TUOMAS K. PERNU

Interactions and Exclusion: Studies on Causal Explanation in Naturalistic Philosophy of Mind

DEPARTMENT OF BIOSCIENCES
FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES
AND
DEPARTMENT OF PHILOSOPHY, HISTORY, CULTURE AND ART STUDIES
FACULTY OF ARTS
UNIVERSITY OF HELSINKI
Interactions and Exclusions: Studies on Causal Explanation in Naturalistic Philosophy of Mind

Tuomas K. Pernu

Department of Biosciences
Faculty of Biological and Environmental Sciences
&
Department of Philosophy, History, Culture and Art Studies
Faculty of Arts

ACADEMIC DISSERTATION

To be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki, in lecture room 5 of the University of Helsinki Main Building, on the 30th of November, 2013, at 12 o'clock noon.
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.

Jerry Fodor
“Making mind matter more”
(1989)
Supervised by
Docent Markus Lammenranta
Department of Philosophy, History, Culture and Art Studies
University of Helsinki

Professor Kristian Donner
Department of Biosciences
University of Helsinki

Reviewed by
Professor Jussi Haukioja
Department of Philosophy and Religious Studies
Norwegian University of Science and Technology

Professor David Papineau
Department of Philosophy
King’s College London

Publicly examined by
Professor David Papineau
Department of Philosophy
King’s College London
ABSTRACT

The notion of causal explanation is an essential element of the naturalistic world view. This view is typically interpreted to claim that we are only licensed to postulate entities that make a “causal difference”, or have “causal power”. The rest are epiphenomena and hence eliminable from the correct view of reality. The worry that some entities and phenomena that we take for granted – mental properties in particular – turn out to be epiphenomenal, can be seen as stemming from this sort of naturalistic attitude.

This thesis reviews the issue of causal explanation within the context of the naturalistic philosophy of mind. It is argued that there is no single monolithic, unanimously accepted notion of causation that the naturalist should be committed to. Views vary on what this notion amounts to exactly, and fields of science vary with respect to their causal commitments. However, the naturalist can still presume that a scientifically informed philosophical account of causation exists, an account that is fundamentally philosophical, but also sensitive to actual scientific practice and its view of reality.

The central issue of the current naturalistic philosophy of mind is the so-called problem of causal exclusion. According to this, the assumption that mental states could have genuine and autonomous effects on the physical world is inconsistent with physical commitments, namely the idea that mental states are necessarily neurally based and the idea that the physical world is causally complete. The causal exclusion argument claims that mental causes must be reduced to physical causes, as there remains no role for independent mental causes.

The thesis reviews some central responses to the causal exclusion argument. It is shown that within the context of the interventionist notion of causation, inter-level causation can be ruled out. The causal exclusion argument would thus find support, contrary to what the proponents of the interventionist view typically claim. However, the result is also shown to have the corollary that purely higher-level, mental-to-mental causation is possible. The thesis suggests that this offers a consistent view of mental causation for a naturalist to hold.
CONTENTS

7  List of original articles

9  Preface

PART I  INTRODUCTORY ESSAY

15  1. Introduction

17  2. Naturalism

25  3. What is the problem of mental causation?

29  4. Causal exclusion

30    4.1 Distinctness

32    4.2 The causal completeness of the physical

35    4.3 Overdetermination and exclusion

38  5. Responses to causal exclusion

38    5.1 Mental causation and physical causation

41    5.2 The counterfactual response

43    5.3 Interventionism and difference-making

47  6. The autonomy solution and interventionist parallelism

50  7. Conclusions

51  References

PART II  ORIGINAL ARTICLES

57  Is knowledge a natural kind?

73  Minding matter: how not to argue for the causal efficacy of the mental

99  Causal exclusion and multiple realizations

105  Interventions on causal exclusion

115  Does the interventionist notion of causation deliver us from the fear of epiphenomenalism?
LIST OF ORIGINAL ARTICLES

ARTICLE I

“Is knowledge a natural kind?”.  
*Philosophical Studies* 142, p. 371–386.

ARTICLE II

“Minding matter: how not to argue for the causal efficacy of the mental”.  

