these beliefs are not mere "boxological" speculations without empirical grounding. The neuroscientific techniques just mentioned have contributed strongly towards anchoring the model in biological knowledge. The interplay among structures depicted is complex, as one would expect, and the model is partial – much being known about some structures, less about others. We cannot be delayed by the interesting details here.[70] I hope the figure conveys its message anyway. The arrows, indicating simplified anatomical connections, are really what matters to us now. They illustrate how the interactions of different systems serving different functions, can implement the strategy, tactics and execution of bodily movements.

[70] I describe what I take to be the nature of neuroscientific explanations of behavior more fully in paper (#2).

So it is far from clear that neuroscience cannot offer an explanation of "why […] chains of neural events should have converged upon the bodily movements in question" (Lowe 2000, p. 581), even if the "why"-question is interpreted as a call for a cognitive or intentional answer. The perspective of the figure is, after all, cognitive. In fact, the figure would hardly have been available if research had not been guided by a cognitive-representational perspective. But this perspective is not one of invisible mental causation. The underlying theoretical assumption is that cognitive functions and control of voluntary behavior is implemented by underlying neural causal chains, without any invisible mental links. And these causal relations ultimately boil down to the principles of neurotransmission discussed above. This is very important. Physicalists and neuroscientists need not deny that cognitive perspectives and concepts are needed to see bodily movements as behaviors. What separates Lowe from most reductive and antireductive physicalists is that he thinks extra *causal* work done by mental causes is necessary to get sufficient causes of some effects like bodily movements, or at least to link the neural chains leading up to such effects.[71] Physicalism and neuroscience make no such assumption, and the knowledge embedded in the figure was arguably arrived at without assuming invisible mental causes.

But even setting aside the point that neuroscience offers putative explanations of *behavior*, and not just bodily movements without invoking mental causes that do extra *work*, there is a weaker point to be made against Lowe. His concept of non-coincidental effects is not inherently mental or cognitive. Accordingly, just the large number of interconnections between processes leading up to bodily movements should be sufficient to make plausible the claim that bodily movements are not rendered coincidental, even without invoking a cognitive perspective.

With this figure as a model for neural causation in place I can therefore conclude with what has perhaps been clear for a while. I do not think the neural system it is a

---

[71] Lowe (2000, p. 577) explicitly suggests that the causal chains running via invisible mental causes are needed to get the relevant effect. On the other hand, Lowe (1999) suggests that effect might occur even in the absence of invisible chains, but that occurrence of the effect would then be *coincidental*. But again, in both cases invisible mental causes do some kind of causal work that is *not* done by physical causes.

model *of* is like Lowe's neural maze. It is at best unclear whether we, by adopting a neuroscientific perspective on bodily movements, render bodily movements coincidental. The causal chains leading up to muscle contractions are interconnected, and non-independent. Furthermore, since each chain plays a role in a carefully orchestrated whole, it does not appear arbitrary that they should converge on the bodily movement in question. What we know about the neurophysiology of behavior, then, makes it far from clear that bodily movements are in need of invisible mental causes.

5.6. Conclusions

I have argued that Lowe fails to make a sufficient case for invisible mental causation. First, his description of the nature of the "binding problem of behavior" and its relation to the notion of non-coincidental effects is not sufficiently developed to motivate invisible mental causation. Second, and most importantly, while Lowe thinks invisible mental causation is immune to empirical refutation, his argument for it does depend on an *empirical* premise to the effect that neural causes *do* render bodily movements coincidental. Lowe does not justify this premise, so his argument is inconclusive. His characterization of the binding problem is compatible with its being solved by physical causes. Finally, I have offered reasons why neural causes do not appear to render their effects coincidental. Hence, it appears that Lowe's empirical premise might be false, and if it is the motivation for postulating invisible mental causes would not just be unclear, but also absent.

6. Paper (#4): What Is Closed in Causal Closure?

ABSTRACT: The Causal Closure of the Physical Domain (Closure) is a central tenet of physicalism, and plays an important role in what is known as the Exclusion Argument for mind-body reductionism. While (Closure) is widely accepted, there *are* potential problems with it. First, judging from the mental causation literature it is not clear what the physical domain considered to be closed *is*, nor why precisely we should believe in (Closure). Second, Scott Sturgeon has argued that the Exclusion Argument is invalid because it equivocates between a broad and a narrow sense of "the physical domain." Finally, Sturgeon and others think that given the current state of science we cannot tell whether (Closure) holds for *any* physical domain. I review these problems and argue that for the purposes of debating the Exclusion Argument they can be avoided by formulating an alternative "quasi-closure" pertaining to the neurobiological domain.

Looking for causal closure? eBay has great deals on new & used electronics, cars, apparel, collectibles, sporting goods and more.
If you can't find it on eBay, it doesn't exist
– Sponsored link, encountered by the author during online research

6.1. Introduction

Subtleties aside, the Causal Closure of the Physical Domain – henceforth "(Closure)" – claims that all physical events have sufficient physical causes. If this is right, the physical domain contrasts sharply with the mental domain, as mental events often depend on physical causes. To assent to (Closure), then, is to ascribe a radical causal self-sufficiency to the physical domain. It is therefore not surprising that (Closure) is a central assumption shared by most reductive and antireductive physicalists alike. As such, it may be said to be a part of the *substance* of physicalism. But as we shall see, it also plays a role in the argumentative foundations of various forms of physicalism. In particular, it figures in the so-called Exclusion Argument for *reductive* physicalism, according to which mental events are identical with physical events. But what is the "physical" domain that physicalists take to be closed? Is it the broadly physical domain consisting of ordinary medium-sized objects? Or is it the narrowly physical domain of microphysics? Or is it

something else again? These questions are important not only because the content of a central physicalist claim turns on them, but also because of a challenge due to Scott Sturgeon (1998). Sturgeon argues that reflection on the empirical plausibility of (Closure) for the narrowly and the broadly physical domains undermines the Exclusion Argument. He also questions whether any physical domain is closed. I review this challenge and argue that more work has to be done to settle the domain question. However, for the purpose of the debate over the Exclusion Argument, Sturgeon's challenge can be met by shifting our attention from the science of physics to neuroscience. I show how the Exclusion Argument can be reformulated with a more limited, but empirically plausible, "quasi-Closure" pertaining to the neurobiological domain. To this end I adopt a historical approach and trace the emergence of current models of neural causation.

6.2. The Significance of Causal Closure

For our purposes (Closure) can be formulated as follows. *Any physical event that has a sufficient cause at t, has a sufficient physical cause at t*.[72] Why should we care about this principle? For one thing, it is a central tenet of physicalism, which is perhaps the most widely held world-view in contemporary analytic philosophy. By assenting to (Closure) we ascribe a radical causal self-sufficiency to the physical domain. Whereas there plausibly are mental events like perceptions that lack sufficient mental causes, we will never have to leave the physical domain to find sufficient causes of physical effects. (Kim, 2005, p. 16) On the other hand (Closure) does not rule out the possibility of some

[72] The closure principle, a.k.a. the completeness principle, has attracted a variety of formulations. According to Sturgeon "Every physical effect has a fully revealing, purely physical history." (Sturgeon, 1998, p. 413) To allow for indeterminism it is sometimes said that physical events have their *probabilities* fixed by prior physical events or states. (Papineau, 2001) Kim at one point considers a stronger version according to which physical events have *only* physical causes. (Kim, 2005, p. 50) But as he notes (2005, p. 51) this formulation has the disadvantage of saddling antireductionists with epiphenomenalism at the outset. An even stronger version demands that physical events also have only physical *effects*. While this principle may be relevant to understanding physical-to-mental causation in phenomena like perception, it can be set it aside for present purposes, as I will be focusing on mental causation in actions. See Lowe (2000) or Montero (2003) for more discussion of closure principles.

physical events like bodily movements *also* having irreducibly mental causes. As such, it can figure as a shared assumption for reductive and antireductive physicalists alike. In fact, (Closure) plays an important role both in arguments for physicalism as such, and in arguments for reductive physicalism. It is commonly assumed that the minimal commitments shared by reductive and antireductive physicalists are defined by some form of supervenience. (Jackson, 1998, ch. 1; Kim, 1998, ch. 1; Lewis, 1983) Supervenience is perhaps best viewed as a technical way of saying that all non-physical facts obtain in *virtue of* physical facts. (Witmer, 2001) Basically, supervenience means that worlds (or perhaps objects or regions)[73] that are indiscernible in physical respects are also indiscernible in all respects, including mental respects.[74] If physicalism is defined in this minimal way, it does not require (Closure). But even though physicalism is a broad church, most believers in physicalism are also believers in (Closure). In fact, (Closure) plays an essential role in a popular argument for supervenience. (Loewer, 2001a; Papineau, 1990) So if physicalism is understood in terms of supervenience, its fate may turn at least partly on that of (Closure). Anyway, Jaegwon Kim comes close to saying that only heretics would doubt (Closure), as "no serious physicalist could accept" the possibility of its failure. (Kim, 1998, p. 40) I shall follow Kim in taking (Closure) to be part of the physicalist orthodoxy, while bearing in mind that the status of the principle is interesting even if physicalism is characterized independently of it.

(Closure) also figures prominently in most formulations of what goes under the name of the Exclusion Argument for *reductive* physicalism.[75] This argument, on which I

---

[73] See Kim's "strong" and "weak" supervenience for two formulations in terms of indiscernible *objects*. (Kim, 1984a) Horgan's "regional" supervenience appeals to indiscernibility of spatiotemporal *regions*. (Horgan, 1982) Jackson (1998, ch. 1) and Lewis (1983) offer definitions of supervenience in terms of indiscernible *worlds*. This latter type of formulation, sometimes called "global supervenience," appears to be the most widely accepted one.

[74] There are reasons for doubting whether supervenience in itself is sufficient to capture the idea that everything non-physical obtains "in virtue of" claim of the physical. (See, e.g. Kim, 1998, pp. 9-15)

[75] Although Sturgeon (1998) treats the Exclusion Argument as an argument for physicalism *per se,* he formulates the argument as yielding the reductive conclusion that mental events are physical events, thus implicating that he has what I call reductive physicalism in mind.

shall focus here, starts from the plausible assumption that mental events like the occurence of beliefs and desires have physical effects like bodily movements. By (Closure), these bodily movements must have sufficient physical causes. Since bodily movements are – typically at least – not causally overdetermined by more than one sufficient cause, the mental cause must be identical with the physical cause.[76] (Cp. Kim, 1998, ch. 2; 2005, ch. 2; Papineau, 2001) Here is one way of formulating this argument:

> (Closure) Any physical event that has a sufficient cause at t, has a sufficient physical cause at t.[77]
>
> (Impact of the mental) Mental events have physical effects.[78]
>
> (No Overdetermination) The physical effects of mental events are not generally overdetermined.[79]
>
> ---
>
> (Reductionism) Mental events are physical events.[80]

Summing up, not only is (Closure) a strong physicalist claim in its own right; it also plays a role in the foundations of both reductive and antireductive physicalism. It therefore becomes a matter of some importance to find out (a) whether (Closure) holds and (b) if it does hold, what *is* the physical domain that is closed.

---

[76] Strictly speaking Kim's version of this argument contends that mental events are either physical or epiphenomenal. This argument nevertheless serves to motivate reductionism, as epiphenomenalism is typically taken to be an immensely unattractive option.

[77] As indicated above, this is not Sturgeon's formulation.

[78] The phrase "Impact of the mental" – henceforth "(Impact)" – as well as the present formulation is due to Sturgeon (1998, p. 414). As is often remarked, the Exclusion Argument's reductive conclusion does not apply to mental events that do not cause physical events, if such there be.

[79] This formulation originates with Sturgeon (1998, p. 414). The "generally" phrase allows for occasional overdetermination, and strictly speaking the reductionist conclusion would not apply to mental events involved in such cases.

[80] This could be interpreted either as a token-identity claim (every token mental event is identical with *some* token physical event), or as a type-identity claim (every type of mental events is identical with some type of physical events). If Kim's (1976) theory of events as property instantiations is assumed, token-identity entails type-identity.

6.3. The Empirical Status of Closure

What evidence is there for (Closure)? Somewhat surprisingly, given its almost universal acceptance by physicalists, the mental causation literature is relatively unforthcoming on why we should believe in the principle. This is unfortunate, as it is not an obvious truth. Arguments for (Closure) appear to fall into two classes. Some appeal explicitly to empirical evidence, whereas others construe (Closure) as following more or less straightforwardly from abstract explanatory features attributed to physics. According to the former type of argument (Closure) is far from obvious, and robust evidence for the principle only became available quite recently. According to the latter, "simpler" type, (Closure) turns out to be a relatively obvious truth which can be arrived at without much reflection on empirical evidence.