ARTICLE III

“Causal exclusion and multiple realizations”.  
*Topoi*, forthcoming.

ARTICLE IV

“Interventions on causal exclusion”.  
*Philosophical Explorations*, forthcoming.

ARTICLE V

“Does the interventionist notion of causation deliver us from the fear of epiphenomenalism?”.  
Completing this thesis has been an unexpectedly long and winding journey. Many things have happened and I am in a totally different place now than where I was when I started, both academically and personally.

I started this work with a metaphilosophical agenda. In my master’s thesis I had studied the foundations of empiricism, namely the question of how pure sensory data (supposing that there is such thing) could mesh with the conceptual elements of cognition to produce reliable propositional information. This project already had a distinct naturalistic attitude and much of the discussion drew from the psychology of perception and linguistics – and all the of naturalistic philosophy of mind that is sensitive to these matters. However, upon finishing the project it became apparent that the elements of the thesis did not go together. Although I could not immediately put my finger on it, there was definitely something fundamentally wrong. If I was ever going to do philosophy again, I had to investigate the source of this uneasiness. This is what I originally set out to do in this thesis.

The period after I had finished my master’s thesis was marked by personal turmoil and soul-searching. Like many others, I found the chances of doing fruitful philosophical analysis under the strain of empirical science very bleak. Such feelings can become very concrete for a young philosopher who has no solid conception of philosophy, science or life in general. In different circumstances, with slightly different turns of events and decisions – if, for example, I would have not become frustrated with the management of the bakery where I was working at the time – things might have turned out entirely differently. It was by no means clear that I would pursue an academic career.

It is one of the first lessons in metaphysics of causation that each turn of an event is preceded by an astronomical number of causal antecedents that are necessary for bringing about that event in exactly the way it happened to occur. There is, however, a very distinct moment in history that can be pointed to as an event that initiated the process that led into the thesis as it appears here. I wrote my master’s thesis under the supervision of docent Markus Lammenranta and Professor Ilkka Niiniluoto, and upon finishing the project, they both encouraged me to continue and start to work on a PhD. This support alone was not enough to convince me that it was prudent to commit to such an undertaking, but after a few months I was notified that the department of philosophy was going to grant small scholarships to a few promising postgraduate students, and that I would be one of the recipients, if I just submitted an application. I remember professor Gabriel Sandu being particularly encouraging and active on this matter.
I am grateful to all three of these people, and to the department of philosophy, for taking the right actions at the right time. Receiving that small start-up grant can definitely be singled out as an event that set this project in motion. Markus also took on the role of a supervisor of the thesis. Years have passed, but his encouragement has never waned and his support has always been steadfast.

The first article in the thesis derives from this time. The first version was written ten years ago and it appears here largely in the form that it was originally conceived. The article is an attempt to couch in rational terms some of the uneasiness that the work on the master’s thesis had prompted. Its main insight is that the possibility of eliminative naturalism has been largely overlooked. The focus of this article was epistemology, but the initial idea was to go further and apply this approach to issues in philosophy of mind and philosophy of language. This would require taking seriously the sciences that are essential to naturalistic philosophy, biology in particular.

I first came to Viikki, to the faculty of biological and environmental sciences, to study ecology and evolutionary biology under the supervision of docent Kari Vepsäläinen. His seminar series on current questions in evolutionary biology were legendary and I attended one with little previous knowledge of biology. Rather than focus on my obvious inabilities, Kari welcomed me as philosophical advisor, and I soon found myself assisting him in organizing new seminars. I am grateful for his open attitude and all of the help I received from him. I would not be where I am had I not had his support.

At this time I was searching for a place to complete my civilian service. Kari introduced me to professor Kristian Donner, who was heading the division of animal physiology. There was a civilian service position at the division and it was soon decided that I would fulfil my duty there. This gave me the opportunity to deepen my knowledge of biology and teach philosophy to biologists. During the year I was fortunate to encounter many motivated and bright students, some of whom are still my friends.

This time was so fruitful for all parties involved that I subsequently stayed on in Viikki and resumed my work on this thesis. Kristian also assumed the role of a supervisor of my thesis. It has been a great pleasure and honour to work as his student, colleague and friend. Our regular meetings have always been sincere and sharp, with discussion topics ranging from culture and politics to philosophy, science and life in general. Kristian’s help in practical matters has been simply invaluable. If only all of us were that well equipped both theoretically and practically.

During my stay in Viikki I have had the privilege to enjoy the stimulating company of a number of students and researchers. I have learned much from Vesa Selonen, Tiina Huttu, Tero Viitanen, Juuso Juuri, Tomi Maila, Kai Kaila, Roosa Laitinen, Isaac Salazar-Ciudad, Heikki Helanterä, Arto Annila, Petri Ala-Laurila, Neil Bell and Gunther Jansen. I remember many nightly discussions on science and philosophy with Roosa in parties and in the hallways and corridors of the Biocentres. Although we may have had
a somewhat bumpy ride, I still remember those days with warmth and respect. Arto has infected me with the need to see everything in terms of the 2nd law. I have yet to decide whether this is a good or a bad thing. Heikki has been my collaborator in theoretical and philosophical evolutionary biology and I’m looking forward to working on our future projects (and finalizing some that we have already started). Neil has revised the English of several of my manuscripts, including a draft of this thesis. I’m deeply grateful for all the work he has done. I think I owe him at least one more crazy dinner party.

Gunther came to study systematics under the supervision of Kari Vepsäläinen and Riitta Savolainen, and we were quickly introduced to each other. His wide interest in biology, science, philosophy and culture ensured that we always had things to discuss and projects to plan. Together with Gunther we founded the Viikki Study Group on Philosophy of Science, and I want to thank Gunther, Kristian, Heikki, Neil, Wouter Vahl, Jonathan Jeschke and all the other participants for enjoyable meetings and stimulating discussions. I think I have always learned something new.

After resuming my work on this thesis two things happened. First, it soon became apparent that my original idea for a thesis rested on assumptions that were in a need of thorough reanalysis. Although the issue of causality and causal explanation lies at the heart of naturalism and naturalistic philosophy of mind, much of the discussion relies on a rather vague conception of these. I understood that in order to address the issues that I was originally interested in, this is where my focus should be. Second, I became more and more involved with the sport of cycling.

For two years I served as president of Prologi, the Helsinki University Cycling Club, and a further year as vice-president. In 2007 I was in charge of establishing the Tour de Helsinki, an open cycle race taking place in Helsinki and its vicinity. In 2008 I founded Velocitor Ltd, which has since been the organizing body of the race. And in 2010 I founded Team Velocitor, an elite class road racing team sponsored by Velocitor Ltd. Although looking back now all of these ventures have been rewarding in many ways, they have definitely taken a toll on my work on this thesis. At times I have felt that cycling has meant nothing but teeth grinding and sleepless nights. I could certainly have had a more relaxing hobby than running a sporting event promotion business. Nevertheless, a mark has been left on the Finnish sporting scene, and the experience has definitely been edifying. I want to thank Olli Heikurinen and Tuomas Turunen who have stood by me in these endeavours.

From 2010 to 2011 I served two years as the Chairman of the Committee for Educational Affairs of the Student Union of the University of Helsinki (HYY). This allowed me to see the administrative and political side of the university and its educational processes. I want to thank all of the numerous friends and colleagues I got to know and work with during these years and with whom I had the pleasure of enjoying countless nocturnal adventures in various parties, student houses, bars and saunas.
All work, even writing a philosophy thesis, occurs within a larger social context. I want to thank all my old friends who never gave up teasing me about the status of this work. Special thanks are due to Aleksi Salokannel and Sisin for helping me to prepare the final layout of the thesis. I am grateful beyond words to Ida Vartiala for her love and support. She has given me the strength to carry on through to the final stages of this project. Finally, I want to thank my parents, Kalevi and Ulla Pernu, for their encouragement and relentless support.