There are at least two simple arguments for (Closure).[81] First, Kim (1992) thinks that if (Closure) is violated by irreducibly mental causes, the laws of physics, too, would be violated. He takes such violation to be intolerable. Second, Kim frequently tells us that unless (Closure) holds, physics will not be "completable." Again, he takes this to be an unacceptable consequence. (Cp., e.g., Kim, 1993; 1998, p. 40) Now most physicalists would probably – and unlike antiphysicalists like Nancy Cartwright (1980) – follow Kim in taking the *truth* of the laws of physics for granted. Furthermore, physicalism is, perhaps, by its very nature wedded to the idea that physics is in some sense *completable*. Kim's arguments therefore appear to provide simple and relatively *a priori* routes to (Closure).

However, such simple arguments are probably not successful unless supplemented with substantial empirical evidence. For Brian McLaughlin (1992) and David Papineau (2001) argue persuasively that the laws of physics could remain true even if some physical effects lack sufficient physical causes. Briefly, laws like Newton's

---

[81] In addition to the two arguments I mention here, Donald Davidson (1980) might be taken to argue that unless (Closure) holds, the laws of physics cannot be maximally general and deterministic, as he thinks they must be. (See Ramberg (1999) for discussion of Davidson's concept of the physical and its connection with generality and determinism.) Davidson's argument faces basically the same problems as those mentioned below.

second, **F** = m**a,** appear to be "causally neutral"[82] insofar as they are indifferent to what forces or causes enter into them. **F** = m**a**, then, could remain true even if some accelerations were due to non-physical causes or forces.[83] If there are such causes, physics would presumably not be completable in the sense envisioned by Kim either. The problem with just asserting completability, as Kim appears to do, is this. As McLaughlin (1992) and Papineau (2001) make clear, causes like irreducibly vital forces have been proposed by scientists until quite recently. We shall encounter something similar when we considering the history of theorizing about neural signaling below. Scientists, in fact, appear to have changed their minds about the completability of physics several times over history. (Papineau, 2001) What is needed is therefore some sort of inductive argument from current theories to the effect that physics must be completable. We are left, then, with the first type of arguments for (Closure).

McLaughlin and Papineau offer such arguments. In outline they contend that irreducibly mental or vital causes of physical events are not needed because: (1) the quantum-mechanical explanation of chemical bonding historically undermined a prime case of emergent, non-physical forces. (McLaughlin, 1992) (2) The reduction of forces like friction to more fundamental physical forces inductively supports the claim that all forces reduce to fundamental physical forces. (Papineau, 2001) (3) Detailed and successful investigations of biological systems have not revealed any events that cannot be attributed to physical forces or causes. (Papineau, 2001)

Non-physicalists may harbor worries about this evidence. Some like Robin Hendry (2005) and Sturgeon (1998) doubt that there is conclusive evidence for (Closure). Others, like Cartwright (1999, ch. 1) think of scientific theories as models with limited scope, and are skeptical to any claims about theoretical completeness outside the models' current scope. I am not convinced that these problems cannot be effectively countered by physicalists. In fact, I am inclined to believe that authors like McLaughlin (1992) and Papineau (2001) make a strong empirical case for the claim that some (Closure) principle

---

[82] Cartwright (1979) uses a slightly different sense of "causal neutrality" to characterize laws like **F** = m**a**.

[83] Admittedly, this argument may turn out to depend on whether we chose to reify forces or not. See McLaughlin (1992, pp. 64-65) and Papineau (2001, p. 17n11; 2002, pp. 242-243) for discussion.

must hold. Nevertheless, the problems suggest that detailed considerations of empirical evidence are needed to fully assess the principle. And quite independently of this, we still need to consider empirical evidence to get clear about what "the physical domain" that is supposed to be closed is.[84] This question is interesting in its own right, but also because it is the starting point of Sturgeon's (1998) challenge to the Exclusion Argument.

6.4. Sturgeon's Challenge

The Exclusion Argument mentions the physical domain both in the (Closure) and (No Overdetermination) premises as well as in (Impact), the premise stating that mental events have physical effects. Ours and Sturgeon's central question is: what does "physical" mean in these premises? Under the plausible assumption that the premises must be supported by explanatory practice or common sense (1998, p. 412), Sturgeon argues that the Exclusion Argument equivocates. (Closure) and (Impact) are only plausible under different readings of the "physical." (1998, p. 415). According to Sturgeon (1998, pp. 415-416), the physical domain could mean the "narrowly physical" or microphysical domain, which he takes to be that of quantum mechanics. But it could also mean the "broadly physical" domain containing in addition to microphysics, macro objects like tables and rocks. I call the corresponding premises "(Broad-Closure)," "(Narrow-Closure)" and so on. Sturgeon argues that only (Broad-Impact) and perhaps, but *only* perhaps, (Narrow-Closure) are supported by current science or common sense. Hence, the alleged equivocation. These, then, are Sturgeon's central claims:

(S1) (Broad-Impact) is supported by science/common sense
(S2) (Narrow-Impact) is not supported by science/common sense
(S3) (Broad-Closure) is not supported by science/common sense

---

[84] This problem is related to, but different from, the one posed by "Hempel's Dilemma" to the effect that physicalism is either false or trivial, depending on how physicalists characterize the physical domain. (Crane & Mellor, 1990) Even assuming, as I do, that one or more of the proposed solutions to this dilemma will work – see, e.g., Melnyk (1997); Papineau (2001); Smart (1978) and Stoljar (2005) – it is still interesting to determine *which* physical domain is closed.

(S4) (Narrow-Closure) may or may not be supported by science/common sense

Consider briefly (S1) and (S2) first. Sturgeon contends that "everyday experience" and "macro science" indicate "that mental events have macrophysical effects", so (Broad-Impact) seems solid. (Sturgeon, 1998, pp. 416) However, he finds (Narrow-Impact) dubious. "No working scientific theory postulates a pervasive causal link between mental events and quantum events. Neither does commonsense." (Sturgeon, 1998, p. 415)

What about (S3) and (S4)? Sturgeon thinks that

> [(Broad-Closure)] is *not* part of extant science; nor is it part of everyday experience. No working scientific theory says broadly physical effects have fully revealing broadly physical histories. And neither does commonsense. Quite the contrary: both macro science and everyday experience rely upon mental causes for broadly physical effects. (Sturgeon, 1998, p. 416)

He portrays the evidence presented in favor of (Narrow-Closure) as follows: "[Quantum mechanics] says quantum events have their chances fully determined by quantum states. This is said to render the scientific bona fides of [(Narrow-Closure)] beyond question." (Sturgeon, 1998, p. 415). However, Sturgeon also believes that, given what we currently know about quantum mechanical theory, (Narrow-Closure) might be false. As he (1998, p. 426) – and, as far as I can tell, most popularized expositions – describes quantum mechanics, the development of a quantum system is subject to two rules. First, as long as the system is not being measured, its state evolves in accordance with the so-called wave function. This function can be obtained by solving the Schrödinger equation for the system. Strangely, the wave function allows the system to be in a "superposition" of several states. For instance, "if a particle can be located at P1 or P2 or P3, then it can also be characterized by a combination such as (1/3P1 + 1/3P2 + 1/3P3)." (Sturgeon, 1998, p. 425) Second, there is a "projection" rule stating that during measurement the wave function collapses into one of the superposed states, yielding a definite measured state, e.g., P1 or P2 or P3. Put briefly, the "measurement problem" in quantum mechanics is to say what happens to the system when it is being measured. Sturgeon argues that some, but not all, proposed solutions to this problem render (Narrow-Closure) false. For instance, according to Eugene Wigner's and Niels Bohr's interpretations, the collapse of

the wave function is due to an interaction between quantum mechanical systems and *consciousness* or *classical/macro physical systems*, respectively. But if any of these interpretations are correct, the collapse into a quantum state is caused by something that is not itself quantum mechanical. Hence, (Narrow-Closure) would fail under these interpretations. Since it is still an open question which interpretation is right, there is no conclusive evidence for (Narrow-Closure). (Sturgeon, 1998, pp. 427-428)[85]

Summing up, Sturgeon poses a two-fold challenge to the Exclusion Argument. First, there is the *Equivocation Problem*. If (S1)-(S4) are right, the Exclusion Argument is not valid as it stands, because it equivocates between the narrowly and the broadly physical. Second, if (S4) is right, there is also what we might call the *No Closure Problem*. That is, it is not even obvious that (Narrow-Closure) holds, so there might be *no* empirically plausible closure principle. This is a threat to any physicalist – reductive or antireductive – who endorses (Closure).

With respect to the Equivocation Problem, Sturgeon is well aware that even though (Broad-Closure) is not *postulated* by science or common sense, it might nevertheless be *argued* that broadly physical events have sufficient physical causes. (Sturgeon, 1998, p. 416) He therefore offers a revised Exclusion Argument, only to dismiss it.[86] This argument contains two additional assumptions. First, there is (Composition). *Broadly physical events are mereologically composed by narrowly physical events, that is, have narrowly physical events as their parts*. (Sturgeon, 1998, p. 417) In addition, one might think that to cause the parts composing an event is to cause the event itself. For instance, if you cause the bricks constituting a wall to fall down, would you not *eo ipso* also cause the wall to fall down? The argument's second assumption therefore provides a kind of mereological bridge from the micro to the macro,

---

[85] However, as Paul Noordhof (1999, p. 371n4) notes, Bohr's interpretation is consistent with (Broad-Closure), as the collapse would be caused by a broadly physical event.

[86] He actually offers *two* revised arguments. The second argument aims to show that mental events do have narrowly physical effects after all, as demanded by (Narrow-Impact). Sturgeon considers this argument, but dismisses it as invalid. (Sturgeon, 1998, p. 425n13) I shall set this aside, as the basic problems raised by Sturgeon can be appreciated from the first argument.

across which causal influence can flow. I call this (Upward Causation). *If C causes E and E composes into E*, then C causes E**. (Sturgeon, 1998, p. 417)

The argument then runs as follows.[87] By (Broad-Impact), mental events have broadly physical effects. By (Composition), these effects have narrowly physical events as their parts. Assuming (Narrow-Closure), these parts have sufficient narrowly physical causes. But, according to (Upward Causation), these narrowly physical causes are also causes of the broadly physical events. By (No Overdetermination) these events are not causally overdetermined by their narrowly physical and mental causes. So mental events must be – not broadly physical events – but *quantum mechanical* events. (Sturgeon, 1998, p. 417)

I can only provide an outline of Sturgeon's attempt to refute this argument here. This will nevertheless be instructive, as it shows what difficulties formulating the Exclusion Argument in terms of physics rather than neuroscience may lead to. Briefly, Sturgeon thinks that (Upward Causation) is false, because he holds a (Cause & Essence) principle. *C causes E iff C is sufficient to bring about what is essential to E*.[88] (Sturgeon, 1998, p. 422) This idea can be brought out by considering one of Sturgeon's counterexamples to (Upward Causation).

> 1000 ducks are on a lake. All are normal save $Duck_{10}$. $Duck_{10}$ is deaf. As it happens, $Duck_{10}$ is bitten by a turtle just as a shotgun is fired nearby. The flock takes flight en masse. (Sturgeon, 1998, p. 419)

Sturgeon contends that, contrary to (Upward Causation), the turtle bite does not cause the flock's flight, even though it causes a part of that flight, namely the flight of $Duck_{10}$. (Sturgeon, 1998, p. 369) The explanation is that to cause the flock's flight is to cause the essential number of individual duck flights. We do not treat the turtle bite as a cause,

---

[87] See Sturgeon (1998, p. 418). He subsequently revises the argument (1998, p. 424), but the differences do not matter for our purposes.

[88] As Noordhof (1999, p. 369n3) notes this may be a bit too strict, because C might qualify as a cause even though it only brings about a part of E's essence. (Provided, presumably, the remainder of the essence is brought about by other causes.)

because the flight of $Duck_{10}$ is inessential to the flock's flight. On the other hand we accept the claim that the shot causes the flock's flight, because it causes "enough salient duck-movements" (Sturgeon, 1998, p. 421), in accordance with (Cause & Essence).