The articles collected in this thesis span several years and several different stages of my life. None of this would have been possible without the generous support I have received from various institutions. Apart from the start-up grant I received from the department of philosophy, I am particularly grateful to the Emil Aaltonen Foundation for a three year grant that I was awarded during the early stages of this work. The work for this study has been made financially possible by (in alphabetical order):

- The Emil Aaltonen Foundation
- The Department of Philosophy at the University of Helsinki
- The Ella and Georg Ehrnrooth Foundation
- The Finnish Cultural Foundation
- The Waldemar von Frenckell Foundation
- The Institute of Biotechnology at the University of Helsinki
- The Alfred Kordelin Foundation
- The Otto A. Malm Foundation
- The University of Helsinki Funds
- The University of Helsinki
- The Oskar Öflund Foundation

I hereby express my deepest and most sincere gratitude to all these institutions that deemed my work worthy of supporting.

Tuomas Pernu
Helsinki, November 2013
Part I

Introductory Essay
1. INTRODUCTION

Take two natural and *prima facie* metaphysically innocent assumptions. Suppose first that there are genuine causal relationships between properties. Suppose next that these properties are divided into distinct levels: they are organized hierarchically in such a way that there are higher-level properties that are composed out of lower-level properties. Now burning questions start to arise. What is the relationship between higher and lower-level properties? Can higher-level properties figure in autonomous causal relationships even if higher-level properties are composed of lower-level properties? Or are the relationships epiphenomenal or derivative of lower-level, ultimately real, causal relationships?

These questions are pervasive. They are central to metaphysics and the philosophy of science, and resonate through practically all fields of science and philosophy. Most notably the issue is present in debates over the autonomy of the special sciences – the question of whether chemistry reduces to physics, biology to chemistry, psychology to biology, the social sciences to psychology and so on. In each case one typically ends up asking whether the phenomena studied at a particular level of science can be fully accounted for solely by reference to the dynamics of the underlying science. Within the particular disciplines themselves, the issue often takes the form of asking whether a certain collective or group that is composed of individual entities can enjoy autonomous existence and engage in interactions independently of, or over and above, these individuals and their interactions. To take an example from evolutionary biology, one such prominent issue is to what degree different levels, e.g. genes, individuals and groups of individuals, need to be taken into account in explaining evolutionary changes. Often such issues mark the boundaries between disciplines or sub-disciplines.

Even formal philosophical disciplines such as logic and philosophy of language are not left untouched. They become relevant as soon as one tries to get a more rigorous
grip on the issue. But more importantly, through a metaphilosophical twist, all normative philosophical disciplines, including ethics, are affected. For as soon as one adopts the naturalistic stance favoured by many philosophers today, one is left wondering how such normative disciplines could fit within the causal scheme of things. Epistemology, in its naturalized form, is also immediately affected, for it becomes pertinent to ask how epistemology and epistemic attitudes could have causal relevance.

Nowhere, however, are these questions more tangible than in the field of philosophy of mind. On the one hand we have a clear conception of our bodies as physical entities. Our bodies are not dead, inanimate objects, of course, but living biological organisms. But they are nevertheless physical in the sense that the various processes and actions that they undergo can be accounted for in wholly physicochemical terms. This, in essence, is what the last one hundred years of physiology and neuroscience has taught us. There are no vital powers or mysterious forces that need to be evoked to explain the functioning of our bodies. Whether this amounts to claiming that physiology may be reduced to biochemistry and biophysics is already one formulation of the issue at hand. The issue becomes more concrete, however, in a form of another question that is bound to spring up immediately. For, on the other hand, we have an equally clear conception of our minds as mental, i.e. non-physical, entities. Obviously, this apparent dualism cries for elaboration, but let us assume for now that there is a naturalistically acceptable account of an “emergence”, “constitution” or “realization” relation that explains how such an entity can be related to a physical basis. As already noted, the view of a hierarchical ordering of nature is at least intuitively appealing. But relying on such a view seems to force problems on us rather than solve them. For now we are faced with the very issue of mental causation or mind-body interactionism: the question of how such metaphysically distinct entities as bodies and minds can causally influence each other. The basic conflict here is not difficult to grasp. On the one hand nothing seems more natural to us than the idea that with our thoughts, desires and conscious decisions we can control our actions and have an impact on the physical world around us. On the other hand our behaviour and its influence on the world can be fully explained in physical terms, in the vein sketched above. In consequence, our mental lives seem to be left with a rather idle role. But such a conclusion seems to contradict not only the very image of ourselves as active, personal and morally responsible agents, but also our common sense and our everyday social and legal practices. Naturalizing the mind does not come cheap.

Although the basic problem of mental causation is fairly easy to grasp, when looked at more closely one is soon faced with diverse philosophical issues that intertwine. One is led to ponder the scientific status of propositional attitudes and intentionality, the qualitative and subjective nature of conscious experiences, and the ancient and notorious problem of free will. All of these issues are entangled with deep and equally notorious questions about the metaphysical nature of causation, varieties of scientific expla-
nation and the scope and limits of reductionism. Philosophers are quite careful to keep these issues apart, and often tend to concentrate on only one particular facet of this manifold problem. But outside the domain of philosophy proper, this systematic attitude soon evaporates and all these diverse issues merge into one big and mysterious puzzle that seems to elude all rational scrutiny and discussion.

Since the issues at hand are so diverse and the analysis prone to confusion, it is especially important to be as clear as possible about the aim and scope of this study: it treats the problem of mental causation as a general problem for the philosophy of science at an abstract metaphysical level. The emphasis is on the causality side of the problem, rather than on the mentality side. That is, the main topic is the metaphysical nature of causation and the nature of causal explanation in science with respect to mental states and psychological explanation. The exact nature of the mental – as distinct from the physical – is not the prime focus here. It is simply taken for granted that mental states or psychological explanation constitute a prima facie genuine and independent domain. What is of concern is the causal status of this domain with respect to the physical. The core question can be easily posed: supposing that the physical domain is complete, with each physical effect having sufficient physical causes, and supposing that the mental domain is something distinct from the physical domain, what causal role, if any, is left for the mental to play?

2. NATURALISM

Let us start by elaborating the metaphilosophical background of this study. It falls within the scope of naturalistic philosophy or naturalism. Naturalism, in turn, can have a diversity of meanings depending on the kind of philosophy or historical period one is focusing on. There is the metaphysical naturalism of Aristotle and its scholastic descendants, there is the tradition of natural philosophy starting from the inception of modern science – perhaps most notably represented by Newton’s *Philosophiae Naturalis Principia Mathematica* (1687) – and there is the pragmatic tradition of naturalism, stemming from Dewey (1903), for example. However, the type of naturalism that this study is concerned with is the modern scientific philosophical naturalism that has arisen from – or was born as a reaction to – the logical positivism and analytic philosophy of the early twentieth century. This is the naturalistic tradition outlined by Quine (1951, 1969).

Within this tradition there are two different strains or facets that need to be distinguished. The most fundamental idea, and one that can be taken as the connecting theme in the different historical variations of naturalism, is the rejection of the supernatural as an explanation of the world. This includes not just the physical realm, but the entire animate world as well, including human action and psyche. According to this *ontological* or *metaphysical naturalism*, the world is one all-encompassing natural system with no su-
pernatural ingredients. As such, this characterization is clearly not very informative. It is hardly very revealing to say that naturalism is not supernaturalism. To make this stance more substantial, it needs to be supplemented with some criterion for distinguishing the “natural”, and thus “real”, from the “supernatural”, or “unreal”. The criterion that the naturalist is bound to resort to is the conviction that only entities and phenomena that are acknowledged by science, or at least in principle possible to investigate by empirical means, are considered real. That is, the scientific, empirical method is our only source of information and means of postulating real entities, phenomena and processes. There are no privileged intuitive, contemplative or philosophical sources of information. This is the methodological component of modern naturalism.

The methodological side of naturalism quickly extends towards a more radical and controversial form of naturalism. If empirical science is given the exclusive right to serve us information and shape our view of reality, this will leave philosophy, and epistemology in particular, in a rather straitened situation. According to the traditional and intuitively appealing view, the scientific study of reality is preceded by philosophical contemplation and analysis – does not our confidence in the empirical method itself rely on a particular philosophical view of epistemology, namely empiricism? But now this naturalized epistemology seems to require us to replace philosophical analysis with empirical, scientific research. Or, if this is too radical, at least traditional epistemology is required to yield up its privileged position and admit that it needs to be empirically informed and corrigible. It is easy to appreciate why many may find this philosophically uncomfortable.