As Paul Noordhof points out in his critique of Sturgeon, this does not at first blush seem like much of a problem for the Exclusion Argument. For quantum events might well be sufficient to cause "a certain minimum number of quantum events" (Noordhof, 1999, p. 371) that are essential to broadly physical events like bodily movements. However, Sturgeon's (1999) rejoinder makes it clear that the problem he was originally trying to raise was that quantum events may compose into broadly physical events and yet be *inessential* to them. That is, we have no

> […] right to assume micro phenomena are essential to macro phenomena. Maybe they are. Maybe they aren't. It's an open scientific issue. […] A systematic account of macro movements may ultimately see quantum events as compositional tag-alongs, inessential dust on irreducibly macro shoes. (Sturgeon, 1999, p. 378, see also 1998, p. 422)

The putative reason why quantum events may be inessential to bodily movements is that there is a huge "conceptual gap" between the quantum mechanical domain and the broadly physical domain. (Sturgeon, 1998, p. 422) And the reason why there is such a gap is that the spatial realities of the domains differ. The quantum mechanical image contains strange phenomena like superposed positions and projection. Particles can be in a superposed combination of several positions. Measuring them can project them into a definite position. Things like that do not happen in macro spatial reality.[89] (Sturgeon, 1998, p. 426) Summing up, the micro and macro levels are conceptually so far apart that we do not know whether the former is essential to the latter. But if it is not essential, there are no sufficient microphysical causes of broadly physical events. Thus, this version of the Exclusion Argument would collapse.

---

[89] Sturgeon may be taken to supplement this idea with some brief modal considerations about what kinds of counterfactual dependencies are relevant to micro causation and macro causation. I set this issue aside. (Sturgeon, 1998, pp. 428-429)

6.5. Lessons from Sturgeon's Challenge

There is a lot to be learnt from Sturgeon's challenge. Most importantly, his Equivocation Problem puts two constraints on the formulation of the Exclusion Argument. (i) *The Matching Constraint*. A *valid* Exclusion Argument requires that we find some way of matching the physical domains mentioned in (Closure) and in (Impact). Physical and mental causes had better cause the same events. (ii) *The Empirical Constraint*. A *sound* Exclusion Argument requires that domain to be such that (Closure) and (Impact) are empirically plausible. In hindsight, (i) may appear obvious, but Sturgeon's important contribution is to show that domain matching is not necessarily a trivial thing to achieve. The Empirical Constraint is supported independently by the apparent failure of simple explanatory arguments for (Closure), e.g., the argument that the failure of (Closure) would lead to the violation of physical laws. (Section 6.3 above)

In my view the Empirical Constraint should be read strongly, because it would be preferable to avoid the Cartwrightean worries about completeness claims mentioned above. (Section 6.3, Cartwright, 1999, ch. 1) One may of course have little patience with these worries. There is a fine line between sensitivity to the limitations of scientific models and inductive skepticism. Barry Loewer (2001a) argues that Cartwright crosses this line. But the Exclusion Argument would nevertheless be stronger if it could be run *within* actual scientific models or plausible expansions thereof. To be sure, a certain amount of induction is always involved when we claim that a theory – or some plausible expansion thereof – can actually, or in principle, explain what we have not already explained. But to make the Exclusion Argument's contention that effects like bodily movements can in principle be explained by physical causes plausible, we should as far as possible appeal to detailed considerations of *extant* theories of bodily movements. Now, there do not appear to be any extant quantum mechanical models of causes of bodily movements around, and I doubt whether they will be available any time soon. As we shall see, there *are*, however, good neurobiological models available. This motivates realigning our gaze from physics to neuroscience. At the very least, neuroscience may provide a supplemental approach to causal exclusion.

A second lesson from the challenge is this. Reductive and antireductive physicalists all need to debate the question of the domain of (Closure) more explicitly.

For on closer inspection there is striking disagreement as to whether (Broad-Closure) or (Narrow-Closure) or perhaps even both hold.[90] Consequently, it is not clear whether philosophers assent to the same thing when they assent to (Closure). Take (Broad-Closure) first. If it holds, the Equivocation Problem would vanish. And whereas Sturgeon argues that (Broad-Closure) is not supported by current knowledge, others disagree. Papineau (2001, p. 12), for instance, defines the physical as the "non-mental", thus implying that his Exclusion Argument can be run with a wide reading of "physical" in (Closure). And in a later criticism of Sturgeon he explicitly says that "*The* version of [(Closure)] that I take to be *defensible*, as I said, is the [(Closure)] of the inanimate." (2002, p. 44, my italics)[91] Similarly, Noordhof (1999, p. 372n4) questions whether the Exclusion Argument should be formulated in terms of the "more specific claim" (Narrow-Closure) rather than (Broad-Closure). Kim in one place contends that "Physicalism need not be, and should not be, identified with micro-physicalism." (Kim, 1998, p. 117) He also seems to imply that (Broad-Closure) holds because he assents to (Closure) (1998, p. 40), while simultaneously claiming that the physical domain includes things like "tables," "computers" and "biological organisms". (Kim, 1998, p. 113) The widespread talk about bodily movements – which appear paradigmatically broadly

---

[90] There are many logical possibilities here. Sturgeon defines the broadly physical as "the macrophysical *plus* the quantum mechanical." (1998, p. 415, my italics) This turns the narrowly physical into a proper subset of the broadly physical. Thus (Broad-Closure) need not entail (Narrow-Closure), nor the other way round. One could hold without the other, or they might both hold. One could even in principle define the "*purely* broadly physical" as the broadly physical minus the narrowly physical, and claim, e.g., that, (Pure-Broad Closure) holds without (Narrow-Closure) holding.

[91] This quote may be taken to imply that Papineau denies that (Narrow-Closure) holds. And indeed he goes on to say that "I don't think of quantum mechanics *per se* as asserting completeness, since the basic assumptions of quantum mechanics leave it open what forces (Hamiltonians) there are." (Papineau, 2002, p. 43n14) As we have seen, he (2001) nevertheless argues that there are no non-fundamental forces. Accordingly, he may be taking the failure of (Narrow-Closure) to be a mere possibility. In fact, he suggests that the Exclusion Argument could be sharpened to apply to, e.g., the biological and the chemical by defining "the physical" first as "the non-mental," then as "the non-biological" and finally as "the non-chemical". See, e.g. (Papineau, 2001, p. 11) This would ultimately reduce every domain to the purely physical domain, and suggests that he *does* endorse (Narrow-Closure).

physical – in the exclusion debate suggests that many others share this view. On the other hand a more recent claim of Kim's only adds to the confusion about the plausibility of (Broad-Closure): "It is only when we reach the fundamental level of microphysics that we are likely to get a causally closed domain." (Kim, 2005, p. 65) Here, then, Kim appears to deny (Broad-Closure).

With respect to (Narrow-Closure), some physicalists do seem to appeal primarily to this domain in their discussions of (Closure). For instance, McLaughlin (1992) and Loewer (2001a) both appeal to microphysical forces or causes. Andrew Melnyk, too, indicates that he is operating with a "strict sense of the 'physical'" (2003, p. 158) in his discussion of (Closure), thus apparently endorsing (Narrow-Closure).

This brief review of closure principles at play in the literature, then, reveals that people discussing, to a large extent, the very same (Closure)-related problems (e.g., the Exclusion Argument), appear to operate with different versions of (Closure). This will have to be amended. Furthermore, there is significant disagreement as to what the evidence for (Closure) actually shows, that is whether it supports (Narrow-Closure), (Broad-Closure), both, or no (Closure) at all. This calls for more detailed, and more explicit discussion of the empirical evidence presented by authors like McLaughlin (1992), Papineau (2001) and Sturgeon (1998). Like, I suspect many other philosophers of mind, I am in no position to assess this evidence, as it involves difficult questions from the philosophy of physics. It nevertheless seems safe to say that it is still early days with respect to these questions. As philosophers of mind we can apparently not take it for granted that (Broad-Closure) or (Narrow-Closure) can figure in an Exclusion Argument. I am not saying that the difficulties involved in formulating the Exclusion Argument in terms of physics cannot be sorted out, but they do suggest that it would be useful to supplement this approach with a different one. This is especially true, since we, as we shall see, can bypass the difficulties by appealing to neuroscience instead.

Finally, and perhaps most importantly, Sturgeon is surely right about one thing. There *is* a huge conceptual gap between mental causation and bodily movements on the one hand, and quantum mechanics on the other. Most current scientific approaches to mentality seem to disregard quantum mechanics. It is highly instructive to note that Sturgeon has taken us from the relatively simple Exclusion Argument into hairy

questions about the metaphysics of part-whole relations and quantum mechanics. What is the correct account of superposition and the measurement problem? Can the parts of something be inessential to that thing? Do you have to cause something's essence in order to cause that thing? The first question continues to plague trained philosophers of physics. The latter two are likely to keep metaphysicians engaged for years to come. I contend that what allowed Sturgeon to raise these problems is that we started out with a (Closure) principle formulated in terms of physics. In what follows I suggest a different, neurobiological approach that respects both the Matching Constraint and the Empirical Constraint.

6.6. Neurobiological quasi-Closure

Interesting as the challenging to (Closure) discussed in Sections 6.3-6.5 are, I shall argue that there is a version of the Exclusion Argument that does not turn on how they are resolved. In fact, I suspect that appealing to general (Closure) principles is not the most intuitive or plausible way of motivating exclusion. Picking up a neuroscience text book, and considering what we know about neural processes leading to bodily movements, *is*, however, arguably a good way of intuitively motivating exclusion. Given these neural causes, where do irreducibly mental causes enter the picture? Is there any room for them? Accordingly, I shall be arguing that the Exclusion Argument can be run with a much more limited "quasi-Closure" pertaining to the neurophysiological domain.[92] The rough idea is that setting aside perceptual causal input, and letting circumstances be "normal," neural events have sufficient neural causes. Since what counts as "normal circumstances" arguably does not depend on the presence of irreducibly mental causes, we can formulate the exclusion problem as follows. Mental causes compete with *neural* causes of bodily

[92] Others have also attempted to motivate exclusion by reference to neuroscience and without invoking a general closure principle. John Bickle (2003, ch. 3) argues that causal exclusion is an implicit part of current cellular/molecular neuroscience. Kim (2005, pp. 154-155) also offers a brief appeal to neural causes, and claims that this way of formulating exclusion does not presuppose a general (Closure) principle. However, there is clearly a need to look more closely at what such an Exclusion Argument looks like, and what evidence there is for it. This is what I propose to do.

movements. This may appear to render the resulting Neural Exclusion Argument trivial, but as we shall see it does not. The rest of this paper will be dedicated to spelling out and defending this idea.

6.6.1. Neurobiological quasi-Closure and a Neural Exclusion Argument

Notice first that the principle we might call "(Neuro-Closure)" – which results from substituting "neurobiological" for "physical" in (Closure) – is obviously *false*. It is false because the neurobiological suffers causal influence from the outside that may be divided into at least two (possibly overlapping) kinds: first, in phenomena like perception, non-biological entities like photons impinge on sensory receptors, thus working as "causal input" to neural events at the body's periphery. Second, there are "background conditions" like the presence of oxygen and the absence of ultra-strong electromagnetic fields on which normal biological processes depend, and, we may wish to say, *causally* depend.[93] Perhaps this is part of the reason why so much of the exclusion debate has been formulated in terms of physics rather than biology. But by adding a *ceteris paribus* clause indicating that circumstances are "normal" for organisms like ourselves, we can formulate a very different, and more limited closure principle, claiming roughly that: those neural events that are not directly caused by non-neural input have neural causes. These causes are not absolutely sufficient for their effects, but they are nevertheless sufficient in the circumstances. As Jerry Fodor once remarked in a different context, *ceteris paribus* laws "necessitate their consequents *when* their *ceteris paribus* conditions are discharged". (Fodor, 1990, p. 152, my italics) Let us cast these ideas into a principle.

[93] Is the division of non-neurobiological factors into causal input and background conditions ultimately pragmatic and tuned to an explanatory interest in phenomena like perception? Perhaps. But even if it is, all that matters to my argument is that the non-neurobiological factors on which neural causation depends do not include irreducibly but necessary mental factors, however the factors are classified.