This study is not primarily a treatise on the aim and scope of naturalism or the possibility of philosophy within the naturalistic scheme. Naturalism, in the vein characterized above, is taken for granted. This thesis is a study within this naturalistic tradition. However, it is instructive to point out some problems that this naturalistic philosophy is bound to face. Many of these critical issues will resurface within the naturalistic research program itself. After all, the notion of causal explanation in general, and the problem of mental causation in particular, is ridden with concerns about what the naturalistic attitude is committed to hold as ultimately real.

The most obvious, and most discussed, problem faced by naturalistic philosophy is the question of how normative elements, which form the core of most philosophical disciplines, fit within this empirical and purely descriptive research scheme. As already hinted, epistemology (as traditionally conceived) is not the study of actual belief formation processes. What epistemology is concerned with is the question of how we should form our beliefs. Although the former issue can be seen as interesting and relevant in many ways – for one thing, it might be nice to know whether our actual belief formation processes match the ones we aspire to – the latter question is the fundamental one, and the one that is characteristically philosophical. But epistemology is no exception. Logic
and ethics are also philosophical disciplines that are not primarily concerned with the actual world and our place and behaviour in it, but with describing correct and better ways of reasoning, thinking or behaving. Whether, or how, these disciplines are possible within naturalistic philosophy is an important and widely discussed question in current philosophy. However, this is not an issue that this thesis is primarily concerned with.

Although the problematic situation of normative philosophy is something that must be acknowledged and the extensive discussion surrounding the issue is well deserved, there is another cluster of problematic issues that seems to be largely overlooked. They are also more directly relevant to this study. These are the issues that result from the internal tensions that naturalistic philosophy is bound to be infected with. Although applying the naturalistic attitude may sound attractively simple and straightforward, and something that would help us to solve or bypass some of the traditionally insurmountable philosophical problems, on closer inspection this optimism may turn out to be ill-founded. The problem is that there is no unequivocal and universally accepted conception of what exactly the scientifically informed view of reality amounts to.

The basic observation, which often seems to get overlooked, is that the sort of naturalism discussed here is a type of conventionalism. Science and scientific research comprise highly diverse and variable practices that are in a state of constant change. What science deems to be real now, it can denounce as unreal in the future, and vice versa. Such fallibilism is one of the central characteristics of scientific research. This in itself does not need to be a vice. After all, naturalists are the first to note that renouncing traditional foundationalist philosophical tendencies is one of the central advantages of naturalistic philosophy. However, at the same time, this variability cripples the aspiration that naturalism could offer clear-cut answers to foundational metaphysical and epistemological issues. Different scientific disciplines, at different times, have different methods and postulate different entities and processes. Often this variability is difficult to fit together. There is no single unified scientific stance on reality. Disputes on what kind of entities to postulate, and what are the best and most correct ways of carrying out scientific research, rage constantly, not only between scientific disciplines, but also within them. To settle such disputes, philosophical analysis seems to be called for. But naturalistic philosophy seems to be ill-equipped to offer any help. To reach a unified view on reality you would need to pick out a specific science, with a specific method, at a specific time. But such a choice would seem completely arbitrary from the naturalistic perspective, for there is supposed to be no prescientific philosophical justification for such choices.