(quasi-Closure) Assuming that background conditions are normal, neurobiological events that are not immediately caused by organism-internal or organism-external input to the nervous system have neurobiological causes that are sufficient (in the circumstances).[94]

The formulation of this principle may sound unnecessarily complicated, and a simpler formulation claiming that neural causation does not depend on irreducibly mental causes might do the same job. I shall nevertheless use the present formulation, because I think it mirrors the pattern of explanation in the neural models we shall consider. The principle may also appear to render the Neural Exclusion Argument I will develop in the following trivial. But as will become apparent when we consider the empirical evidence, it does not. Also note that in order for (quasi-Closure) to figure in this argument, the principle's neurobiological domain should be read as including the muscle contractions that constitute bodily movements. After all, in the philosophy of mind, the Exclusion Argument's primary target has been mental causes of bodily movements. Finally, neuroscience also draws on causes from other scientific fields like biochemistry. To the extent that such causes are included in neuroscientific explanations I shall count them as belonging to the neurobiological domain.

Before arguing for (quasi-Closure) I will clarify it and its role in the Exclusion Argument by making three points. (1) (*Quasi-Closure) is weaker than (Closure) in three ways*. First, *no* neurobiological event has absolutely sufficient neurobiological causes. This is due to the necessary but non-neurobiological background conditions mentioned above.[95] For instance, if my arm is restrained, the firing of motor neurons enervating it

---

[94] The distinction between organism-internal and organism-external input is needed because the nervous system also receives causal input from other bodily systems like the endocrinal or hormonal system.

[95] As will be recalled, Papineau (2001) also appeals to biology in his argument for a *general* closure principle. My more detailed considerations of neuroscience may therefore be taken to support his general principle. Especially so if the physical is understood extremely widely as the "non-mental" as Papineau suggests. So, note again that I am not denying that a more general closure principle holds. However, using [Footnote continued on next page]

may not cause a bodily movement, even if they normally would. Second, some neurobiological events lack neurobiological causes. Perceptual input at the body's periphery is again the most natural example.[96] Third, even the neurobiological events that have neurobiological causes do not have such causes at all prior times. Tracing the effects of a neurobiological event backwards will eventually take one outside the neurobiological domain, whereas (Closure) demands there be a physical cause occurring at any time t at which there is a sufficient cause at all.

(2) *(Quasi-closure) can figure in a Neural Exclusion Argument*. Provided the Empirical Constraint and the Matching Constraint are satisfied, the relative weakness of (Quasi-Closure) poses no problems to its use in a Neural Exclusion Argument. For the Empirical Constraint will be satisfied if: (a) there is empirical evidence in favor of the principle's claim that *ceteris paribus* (and setting aside causal input) neural events have sufficient neural causes. (b) The *ceteris paribus* conditions do not include irreducibly mental factors. Of course, the presence of mental states will "normally," perhaps even necessarily, accompany many neural states. So strictly speaking irreducibly mental events are in fact likely to figure in the *ceteris paribus* conditions for neural causation. However, there should be a *minimal* set of *ceteris paribus* conditions in which they do not figure. In a word, their inclusion is not *necessary*. Now, the function of (Closure) in the traditional Exclusion Argument is to ensure that there are sufficient *non-mental* causes of bodily movements that do not include mental events. But switching from the "physical" to the "neurobiological" by substituting (quasi-Closure) for (Closure) does nothing to change this. For in conjunction conditions (a) and (b) entail that there are neural causes of bodily

---

(quasi-Closure) we can see how exclusion can be formulated in terms of more limited, but actual scientific models. Furthermore, the background conditions under which neural causation takes place are likely to be of a theoretically heterogeneous nature, so (quasi-Closure) does not entail that any particular theory is absolutely closed.

[96] By perceptual input I mean the events that trigger signaling from sensory receptors to the central nervous system, not the actual perception, whatever that is. Note that such triggering depends on the receptors' being ready to signal. Such "readiness" arguably involves biological factors, like the presence of ions that can flow through the receptor when it opens. Strictly speaking, then, even perceptual input depends on *some* neurobiological causes.

movements that do not include – or depend on as background conditions – mental events. So (quasi-Closure) can figure as our closure principle, even though it does not provide us with absolutely sufficient neural causes.

The Matching Constraint requires that mental and neural events cause the same events. Focusing on behavioral output from the system, it requires that neural events cause bodily movements, just as we and Sturgeon take mental events to do. Now if the Empirical Constraint is satisfied, neural events surely cause *muscle contractions*. Do they also cause bodily movements? Yes. Here Sturgeon's objections appear to have no bite. Bodily movements can be complex and involve multiple muscles. But the way contractions of individual muscle fibers compose into contractions of the muscles that pull on our bones is well understood. This understanding was arrived at empirically, to be sure, but there is no huge conceptual gap between contractions/relaxations and bodily movements. This contrasts sharply with the relation between quantum mechanical events and bodily movements. If Sturgeon's (Cause & Essence) principle is right, to cause a bodily movement is to cause enough contractions and relaxations of individual fibers. Neural events do that. Since some individual fiber contractions and relaxations are essential to bodily movements, bodily movements fall squarely within the explanatory domain of neuroscience.[97] The Matching Constraint is satisfied. In conclusion, whatever difficulties there are with the Neural Exclusion Argument, they do not appear to be problems with the use of (quasi-Closure).

That is not to say that the Neural Exclusion Argument is ultimately sound. I am not claiming that. After all it must also include some principle of (No Overdetermination) to ensure that mental events *compete* with neural events for the status of being causes of bodily movements. For all I have said overdetermination of bodily movements by mental and neural events might be perfectly acceptable. But this difficulty is unrelated to (quasi-Closure), which is my focus here.

(3) *The domain to which the mental is reduced will be different*. If the argument is sound, the mental will thereby be reduced to the neurobiological, and not to the narrowly

---

[97] Of course, *some* fiber or muscle contractions caused by neural events may be inessential to a bodily movement. (Sturgeon, 1998, p. 420)

physical domain.[98] This is in my view something to be welcomed. My point is not that the conclusion that mental events are quantum events would not be philosophically interesting. It would indeed. But given the current state of science, neuroscience's claim for relevance to mental causation is arguably stronger than that of physics. Neuroscientists are currently trying to establish correlations between mental functions and neural regions with the aid of functional imaging techniques, lesion studies, single-cell recordings and the like. They are interested in how neural systems interact to produce behavior and in how specific neural circuits implement cognitive functions. A sound Neural Exclusion Argument would therefore imply that an actual, and much debated, line of research is reductive. I doubt whether the same can be said of the Exclusion Argument which appeals to physics properly construed, e.g., to quantum mechanics. There does not, after all, appear to be any developed science called "behavioral quantum mechanics." In many ways, the traditional Exclusion Argument, as it was formulated above, is a very simple solution to the question of reductionism. If sound, this Exclusion Argument would provide a convenient and simple way to reductionism from just three premises. But the simplicity has a down-side. The argument contends that the mental *must* reduce, without really indicating specifically *to what* it might reduce. On the other hand a Neural Exclusion Argument could be taken to support the claim that correlations between mental and neural phenomena currently studied by neuroscience are actually identities.

Summing up, (quasi-Closure) can figure in the following Neural Exclusion Argument.

---

[98] Given (quasi-Closure) it would strictly speaking be more correct to say that the mental reduces to something neurobiological *o*r to something in the non-neurobiological background conditions. The latter alternative appears unmotivated, however.

(quasi-Closure) Assuming that background conditions are normal, neurobiological events that are not immediately caused by organism-internal or organism-external input to the nervous system have neurobiological causes that are sufficient (in the circumstances)

(Non-Mental Background) A minimal characterization of the normal background conditions does not include irreducibly mental events

(Neural Impact) Mental events have neurobiological effects

(No Overdetermination) The physical effects of mental events are not generally overdetermined

---

(Reductionism) Mental events are neural events

6.6.2. Arguing for quasi-Closure

As indicated above, I need to argue (a), that there is empirical evidence in favor of (quasi-Closure), and (b), that its minimal *ceteris paribus* conditions do not contain irreducibly mental conditions. With respect to (a), the plausibility of (quasi-Closure) hinges on what is meant by "neurobiological," just as the plausibility of (Closure) will hinge, if Sturgeon is right, on whether the "physical" is given a broad or narrow reading. "Neurobiological" might of course mean many things, partly because neuroscientists study the nervous system at many levels of analysis. These levels include *inter alia cellular/molecular neuroscience*, which studies nerve cells and their chemical components, *systems neuroscience*, which focuses on complex neural circuits and *cognitive neuroscience*, which studies the neural processes underlying higher cognition. (See, e.g., Bear et al. 2001, pp. 13-14) I will focus on cellular/molecular neuroscience, and to a certain extent systems neuroscience, for which I do think (quasi-Closure) is empirically plausible. In contrast, cognitive neuroscience often relies on concepts from the cognitive/intentional vocabulary in its explanations, and hence does not appear to be even quasi-closed *vis-à-*

vis the mental.[99] One might think that we *ought* to formulate the argument in terms of cognitive neuroscience, since that may be the best currently available biological approach to behavior and mentality. However, the neural causes invoked in the Neural Exclusion Argument need not be those that provide us with the best biological explanations of behavior. All that matters is that they are *ceteris paribus* causally sufficient for them.

My argument for (quasi-Closure) consists of three steps. Taken together they indicate that we have relatively detailed models of neural causation that do not implicitly or explicitly rely on mental causes. First, I outline a "connectionist" view of the anatomical causal structure within which neural causation takes place. This causal structure is part of neuroscience's theoretical backbone, and fits well with the formulation of (quasi-Closure). Second, I argue that we currently know enough about the kinds of causes that are at play within the causal structure to exclude the contribution of irreducibly mental causes. Over history a theoretical picture of the kinds of causes and mechanisms that are relevant to neural causation has emerged, from what was once an empty, or very loosely filled canvas. It is the amount of detail that relatively recently filled the picture that turns (quasi-Closure) into more than an off-hand appeal to future or idealized explanations of bodily movements. Third, I argue more briefly that we have no reason to believe that irreducibly mental factors are included in the minimal background conditions. My exposition will in many ways be simplifying, but should be sufficient to render (quasi-Closure) plausible given extant science.

Step 1: The structure of neural causation. I begin, then, by arguing that, while (quasi-Closure) may seem like an *ad hoc* philosopher's construct, its view of neural causation is

---

[99] Notably, there is also representational lingo at play in cellular/molecular neuroscience, for instance the action potential (see below) is frequently referred to as a "signal," and there is also talk of intracellular "second messengers" and the like. Unless we find some way of naturalizing representation, such descriptions may perhaps be taken to be inherently mental. Thus (quasi-Closure) might be jeopardized even for the cellular/molecular domain of neuroscience. But at this low level it nevertheless seems clear that the representational descriptions of causes could be substituted with non-representational (e.g., biochemical and cellular) descriptions. We might not *want* to do so, but (quasi-Closure) only requires that we *can*.

mirrored in neuroscience's broad theoretical perspective on the nervous system. This is related to the picture, familiar even from the popular press, of the brain as a kind of "neural network." I will follow Kandel et al. (2000, pp. 7 & 23) in referring to this network theory as "connectionism."[100] More precisely, connectionism as I portray it involves two theoretical assumptions, early evidence for which is often attributed largely to Santiago Ramón y Cajal (1852-1934). In revised forms they still constitute parts of the backbone of neuroscientific thinking.[101]

First, Cajal provided important microscopic evidence for the *Neuron Doctrine*. This doctrine has been interpreted in many ways (Mundale, 2001), but we can view it as making two claims. (i) Anatomically, neurons are discrete entities that do not fuse with one another. (ii) Physiologically, they are fundamental signaling units of the brain. In contrast, the "reticularists" of the time, like Camillo Golgi (1843-1926) – whose staining technique Cajal paradoxically relied on – believed that the brain forms a continuous web or reticulum. (Finger, 1994, ch. 3; Jones, 1999) In effect, reticularists denied that the Cell Theory, proposed by Mathias J. Schleiden (1804-1881) and Theodor Schwann (1810-1882) in the 1830s, applied to the brain. According to this theory, cells are the fundamental functional building blocks of organisms. (Coleman, 1971, ch. 2) The Neuron Doctrine, then, appears symptomatic of – and may well have directly shaped – the theoretical focus on the structure and function of neurons in much of Twentieth-Century neuroanatomy and physiology.

From the point of view of neural causation and (quasi-Closure) the doctrine's impact is this. One neuron's becoming active ("firing") and signaling to another neuron becomes one of the most important neural events to be causally explained. In short, the firing of individual neurons becomes a privileged kind of causal-explanatory event that is located at the center of our theoretical picture.