This problem is not merely abstractly philosophical, but very concrete. Suppose we settle on the following general characterization of science: scientific research is objective quantitative analysis of empirical data. Empiricism is definitely something that we should deem essential to the scientific method. But how about logic and mathematics? Surely they are not empirical sciences? But it is difficult to imagine any scientific re-
search managing without them. It seems that the naturalist will have to accept at least this many extra-empirical ingredients in her worldview. Objectivity is again something that seems to characterize the scientific method. It is an essential feature of any scientific research that the data and methods are able to be publicly scrutinized and that the results are at least in principle repeatable. But how about the subjective feelings and experiences that are so characteristic of our mental lives? They are not only central to our everyday life, and as such should earn our full scientific attention, but also constitute a concrete research subject for psychology and medicine. Should the naturalist instruct us to refrain from such research? Such advice would seem completely arbitrary and would go against the naturalist’s own commitment to take science at face value. Consider finally the quantitative method. It is certainly characteristic of science that the data and results of research are quantitatively representable and communicable. But much of the research in psychology, humanities, the social sciences and medicine is carried in qualitative terms. Again, forcing quantitative research in line with the rest of the sciences does not seem like an attractive option.

One problematic issue is particularly relevant from the perspective of this study. This is the question how, exactly, naturalism and physicalism relate to each other. Physicalism is a metaphysical thesis according to which all that exists is physical, or determined by the underlying physical reality. Does naturalism entail physicalism? And more importantly, if it does, what sort of physicalism, the reductive or non-reductive kind? The first thing to note here is the obvious close connection of physicalism to physics. So one could reformulate the question thus: does naturalism entail that physics has a somehow privileged or fundamental status among the sciences? There certainly seems to be a strong tendency to answer this in the positive. However, that is clearly a particular interpretation of what naturalism amounts to, not a necessary connection. Many find a more relaxed interpretation more useful.

But even if physics is regarded as a benchmark science, this does not necessarily help naturalism to resolve the issues that it aspires to resolve. First of all, as is widely acknowledged, what physics is and what kind of entities it postulates varies in time. Hence, as before, it is not at all clear what kind view of reality the naturalist is committed to accept, even if naturalism is taken to entail physicalism. The problem the naturalist now faces can be neatly formulated as a dilemma, presumably first posed by Hempel (1969): either the naturalist will have to rely on a view offered by the current physics, which is in all probability false, for as we know from history, physics will change and the current view will be discarded, or alternatively on a view of an ideal future physics, which amounts to a rather empty view, for no-one knows what such a physics will look like. If the views that the naturalist has on offer are either false or purely speculative, naturalism does not seem to amount to a thesis that is any better off than the traditional philosophical doctrines it seeks to replace. Second, even if the naturalist opts for the first horn of the di-
lemma and claims, for example, that the view we should be interested in is one based on current physics and that it does not matter for the current discussion that physics and the view resulting from it might change in the future, this might not improve the situation significantly. The problem is that there is no single currently unanimously accepted view on physical reality, even within physics itself. Physics is a large and variable field of ongoing research. Physicists disagree on what the fundamental nature of reality is, and, more notably, on whether all of the entities and phenomena that we acknowledge as real are ultimately derivable from the basic entities and processes that this fundamental physical research postulates. That is, it is only a physicist with particular philosophical inclinations that might think that physics holds a special and fundamental place among the sciences. To many, this issue is simply not relevant. Some think that physics itself has actually shown how many of the entities and phenomena mundanely acknowledged as real and studied by other empirical sciences simply cannot be derived from the data and theories of fundamental physics.

What this suggests is that even if some sort of physicalism is entailed by, or at least is closely connected to naturalism, it is a genuinely open question whether this physicalism is of the reductive or non-reductive kind. In fact, it is fair to say that although the majority of today’s philosophers would admit to being naturalists of some sort, and even physicalists, only a minority would admit to being reductive physicalists. How exactly reductionism should then be understood, is another problematic and widely discussed issue. However, there is a particular view that receives most of the attention in the current discussion. This is the view that holds that although the basic foundation of reality is thoroughly physical, the higher levels of reality supervene on this physical foundation. Supervenience, in turn, is taken to be a non-reductive, asymmetric dependence relation. That is, higher levels of reality are dependent on, and determined by, the subvening physical basis, but they are not reducible to it. This is the core idea of the current non-reductive physicalism. It is also the view that is most relevant to the present study. The issue of mental causation in naturalistic philosophy largely concerns the question of whether this type of non-reductive physicalism is a stable view.