---

[100] Care should however be taken to distinguish this network theory from the more abstract and less biologically realistic "connectionist" models in computer and cognitive science.

[101] As witnessed by their explicit introduction in, e.g., Kandel et al.'s influential *Principles of Neural Science*. (2000, ch. 2)

Second, Cajal also held a *Law of Dynamic Polarization*, constraining the flow of information in neurons. Basically, neural signals travel only one way, from the receiving sites of neurons (e.g., its dendrites or soma) to the end of its outgoing projection or axon through which it affects other neurons. In contrast, some reticularists like Golgi thought that information could flow in several directions. (See, e.g., Berlucchi, 1999; Rapport, 2005, ch. 9.) The physiologist Charles Scott Sherrington (1857-1952) proposed that this kind of neural communication happens at specialized regions he in 1897 called "synapses." (Finger, 1994, ch. 3) We now know that at synapses neurons are typically, as Cajal predicted, separated by a small "synaptic cleft." Interneuronal communication is typically mediated by chemicals called neurotransmitters. Such transmitters can have an excitatory or an inhibitory effect. That is, the release of neurotransmitter from the first, "presynaptic" neuron can either stimulate the second "postsynaptic" neuron to fire, or it can hinder it from doing so.

All in all, this outline of a connectionist theory constrains how neural causation is to be traced within neural networks. First, it tells us that one neuron's firing constitutes a privileged kind of event. Second, it indicates that – setting aside spontaneous firing – both efficient and preventing causes of this privileged firing-effect are to be found in the firing of other, presynaptic neurons. This view of the direction of neural causation is mirrored in the analysis of neural circuits, where a standard functional taxonomy divides neurons into sensory, motor and interneurons. (Kandel et al., 2000, p. 25) *Sensory neurons* carry sensory information – in particular from sensory receptors at the body's periphery – into the nervous system. *Motor neurons* synapse on muscle fibers and convey commands about movements to be executed. *Interneurons* mediate between sensory neurons, other interneurons and motor neurons. It is not a far leap from this connectionism to view the nervous system as largely *an input-processing-output* device. Information arrives through sensory neurons, is processed and sometimes stored (largely by interneurons), and may ultimately give rise to various sorts of output from the nervous system. The output which will concern us here is of course bodily movements.[102]

---

[102] The nervous system can also output to other bodily systems like the endocrinal system when it causes the release of hormones.

Returning to the Neural Exclusion Argument, this is pretty much what (quasi-Closure) tells us. Tracing the effect of a neuron's firing backwards will typically take us via sensory neurons to the outside of the neurobiological domain – as suggested in the "input" part of (quasi-Closure). Tracing them forwards may take us via motor neurons and muscle contractions, and once again, to the non-neurobiological domain. There are of course also background conditions on which such neurobiological causal chains depend. Some biological conditions – like the presence of normal cerebral blood flow and properly functioning glial cells – are among neuroscience's explicit explanatory concerns. Others, like the presence of normal pressure and the absence of bullets hitting the brain, are normally not of explanatory interest, and can be lumped into non-neurobiological *ceteris paribus* conditions. In spite of the necessary non-biological *ceteris paribus* conditions, neuroscientists go about explaining neural events – in particular neural signaling – in terms of neural events.

(quasi-Closure), then, appears to be in accordance with neuroscience's explanatory aims as they are exercised at the levels of individual neurons and neural circuits built out of these. Within certain limits, neural events are explainable by neural events. This pattern of explanation is likely to be widespread in the special sciences generally, and will perhaps not be perceived as very special. It is certainly true that specifying *ceteris paribus* conditions in any special science will typically take you outside the vocabulary of that science. (Fodor, 1990, p. 147n10) But as we shall see shortly, at least in the case of neuroscience the pattern of explanation is supplemented with a detailed theoretical grasp of the mechanisms by which neural causes must work. Furthermore, neuroscientists also have a partly *quantitative* grasp of how these mechanisms must act together to yield their effects. This lends support to the claim that neural causes are really sufficient in the circumstances, and that we have good models of how such causes work.

Step 2: Causation within this structure does not depend on irreducibly mental causes. The connectionism just sketched provides a framework – pitched largely at the cellular/circuit levels – within which neural events, including bodily movements, can hopefully be explained by other neural events. But this neuroscientific explanatory

ambition should of course not be confused with a neuroscientific achievement. There are two reasons why this ambition should be rendered plausible. First, even though neurons cause other neurons within the network to fire, the mechanisms underlying such neural communication might not be neural through and through. Historically, special vital or mental causes have been proposed to explain neural communication. Such causes are not taken very seriously today. But even after the demise of vitalism it is important to provide a sketch of what is known about the mechanisms, to make sure that (quasi-Closure) is plausible given current scientific knowledge, and not just an appeal to future or idealized explanations. Second, it would be foolish to believe that our understanding of neural causation is currently complete and immune to revision. (Indeed, I shall briefly consider ways in which it may currently be changing below.) It is therefore important to emphasize that we know enough about neural mechanisms to believe that changes in our conception of these mechanisms are not likely to include the addition of irreducibly mental or vital causal factors.

To render the above-mentioned explanatory ambition plausible I return to the metaphor of a causal picture. Such pictures can do three things. First, they depict the kinds of events that are of central explanatory interest. Second, they provide scientists with a grasp of what kinds of causes are relevant to these events. Finally, really good pictures include representations of the mechanisms through which causes depicted therein bring about their effects. A causal picture, then, can figure both as a causal-explanatory toolbox, and help scientists constrain and develop hypotheses. If we want to explain a phenomenon, we can delve into the picture/toolbox for possible causes and mechanisms and check whether they can be made to fit with the explanandum. Or we may look for something different, but similar to what is in the box. Causal pictures need not be static. If thinking outside the box works, we can take whatever we were thinking of and add it to the box. If it works much better and is *very* different, we may want to throw away the old box and start off with a new one. As we shall see the causal picture behind current theories of neural signaling is to an increasing extent specified at the molecular level, and includes things like neurotransmitters, receptors and ion channels as crucial causal players. This kind of causal picture is introduced in text books such as those of Bear et al. (2001) and Kandel et al. (2000), which arguably play a sociological role in shaping

scientific research. (Schaffner, 2006) The picture, then, can be used to explain what goes on within the causal structure I just sketched. In my view, it is the picture's explanatory strength which ultimately lends credence to (quasi-Closure) and its contention that we have good neural models of neural causation.

It is, however, a picture that is relatively new, and throughout history it has not always been obvious whether one could understand neural causation by purely neural/chemical means. As philosophers know, Descartes' biology offered putative explanations of nervous signaling – if I may use that word in a Cartesian context – in terms of a kind of hydraulic system involving the flow of "animal spirits." (Finger, 1994, ch. 2) In fact, to explain human actions he thought the immortal Soul had to move the pineal gland, thus changing the flow of the spirits. (See, e.g. Descartes, 1985, p. 340/AT XI 352, and McLaughlin, 1993, for historical discussion.) (quasi-Closure) and (Non-Mental Background) would definitely not apply to a causal picture which includes Cartesian Souls.

However, neuroscientist Eric Kandel (2006, ch. 5) appears to describe the history of theorizing about neural signaling as involving the gradual exclusion of such mental or vital causes from our picture. This should interest us presently not only because it supports (quasi-Closure), but also because if Kandel is right, that history may be an example of causal exclusion in scientific practice. And yet it certainly does not involve any appeal to absolutely sufficient physical causes, as in the traditional Exclusion Argument for reductionism about the mental. Rather, it is based on actual, but limited scientific models of neural communication.

The historical and empirical emergence of this "soulless" and non-vital causal picture includes: (i) the discovery that signaling within neurons is electrical in nature, (ii) quantitative measurements of this signal's properties, (iii) mechanistic explanations of the signal, and finally (iv), a chemical understanding of communication among neurons. I shall provide a rough outline of these developments, which jointly support (quasi-Closure). I have adopted a historical approach, here, because this serves to underscore that the Neural Exclusion Argument is not trivial or obvious. It has not always been clear that we have good neural models of neural causation, as (quasi-Closure) contends. Neither has it always been clear that neural causation does not depend on irreducibly

mental or vital factors, as (Non-Mental Background) contends. These are both empirical and non-trivial discoveries.

(i) *There was an accumulation of evidence that the signal traveling down the axon of individual neurons is electrical in nature*. This evidence includes Luigi Galvani's (1737-1798) discovery in the late Eighteenth Century that electrical stimulation of a frog's leg could cause the frog's leg to move. The discovery lead Galvani to suggest that muscle contractions are caused by naturally occurring "animal electricity" even outside the experimental situation. (Kandel, 2006, ch. 5; Finger, 1994, ch. 29) The idea that electricity occurs naturally in animals was for a while opposed by Alessandro Volta (1745-1827) and others, but ultimately, Galvani's basic ideas where accepted. (Piccolino, 1998; Schuetze, 1983) Marco Piccolino makes the point in a language friends of the Exclusion Argument might appreciate. "Galvani's work swept away from life sciences mysterious fluids and elusive entities like "animal spirits" and led to the foundation of a new science, electrophysiology." (Piccolino, 1998, p. 381)

Discoveries made by Nineteenth Century scientists like Emile du Bois-Reymond (1818-1896) and Hermann von Helmholtz (1821-1894) added significantly to the acceptance of the electrical nature of the nervous signal. (Kandel, 2006, ch. 5; Piccolino 1998) This means that at least some aspects of the signal we now call the action potential can be – and was – made subject of physical investigations.

(ii) *The electrical signal's properties and form were uncovered, and a view of how it encodes information emerged*. With the arrival of better technology, scientists were able to subject neural signaling to quantitative analysis. Helmholtz, for instance, managed to measure the speed of the nervous signal in 1859. (Kandel, 2006, p. 75; Piccolino, 1998) Most interestingly for our purposes, Edgar Douglas Adrian (1889-1977) recorded and amplified the action potential so that it could be visualized with the aid of an ink writer. At rest (see below) the inside of a neural membrane is more negative than the outside, yielding a "membrane potential" of approximately -65 mV. During the action potential this difference is reversed, and the membrane potential rises to approximately +55 mV before the resting potential is restored. Adrian's measurements of action potentials in sensory neurons revealed *inter alia* that the shape and amplitude of the resulting "spike"-shaped curve is highly similar independently of the intensity and nature

of the sensory stimulus. Action potentials are "all-or-none" rather than graded phenomena, and, furthermore, they look more or less the same across neurons. This stereotypicality is the feature of action potential that will concern us here, as it presents us with a puzzle about how action potentials encode information. If their shape and amplitude do not differ, how is the nervous system able to distinguish between different messages conveyed by them? The answer appears to be that information is encoded in firing rates – that is, in the number of action potentials per time unit – rather than in their amplitudes. For instance, Adrian found that the firing rates in sensory neurons increases with the intensity of the sensory stimulus. (Kandel, 2006, p. 78)

The discoveries of researchers like Helmholtz and Adrian have three implications for (quasi-Closure). First, neural signaling can be studied quantitatively, which many view as a hallmark of natural science. Quantitative measurements may therefore be viewed as demystifying signaling (but see below). Second, the discoveries serve to sharpen an important neurophysiological explanandum. We have already seen that, given the Neuron Doctrine, the signaling of individual neurons becomes a privileged kind of event. Given the further assumption that neural information is encoded in firing frequencies, the more specific event of neurons' firing with certain frequencies is given a special causal/explanatory status within our causal picture. Third, given the connectionism sketched above, where neural causation is portrayed as occurring within a neural network, this constrains what events are causally relevant in cellular/molecular neuroscience. For if an event is to affect what happens in a neural network, it must work through a mechanism that allows it to change the firing rates of neurons. As John Bickle makes the point:

> If action potential rate is the currency of neural causation and information exchange, then the only way an event can elicit neural change is by affecting the processes that underlie action potential generation in individual neurons. That is where the rubber meets the road. (Bickle, 2003, p. 59)

Now, developments (i) and (ii) may, perhaps, as a historical fact, have contributed to the demise of irreducibly vital or mental causes in neuroscience. But when taken in isolation, I do not think they should worry dualists much. Provided they have theoretical reasons for doing so dualists could still maintain that the electrical signal depends on

mental or vital causes for its initiation. In fact, the contemporary dualist E.J. Lowe suggests on theoretical grounds that neural processes need invisible "help" from irreducibly mental events to cause voluntary bodily movements. (Lowe, 2000)[103] Furthermore, such causes might perfectly well be described mathematically, or give rise to a mathematically describable phenomenon. To lend support to (quasi-Closure) developments (i) and (ii) must therefore be supplemented with evidence that the action potential can be accounted for by neurobiological causes without invoking such factors. In other words, the picture should be supplemented with depictions of neural mechanisms. Mechanisms add significantly to the explanatory power of causal pictures. The third and fourth theoretical developments involve the filling-in of mechanisms to the causal picture.

(iii) *Neurobiological models for the generation and conduction of nervous signaling were developed.* Julius Bernstein (1839-1917) described in 1866 the action potential as a "wave of negativity" traveling along the nerve. (Boring, 1950, ch. 2) Several questions had to be answered in order to provide a mechanistic explanation of this signal, now known as the action potential. As mentioned above, the inside of the neural membrane is negative relative to the outside when the neuron is at rest. How does this resting membrane potential arise? During the action potential, the membrane is briefly depolarized, that is its outside becomes negative relative to the inside. What causes this depolarization? How is it conducted down the axon without decreasing or failing? The *membrane hypothesis*, made famous by Bernstein, was an important first step toward answering these mechanism-related questions. (See Boring (1950, ch. 2); Kandel (2006, ch. 5) and Piccolino (1998) for discussions of Bernstein's hypothesis.) In 1902 Bernstein suggested that the resting potential could arise from an uneven distribution of ions in intra and extra cellular space. He knew the inside is dominated by negatively charged organic ions, and positively charged potassium (K^+) ions, whereas the outside is dominated by positive sodium (Na^+) ions and negative chloride (Cl^-) ions. Bernstein proposed that at rest, the membrane contains open ion channels that only allow

[103] Cp. paper (#3).

potassium to pass through. He then explained the resting potential in terms of potassium flow across the membrane. Initially, a diffusion force would push potassium from the inside through the potassium selective channels to the outside, because the concentration of potassium is lower there. But as the inside grows more negative due to the potassium efflux, an electromagnetic force would begin to pull the positive potassium ions back inside. The balancing of these forces would yield a membrane potential of -70 mV. Bernstein also suggested a mechanism explaining how the membrane could be depolarized during the action potential. When stimulated sufficiently, the membrane would become permeable to all ions, and the potential would change from -70 mV to 0 mV, yielding an action potential with an amplitude of 70 mV. While Bernstein's model turned out to be flawed in many ways, the important point for present purposes is that only physical and chemical causes are at play in it. The all-important flow of potassium is due to the physical influences of diffusion and electromagnetic forces. The balancing of these influences can even be modeled mathematically, using a principle from physical chemistry, called the Nernst Equation. (See, e.g., Bear et al., 2001, p. 64)

This explanation of the action potential was revised and supplemented by the *ionic hypothesis*, which is largely due to the work of Alan Hodgkin (1914-1998) and Andrew Huxley (b. 1917). (Kandel, 2006, ch. 5; Piccolino 1998) Working on the experimentally convenient giant axon of squids, Hodgkin and Huxley developed a precise and quantitative model of the action potential. Their measurements confirmed Bernstein's suggestion of -70 mV for the resting potential, but they found that the action potential rises to +40 m, yielding an amplitude of 110 mV, rather than Bernstein's 70 mV. This called for a revision of Bernstein's model.

Nevertheless, the revised explanation was fundamentally in terms of the flow of ions. The characteristic shape of an action potential recording suggests two phases. First there is a rise or upstroke, when the membrane's inside becomes positive relative to the inside, followed by a fall or downstroke, where the original resting potential is ultimately restored. Hodgkin and Huxley's measurements suggested that the upstroke could be due to the influx of positive sodium ions, whereas the downstroke is caused by the efflux of potassium ions. To explain this mechanism, they postulated the existence of voltage gated sodium and potassium channels. These channels would work as switches, allowing

sodium and potassium, respectively, to pass through the membrane at just the right times. They were called "voltage gated," because they were supposed to open as the result of the electrical field resulting from nearby depolarization. First, voltage gated sodium channels would open, and the influx of sodium ions would depolarize the membrane. Slightly later, these channels would close, and the efflux of potassium would take the membrane potential back to a negative level. (Specialized ion pumps transport sodium out and potassium back in to maintain the ionic concentrations necessary for the resting potential.) Once a portion of the axon is depolarized the field thereby created would open sodium channels further down the axon, thus explaining how the signal is conducted down the axon. The predictions of the model fit well with most measurements, provided biological parameters are appropriately set. Furthermore, the existence of voltage gated ion channels has since been confirmed. The details of these mechanisms are beyond the scope of this paper, but can be found in text books like Bear et al. (2001, ch. 3-4) and Kandel et al. (2000, ch. 7-9).

Returning to (quasi-Closure), the emergence of this model has two important consequences. First, it strongly supports the contention of (quasi-Closure) that neural signaling can be accounted for by neural causes found at the cellular/molecular level, and neural mechanisms involving these causes. Indeed, the model would seem to remove the theoretical need to postulate vital or mental causes of neural signaling. Such causes are, as it were, pushed out of our causal picture. In this connection it is very interesting to note that Eric Kandel's remark about Bernstein's historical contribution is also *exclusionist* in spirit:

> In a larger sense, Bernstein's formulation joined those of Galvani and Helmholtz in providing compelling evidence that the laws of physics and chemistry can explain even some aspects of how mind functions – the signaling of the nervous system and therefore the control of behavior. There was no *need* or *room* for "vital forces" or other phenomena that could not be explained in terms of physics and chemistry. (Kandel, 2006, p. 83, my italics.)

There is, presumably, no *need* to invoke irreducibly mental or vital forces, because physical causes are sufficient in the circumstances to explain the phenomenon.[104] (Provided, that is, they are not needed in the background conditions, see below.) But dualists like Lowe will hardly be convinced. Irreducibly mental causes might still be at play even though their contribution is invisible from the point of view of neuroscience. I shall discuss this problem in step 3 of my argument below.

The second implication of Hodgkin and Huxley's model is this. Not only are vital causes excluded as necessary from our causal picture, the picture is also made even more precise as the model tells us *what kinds of causes* are relevant to the generation of action potentials. Certainly the factors that affect whether ion channels open and the forces that determine how ions flow through these channels become neural causes *par excellence*. This specification of the causal picture is all the more important given that neurons' action potential rate is depicted as a privileged kind of effect. Finally, the model provides neuroscientists with a precise quantitative grasp of how such causes yield their effects. Gutkin & Ermentrout go so far as to say that the model's formalism "is the closest that neurophysiologists have to Newton's laws of motion, and it underpins almost all modern models of how neurons work." (Gutkin & Ermentrout, 2006, p. 999) The Hodgkin and Huxley model, then, is the kind of model that inspires great explanatory confidence.

(iii) *An understanding of the chemical nature of interneuronal communication emerged*. So far I have focused on signaling within neurons and the electrophysiological

---

[104] This, then, may perhaps – in retrospect at least – be construed as a case of causal exclusion within explanatory practice. Putative causes of neural signaling like vital forces are excluded on the grounds that there is no *need* to invoke them. And as Kandel's quote suggests, there might be no *room* for them either, perhaps because they would yield a different effect. (See paper (#1) for further discussion.) But to repeat: I am not saying that the Neural Exclusion Argument as applied to mental causes in actions is ultimately sound. Mental causes of voluntary bodily movements need not be analogous with the mental or vital forces once postulated to explain neural signaling. We have independent reasons for believing that mental events cause bodily movements, whereas the postulation of mental/vital forces to explain neural signaling may well have lacked independent motivation. Furthermore, as discussed in paper (#1) and (#2) it is far from clear whether mental causes could be excluded on the grounds that they distort the effect, because it is far from clear that mental causes must be "productive" causes.

understanding of this phenomenon. But how does the action potential affect other neurons when it arrives at the end of the axon? I shall be briefer here than with the exposition of signaling within neurons. Electrophysiologists like John Eccles (1903-1997) long assumed that communication *between* neurons also had to be electrical. In contrast, physiologists working on the autonomic nervous system brought pharmacology into the debate and argued that neurons communicated with the aid of chemical signaling molecule – the above-mentioned neurotransmitters – that were dumped into the synaptic cleft. At the time, these scientists were jokingly referred to as "soups" since they believed that what I call a causal picture had to be supplemented with chemical "soup." The former scientists were called "sparks," because they thought that the picture could be painted using only electrical causes of neurotransmission. (See Kandel, 2006, ch. 6; Valenstein, 2005, ch. 8.) We now know that the sparks where right about certain synapses called "gap junctions" or "electrical" synapses. Here an electrical signal passes through a specialized channel linking the pre and postsynaptic neurons. Interestingly, this discovery can be viewed as a partial vindication of Golgi-style reticularism. (Bullock et al., 2005; Rapport, 2005, ch. 10) Nevertheless, as Kandel and Elliot S. Valenstein explain, arguments from soups like Otto Loewi (1873-1961) gradually made the soups' theory of chemical transmission the accepted view for most synapses. Chemical neurotransmission has since been the subject of intense research. The details cannot concern us here, except to note that the arrival of the action potential at the axon's terminal causes the release of neurotransmitter molecules into the synaptic cleft. These diffuse across the cleft and bind to specialized receptor molecules on the postsynaptic neuron's membrane. In "fast" or "ionotropic" receptors, this causes the receptor to open and allow certain ions to flow across the membrane, producing a change in the membrane potential. Excitatory neurotransmitters cause a depolarization of the membrane, known as an Excitatory Postsynaptic Potential (EPSP). On the other hand, inhibitory neurotransmitters cause a hyperpolarization – where the membrane potential becomes more negative – called an Inhibitory Postsynaptic Potential (IPSP). Notably, these changes in the postsynaptic membrane potential are graded, i.e., vary in amplitude, which contrasts with the all-or-nothing action potentials. The postsynaptic neuron integrates excitatory and inhibitory inputs and "determines" whether it should fire or not by spatial and temporal summation

of incoming IPSPs and EPSPs. (The activation of so-called "slow" or metabotropic receptors can initiate a series of intracellular events leading, e.g., to changes in the strength of synapses.)

Abstracting from these details, pharmacological or biochemical causes are added to our causal picture. These soupy causes bring with them a bewildering array of soupy mechanisms that directly or indirectly affect neurons' behavior. Central in the composition of the picture is now the all-important event of neurons' firing at certain rates. It is surrounded by other events that contribute towards it, like the release of neurotransmitters, and the opening of ion channels, as well as sparky and soupy mechanisms connecting such events. But as important as what the picture contains, is what it does not contain. The picture is exquisitely complex, but neurobiological and chemical through and through.

S*tep 3: Non-mental Background Conditions*. We have, then, ample reasons for believing that in normal circumstances neural events like bodily movements have sufficient neural causes. Whereas Sturgeon may be right that it is an open question whether quantum mechanical events cause bodily movements, it is not an open empirical question whether neural events do so. For this claim is grounded in extant, and detailed scientific models. (That is the primary reason why we waded through all this detail in the first place.) But unlike the Quantum Mechanical Exclusion Argument, the Neural Exclusion Argument offers no hope at all of finding absolutely sufficient causes of bodily movements within the domain considered. Accordingly, we must now briefly consider whether the minimal conditions under which neural causation takes place include irreducibly mental factors.

Offering a demonstrative argument for the absence of such causes seems practically impossible. For one thing it is in the nature of *ceteris paribus* generalizations that their *ceteris paribus* conditions cannot be listed. Second, as Lowe (2000) plausibly argues, there is always a logical possibility that the sufficiency we seem to find at the neural level is only apparent. Irreducible, but necessary, mental causes might be at play, even though their contribution is invisible from the point of view of neuroscience, or

from physics for that matter. Nevertheless I shall offer a brief plausibility argument to rule out such causes.[105]

First, what we do know about the *ceteris paribus* conditions does not suggest that any irreducibly mental causes must be included. We know that neural causation depends on things like normal temperature, blood flow, the presence of properly functioning glial cells etc. These factors are not mental in nature. The problem with claiming that mental causes must be included in addition is that neural causation appears to run normally in systems where we are not inclined to believe that mentality is present at all, e.g. in primitive animals, or even in tissue cultures. Why would they be needed in humans? Second, given that there is no direct empirical evidence for the presence of such irreducibly mental events – a point Lowe (2000) for one, agrees to – their postulation seems wholly *ad hoc*. The same reasoning applies to perceptual input to the neurobiological domain.[106]

6.6.3. Extrapolating from these Models

One final challenge to (quasi-Closure) remains to be addressed. Our interpretation of extant neuroscientific models should avoid the extreme of claiming that we now have arrived at the ultimate truth. Given the unfortunate fate to which such claims have fallen over history, it would be naïve to believe that the causal picture we have been considering will not be revised and expanded in a variety of ways. Indeed, even my own relatively cursory reading of contemporary scientific and popular journals suggests that the picture is *currently* being challenged and changed.

Here are a few examples. The Neuron Doctrine does not *imply* that glial cells are nothing more than a support team for the more important neurons. Nevertheless, it appears to have carried with it a highly "neuron-centric" theoretical outlook on neural signaling. However, some scientists now argue that the interneuronal communication-

---

[105] I provide some further details in paper (#3).

[106] In fact, we have a good understanding of how perceptual input works, and what kinds of input can activate sensory receptors. (See, e.g., Kandel et al., 2000, ch. 21) A distinguishing feature of good models is arguably that they include details about the nature of input to the system.

system we have been considering both interacts with, and is supplemented by a different, slower, kind of communication system involving glial cells. (Bullock et al., 2005, Fields, 2004) Others have suggested that the standard model of interneuronal communication where neurotransmitter is released at specialized synapses must be supplemented with so-called "ectopic release." Here, neurotransmitters can be released at sites remote from the synapses and diffuse along neural membranes. (Coggan et al., 2005) Furthermore, electrical synapses, the "exception" to the Neuron Doctrine noted above, have been thought to be relatively primitive devices, invoked by nature primarily when speed and safety of transmission are crucial. But there is now evidence that electrical synapses, like their more famous chemical cousins, are capable of plasticity or changes in synaptic strength. They may therefore play more sophisticated roles than was originally thought. (Bullock et al., 2005). Finally, even the celebrated Hodgkin and Huxley model of the action potential may not fit the behavior of all neurons. This has lead some scientists to propose a more sophisticated model including a kind of "cooperation" between ion channels, as discussed by Gutkin & Ermentrout (2006). These brief sketches of hypotheses under current discussion, then, suggest ways in which the causal picture may be modified, and supplemented with novel causes and mechanisms.

While we must be sensitive to hypotheses like these, as well as to the possibility of even more radical changes, our interpretation of current models need not fall into the alternative extreme of skepticism about scientific progress. Although changes and revolutions most likely will recur in the future, some doors are closed as science moves – more or less unwaveringly – forward. I contend that necessary but irreducibly mental causes of neural events are likely to be left on the outside for good. Consider the hypotheses I just mentioned. While some of them may change the causal picture significantly, they bear a certain family resemblance to it. For one thing the additional causes and mechanisms proposed are still firmly located within the biological domain. And even if standard interneuronal communication requires in addition a different kind of communication involving glial cells, there is still communication going on between the components of the nervous system. On the other hand, the proposed additions to the picture seem nothing like the mental or vital forces once invoked. Will further research on these or other, perhaps more radical hypotheses that may be forthcoming, lead to the

return of (say) vital causes of neural signaling? Given its empirical nature, this question is likely beyond the scope of conclusive arguments. Nevertheless, based on the neuroscientific causal picture just sketched, and the ways in which it appears to be changing, I believe the answer must be in the negative.

6.7. Conclusions

What is closed in causal closure? This seemingly innocent question is interesting in its own right, and we have witnessed how it appears to threaten the Exclusion Argument if this argument is formulated in terms of physics. More work probably needs to be done, and should be done, on the domain of (Closure). However, I have argued that the difficulties for the Exclusion Argument can be avoided by appealing to neuroscience and (quasi-Closure) instead of physics. (Quasi-Closure) may appear to render he Neural Exclusion Argument trivial. But trivial or not, it will, if I am right, remove the important problem Sturgeon has raised for the Exclusion Argument. This certainly speaks in its favor. Furthermore, I do *not* take it to be trivial. It is not an off-hand appeal to in-principle explanations of an idealized future neurophysiology. After all, its plausibility derives from extant models of causal processes leading up to, *inter alia* bodily movements. Accordingly, even philosophers like Cartwright (1999) and Tim Crane & D.H. Mellor (1990), who are in general skeptical to claims about in-principle completeness should find (quasi-Closure) worthy of their attention. While the impact of (quasi-Closure) is the same as that of the traditional version of (Closure) – there are sufficient non-mental causes of bodily movements – the Neural Exclusion Argument is not wedded to the idea that any particular, privileged science is credited with absolute completeness. The background conditions under which neural causation takes place, may after all be of a theoretically quite heterogeneous nature. In short, if the remaining premises of the Neural Exclusion Argument are sound, causal exclusion could be run within actual, but limited causal models.

References

Adam, C. & Tannery, P. (Eds.) (1964-1976). *Oeuvres de Descartes*. Paris: Vrin/CNRS.

Andersen, P.B., Emmeche, C., Finnemann, N.O. & Christiansen, P.V. (2000). *Downward causation: Minds, bodies and matter*. Aarhus: Aarhus University Press.

Anscombe, G.E.M. (1993). Causality and determination. In E. Sosa & M. Tooley (Eds.), *Causation* (pp. 88-104). Oxford: Oxford University Press.

Arnsten, A. (2003). Patricia Goldman-Rakic: A remembrance. *Neuron*, *40*, 465-470.

Aronson, J. (1971). On the grammar of 'cause'. *Synthese*, *22*, 414-430.

Ayer, A.J. (1954). Freedom and necessity. In A.J. Ayer, *Philosophical essays* (pp. 15-23). London: Macmillan.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829-839.

Bear, M.F., Connors, B. & Paradiso, M. (2001). *Neuroscience: Exploring the brain*. Baltimore, Md: Lippincott Williams & Wilkins.

Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs*, *37*, 471-497.

Berlucchi, G. (1999). Some aspects of the history of the law of dynamic polarization of the neuron. From William James to Sherrington, from Cajal and Van Gehuchten to Golgi. *Journal of the History of the Neurosciences*, *8*, 191-201.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA: MIT Press.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht: Kluwer Academic Publishers.

Bickle, J., Mandik, P., Landreth, A., (2006). The philosophy of neuroscience. *The Stanford Encyclopedia of Philosophy*, *Spring 2006 Edition* (E.N. Zalta, Ed.). Retrieved August 29, 2006, from http://plato.stanford.edu/archives/spr2006/entries/neuroscience/

Blake, W. (1966). *Complete writings with variant readings* (G. Keynes, Ed.). London: Oxford University Press.

Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research*, *67*, 133-150.

Boring, E.G. (1950). *A history of experimental psychology*. New York: Appleton-Century-Crofts.

Broad, C.D. (1925). *The mind and its place in nature*. London: Routledge and Kegan Paul.

Bullock, T.H., Michael, V.L.B., Johnston, D., Josephson, R., Marder, E. & Fields, R.D. (2005). The neuron doctrine, redux. *Science*, *310*, 791-793.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*, 73-121.

Burge, T. (1986). Individualism and psychology. *Philosophical Review*, *95*, 3-45.

Burge, T. (1989). Individuation and causation in psychology. *Pacific Philosophical Quarterly*, *70*, 303-322.

Burge, T. (1992). Philosophy of language and mind: 1950-1990. *The Philosophical Review*, *101*, 3-51.

Burge, T. (1993). Mind-body causation and explanatory practice. In J. Heil & A. Mele (Eds.), *Mental Causation* (pp. 96-120). Oxford: Clarendon Press.

Burge, T. (2003). Epiphenomenalism: Reply to Dretske. In M. Hahn & B. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of Tyler Burge* (pp. 397-403). Cambridge, MA: MIT Press.

Campbell, D.T. (1974). 'Downward causation' in hierarchically organised biological systems. In F.J. Ayala & T. Dobzhansky (Eds.), *Studies in the philosophy of biology* (pp. 179-186). Berkeley, Los Angeles: University of California Press.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, *4*, 20-40.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, *13*, 419-437.

Cartwright, N. (1980). Do the laws of physics state the facts? *Pacific Philosophical Quarterly*, *61*, 75-84.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Churchland, P.M. & Churchland, P.S. (1991). Intertheoretic reduction: A neuroscientist's field guide. *Seminars in the Neurosciences*, *2*, 249-256.

Churchland, P.S. (1980). Language, thought, and information processing. *Noûs*, *14*, 147-170.

Churchland, P.S. & Sejnowski, T.J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Coggan, J.S., Bartol, T.M., Esquenazi, E., Stiles, J.R., Lamont, S., Martone, M.E., Berg, D.K., Ellisman, M.H. & Sejnowski, T.J. (2005). Evidence for ectopic neurotransmission at a neuronal synapse. *Science*, *309*, 446-451.

Coleman, W. (1971). *Biology in the nineteenth century: Problems of form, function, and transformation*. New York: John Wiley & Sons.

Collins, J., Hall, N. & Paul, L.A. (Eds.) (2004a). *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Collins, J., Hall, N. & Paul, L.A. (2004b). Counterfactuals and causation: History, problems and prospects. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and Counterfactuals* (pp. 1-57). Cambridge, MA: MIT Press.

Crane, T. & Mellor, D. H. (1990). There is no question of physicalism. *Mind*, *99*, 185-206.

Craver, C.F. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science Supplemental*, *69*, S83-S97.

Craver, C.F. (2003). The making of a memory mechanism. *Journal of the History of Biology*, *36*, 153-195.

Craver, C.F. & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P.K. Machamer, R. Grush & P. McLaughlin, P. (Eds.), *Theory and method in neuroscience* (pp. 112-137). Pittsburgh, PA: University of Pittsburgh Press.

Creary, L.G. (1981). Causal explanation and the reality of natural component forces. *Pacific Philosophical Quarterly*, *62*, 148–157.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and, the human brain*. New York: G.P. Putnam.

Davidson, D. (1980). Mental events. In D. Davidson, *Essays on Actions and Events* (pp. 207-227). Oxford: Clarendon Press.

Descartes, R. (1985). *The Philosophical Writings of Descartes vol. I* (J. Cottingham, R. Stoothoff & D. Murdoch, Trans.). Cambridge: Cambridge University Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dretske, F. (2003). Burge on mentalistic explanations, or why I am still epiphobic. In M. Hahn & B. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of Tyler Burge*. Cambridge, MA: MIT Press.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Fair, D. (1970). Causation and the flow of energy. *Erkenntnis*, *14*, 219-50.

Feyerabend, P.K. (1988). How to be a good empiricist – A plea for tolerance in matters epistemological. In M. Curd & J. Cover (Eds.), *Philosophy of science: The central issues* (pp. 922-949). New York: W.W. Norton.

Field, H. (2003). Causation in a physical world. In M. Loux & D. Zimmerman (Eds.), *Oxford handbook of metaphysics* (pp. 435-460). Oxford: Oxford University Press.

Fields, D.R. (2004). The other half of the brain. *Scientific American*, *290*, 55-61.

Finger, S. (1994). *Origins of neuroscience: A history of explorations into brain function*. New York: Oxford University Press.

Fodor, J. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese*, *28*, 97-115.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. (1990). Making mind matter more. In J. Fodor, *A theory of content and other essays* (pp. 137-159). Cambridge, MA: MIT Press.

Fodor, J. (1991a). A modal argument for narrow content. *Journal of Philosophy*, *88*, 5-26.

Fodor, J. (1991b). You can fool some of the people all of the time, everything else being equal; hedged laws and psychological explanations. *Mind*, *100*, 19-34.

Galaaen, Ø.S. (2006). Historien gjentar seg: Hva kan vi lære av Descartes' svar til Elisabeth? (History repeats itself: Lessons from the Descartes-Elisabeth correspondence). *Norsk Filosofisk Tidsskrift (Norwegian Journal of Philosophy)*, *41*, 229-240.

Garcia-Molina, H., Ullman, J.D. & Widom, J. (2002). *Database systems: The complete book*. New Jersey: Prentice Hall.

Genovesio, A., Brasted, P., Mitz, A. & Wise, S. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, *47*, 307–320.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*, 49-71.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science Supplemental*, *69*, S342–S353.

Goldman-Rakic, P. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum (Ed.), *Handbook of physiology: The nervous system*, *section 1, vol. 5, part 1* (pp. 373-417).Bethesda: American Physiological Society.

Goldman-Rakic, P. (1992). Working memory and the mind. *Scientific American*, *267*, 110-117.

Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron*, *14*, 477-485.

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience*, *4*, 847-850.

Gutkin, B. & Ermentrout, G.B. (2006). Spikes too kinky in the cortex? *Nature*, *440*, 999-1000

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals* (pp. 225-276). Cambridge, MA: MIT Press.

Hansen, C.M. (2000). Between a rock and a hard place: Mental causation and the mind-body problem. *Inquiry*, *43*, 451-491.

Hedström, P. & Swedberg, R. (1998). Social mechanisms: An introductory essay. In P.

Hedström & R. Swedberg (Eds.), *Social mechanisms: An analytical approach to social theory* (pp. 1-31). Cambridge: Cambridge University Press.

Hendry, R. (2005). Is there downward causation in chemistry? In D. Baird, L. McIntyre & E. Scerri (Eds.), *Philosophy of chemistry: Synthesis of a new discipline* (pp. 173-189). Dordrecht: Kluwer Academic Publishers.

Horgan, T. (1982). Supervenience and microphysics. *Pacific Philosophical Quarterly*, *63*, 29-43.

Hume, D. (1978). *A treatise of human nature*. Oxford: Clarendon Press.

Jackson, F. (1996). Mental causation. *Mind*, *105*, 377-413.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Jackson, F. & Pettit, P. (1988). Functionalism and broad content. *Mind*, *97*, 381-400.

Jammer, M. (1957). *Concepts of force: A study in the foundations of dynamics*. Cambridge, MA: Harvard University.

Jones, E.G. (1999). Golgi, Cajal and the neuron doctrine. *Journal of the History of the Neurosciences*, *8*, 170–178.

Kandel, E.R. (2000). The molecular biology of memory storage. A dialog between genes and synapses. In H. Jörnvall (Ed.), *Nobel lectures, physiology or medicine 1996-2000* (pp. 392-439). Singapore: World Scientific Publishing Co.

Kandel, E.R. (2006). *In search of memory: The emergence of a new science of mind*. New York, London: W. W. Norton & Company.

Kandel, E.R., Schwartz, J.H. & Jessell, T.M. (Eds.) (2000). *Principles of neural science*. New York: McGraw-Hill.

Kenny, A. (1970). *Descartes, philosophical letters*. Oxford: Clarendon Press.

Kim, J. (1976). Events as property exemplifications. In M. Brand & D. Walton (Eds.), *Action theory* (pp. 159-177). Dordrecht, Holland: D. Reidel Publishing Co.

Kim, J. (1984a). Concepts of supervenience. *Philosophy and Phenomenological Research*, *45*, 153-176.

Kim, J. (1984b). Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy*, *9*, 257-270.

Kim, J. (1989). Mechanism, purpose and explanatory exclusion. *Philosophical Perspectives*, *3*, 77-108.

Kim, J. (1992), Downward causation in emergentism and nonreductive physicalism. In A. Beckermann, J. Kim & H. Flohr (Eds.), *Emergence or reduction? Essays on the prospects of nonreductive physicalism* (pp. 119-138). Berlin, New York: Walther de Gruyter.

Kim, J. (1993). The non-reductivist's troubles with mental causation. In J. Heil & A. Mele (Eds.), *Mental Causation* (pp. 189-210). Oxford: Clarendon Press.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.

Kim, J. (2000). An interview with Jaegwon Kim: Fall 2000. *ephilosopher*. Retrieved August 29, 2006, from

http://www.ephilosopher.com/modules.php?op=modload&name=Sections&file=index&req=viewarticle&artid=3

Kim, J. (2002a). Response to Barry Loewer. *Philosophy and Phenomenological Research*, *65*, 674-677.

Kim, J. (2002b). The layered model: Metaphysical considerations. *Philosophical Explorations*, *5*, 2-20.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton: Princeton University Press.

Kim, J. (forthcoming). Causation and mental causation. In B. McLaughlin (Ed.), *Contemporary debates in philosophy of mind*. Oxford: Blackwell.

Kistler, M. (1998). Reducing causality to transmission. *Erkenntnis*, *48*, 1-24.

Kuhn, T. (1964). *The structure of scientific revolutions*. Chicago: Phoenix Books.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Lee, D. (2001). *Cognitive linguistics, an introduction*. Oxford: Oxford University Press.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*, 343-377.

Lewis, D. (1986). Causation. In D. Lewis, *Philosophical papers, vol. II* (pp. 159-213). New York: Oxford University Press.

Lewis, D. (2004). Void and object. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and counterfactuals* (pp. 277-290). Cambridge, MA: MIT Press.

LHermitte, F. (1986). Human autonomy and the frontal lobes. Part II: Patient behavior in complex and social situations: The "Environmental Dependency Syndrome." *Annals of Neurology*, *19*, 335-343.

Loewer, B. (2001a). From physics to physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its Discontents* (pp. 37-56). Cambridge: Cambridge University Press.

Loewer, B. (2001b). Review of mind in a physical world. *Journal of Philosophy*, *98*, 315-324.

Loewer, B. (2002). Comments on Jaegwon Kim's mind in a physical world. *Philosophy and Phenomenological Research*, *65*, 655-662.

Loewer, B. (forthcoming). Mental causation: The free lunch. In B. McLaughlin (Ed.), *Contemporary debates in philosophy of mind*. Oxford: Blackwell.

Loewer, B. & Lepore, E. (1987). Mind matters. *Journal of Philosophy*, *84*, 630-642.

Lowe, E.J. (1999). Self, agency and mental causation. *Journal of Consciousness Studies*, *6*, 225-239.

Lowe, E.J. (2000). Causal closure principles and emergentism. *Philosophy*, *75*, 571-585.

Lowe, E.J. (2003). Physical causal closure and the invisibility of mental causation. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 137-154). Exeter: Imprint Academic.

Lučić, V. & Baumeister, W. (2005). Monte Carlo places strong odds on ectopic release. *Science*, *309*, 387-388.

Machamer, P.K., Darden, L. & Craver, C.F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1-25.

Mackie, J.L. (1993). Causes and conditions. In E. Sosa & M. Tooley (Eds.), *Causation* (pp. 33-55). Oxford: Oxford University Press.

Malenka, R.C. & Bear, M.F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*, 5-21.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W.H. Freeman.

McDowell, J. (1994). *Mind and world.* Cambridge, MA: Harvard University Press.

McLaughlin, B. (1992). The rise and fall of British emergentism. In A. Beckermann, J. Kim & H. Flohr (Eds.), *Emergence or reduction? Essays on the prospects of nonreductive physicalism* (pp. 49-93). Berlin, New York: Walther de Gruyter.

McLaughlin, B. (2001). In defence of new wave materialism: A response to Horgan and Tienson. In C. Gillett & B. Loewer (Eds.), *Physicalism and its Discontents* (pp. 319-330). Cambridge: Cambridge University Press.

McLaughlin, B. (forthcoming). Does mental causation require psychophysical identities? Forthcoming in a volume of essays in honor of Jaegwon Kim.

McLaughlin, P. (1993). Descartes on mind-body interaction and the conservation of motion. *Philosophical Review*, *102*, 155-182.

Melnyk, A. (1997). How to keep the 'physical' in physicalism. *Journal of Philosophy*, *94*, 622-637.

Melnyk, A. (2003). Some evidence for physicalism. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 155-172). Exeter: Imprint Academic.

Menzies, P. (2003). The causal efficacy of mental states. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 195-223). Exeter: Imprint Academic.

Montero, B. (2003). Varieties of causal closure. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 173-187). Exeter: Imprint Academic.

Mundale, J. (2001). Neuroanatomical foundations of cognition: Connecting the neuronal level with the study of higher brain areas. In W. Bechtel, P. Mandik, J. Mundale & R.S. Stufflebeam (Eds.), *Philosophy and the neurosciences: A reader* (pp. 37-54). Malden, Mass: Blackwell Publishers.

Newton, I. (1962). *Unpublished scientific papers*. Cambridge: Cambridge University Press.

Noordhof, P. (1999). The overdetermination argument versus the cause-and-essence principle – no contest. *Mind*, *108*, 367-375.

Nye, A. (1999). *The princess and the philosopher: Letters of Elisabeth of the Palatine to René Descartes*. Lanham, Md.: Rowman & Littlefield Publishers.

Papineau, D. (1990). Why supervenience? *Analysis*, *50*, 66-70.

Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents* (pp. 3-36). Cambridge: Cambridge University Press.

Papineau, D. (2002). *Thinking about consciousness*. Oxford: Clarendon Press.

Piccolino, M. (1998). Animal electricity and the birth of electrophysiology: The legacy of Luigi Galvani. *Brain Research Bulletin*, *46*, 381-407.

Psillos, S. (2004). A glimpse of the secret connexion: Harmonizing mechanisms with counterfactuals. *Perspectives on Science*, *12*, 288-319.

Oppenheim, P. & Putnam, H. (1958). The unity of science as a working hypothesis. *Minnesota Studies in the Philosophy of Science*, 2, 3-36.

Quiroga, R.Q, Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102-1107.

Rains, G.D. (2002). *Principles of human neuropsychology*. Boston: McGraw-Hill.

Ramberg, B. (1999). The significance of charity. In L.E. Hahn (Ed.), *The philosophy of Donald Davidson* (pp. 601-618). Chicago: Open Court Publishers.

Rapport, R. (2005). *Nerve endings, the discovery of the synapse*. New York, London: W.W. Norton & Company.

Rey, G. (2001). Physicalism and psychology: A plea for substantive philosophy of mind. In C. Gillett & B. Loewer (Eds.), *Physicalism and its Discontents* (pp. 99-128). Cambridge: Cambridge University Press.

Reynholds, J.H. & Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron*, *24*, 19-29.

Robertson, L.C. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, *4*, 93-102.

Roskies, A.L. (1999). The binding problem. *Neuron*, *24*, 7-9.

Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society*, *13*, 1-26.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton: Princeton University Press.

Salmon, W. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*, 297-312.

Schaffner, K.F. (1993). *Discovery and explanation in biology and medicine*. Chicago, London: University of Chicago Press.

Schaffner, K.F. (2006). Reduction: the Cheshire cat problem and a return to roots. *Synthese*, *151*, 377-402.

Schuetze, S.M. (1983). The discovery of the action potential. *Trends in neuroscience*, *5*, 164-168.

Sider, T. (2003). What's so bad about overdetermination? *Philosophy and Phenomenological Research*, *67*, 719-726.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, *24*, 49-65.

Smart, J.C.C. (1978). The content of physicalism. *Philosophical Quarterly*, *28*, 339-341.

Sperry, R.W. (1986). Discussion: Macro- versus micro-determinism. *Philosophy of Science*, *53*, 265-270.

Stoljar, D. (2006). Physicalism. *The Stanford Encyclopedia of Philosophy*, *Winter 2005 Edition* (E.N. Zalta, Ed.). Retrieved August 29, 2006, from http://plato.stanford.edu/archives/win2005/entries/physicalism/

Sturgeon, S. (1998). Physicalism and overdetermination. *Mind*, *107*, 411-432.

Sturgeon, S. (1999). Conceptual gaps and odd possibilities. *Mind*, *108*, 377-380.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*, 49-100.

Tooley, M. (1990). Causation: Reductionism versus realism. *Philosophy and Phenomenological Research*, *50 supplement*, 215-236.

Treisman, A. (1996). The Binding Problem. *Current Opinion in Neurobiology*, *6*, 171-178.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, *24*, 105–110.

Valenstein, E.S. (2005). *The war of the soups and the sparks. The discovery of neurotransmitters and the dispute over how nerves communicate*. New York: Columbia University Press.

Van Gulick, R. (1992). Nonreductive materialism and the nature of intertheoretic constraint. in: Beckermann, A., Kim, J. and Flohr, H. (Eds), *Emergence or Reduction?*

Essays on the Prospects of Nonreductive Physicalism, Walther de Gruyter, Berlin, New York, pp. 157-179

Van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies*, *27*, 185-199.

Von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, 24, 95–104.

Wallis, J.D. (2006). Evaluating apples and oranges. *Nature Neuroscience*, *9*, 596-598.

Williams, G. & Goldman-Rakic, P. (1995). Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature*, *376*, 572-575.

Witmer, D.G. (2001). Sufficiency claims and physicalism: A formulation. In C. Gillett & B. Loewer (Eds.), *Physicalism and its Discontents* (pp. 57-73). Cambridge: Cambridge University Press.

Wolff, P. (2002). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*, 1-48.

Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal for the Philosophy of Science*, *51*, 197-254.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science Supplemental*, *69*, S366-S377.