Although the preceding discussion connected the metaphysical facet of naturalism to the methodological one via the latter’s ability to supply the former with a criterion for separating the real and the unreal, there is also another route open for the metaphysical naturalist to substantiate her thesis. Rather than directly resort to the authority of the sciences, the naturalist can stress that there is a particular form of explanation that is characteristically scientific and hence can act as our guide to what is ultimately real. This is the causal form of explanation. What this view holds, basically, is that we are licensed to postulate only entities that make a “causal difference” or have “causal power”. The rest is epiphenomena and eschewable by science, and hence eliminable from the correct view of reality. Although perhaps not that often explicitly acknowledged as such,
this is certainly an intuitively attractive and consequently widely held view. The concern that some entities and phenomena that we take for granted—mental properties in particular—may turn out to be epiphenomenal can be seen as stemming from this sort of naturalistic attitude.

However, this view is plagued by familiar problems. For example, why should causal explanation enjoy such a privileged status? Are there not other ways of making sense of the world that are also scientifically legitimate? Once again we turn to logic, mathematics and other theoretical or abstract ways of explaining the world. Surely there are perfectly cogent, purely mathematical explanations of some features of reality? And being purely mathematical, surely such explanations are not causal? If that is the case, then the naturalist must admit that appealing to causal explanation as a criterion for demarcating the real from the unreal cannot be the whole story.

But there are even more deeply problematic issues related to this view. Even if we would be prepared to grant causal explanation this privileged status, it is not at all clear what exactly this “causal difference-making” or “causal power” would amount to. Although causal explanation has a seductively naturalistic ring to it, and undeniably many scientists would grant that causal explanation is something that they are essentially pursuing in their research, it is also widely acknowledged that causality is something that notoriously eludes the sciences. Again, there are many different problems here.

First and foremost, the question of what causation is exactly is one of the most vehemently discussed metaphysical issues in the history of philosophy. The debate is no less heated today. There are many different accounts on the table, each with their pros and cons, with different fields of philosophy and science favouring different notions. In consequence, there is no single and unanimously accepted account of causation that the naturalist can simply put into use.

Second, assuming that naturalism is closely connected to physicalism, it becomes immediately relevant to enquire about the exact nature of physical causation. But there are in fact well-founded reasons to suspect that there is no such thing as physical causation. As Russell (1913) famously pointed out, if we look at the “advanced sciences such as gravitational astronomy”, the word “cause” never occurs there, which, along with other reasons, should lead us to conclude that the notion is simply “a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm” (p. 1). Now, perhaps “gravitational astronomy” is not the best example of an “advanced science” of today, but the argument is still compelling: if we look at dynamical physical theories, what we tend to find is not causal relationships, but rather functional relationships between variables and data-points. Changes in physical systems are not due to different factors being in causal interaction so much as to the system as a whole unfolding. What this should lead the physicalist to do (as it led Russell) is to conclude that the notion of causation should be abandoned.
Third, even if we would not be so confined to fundamental physics, it is not at all clear that there is some well-defined and universally accepted notion of causation to be found in the special sciences. Here the situation is reversed: if the word “cause” never appears in physical theories, the opposite can be said to hold in all other sciences – the problem is not that the word does not appear, but rather that the sciences are at a loss as to what the word means. Biology, medicine and psychology, for example, can be described as being in the business of deciphering causal relationships. However, at the same time, the often recurring theme in discussions on data and research results is the question how these should be understood causally. What we see are complex patterns of statistical connections and correlations that do not have any straightforward causal interpretation. Such an interpretation awaits a philosophically informed account of what causation fundamentally is. The naturalist cannot just go and have a look at scientific practice and pick out the proper notion of causal explanation from there, even if the word itself is abundantly used, because there simply is no solid notion behind the word.

With all this criticism marched against naturalism, a rational person might start to wonder whether it would be a better to stay away from naturalistic philosophy. The problem is that there are no real options. Empirical science, with natural sciences and physics at the forefront, is, and has always been, closely connected to philosophy. The two will have to get along, whether we like it or not. Which does not mean, of course, that the details of their coexistence should not be exposed to critical analysis and thorough discussion. That is what today’s philosophy is largely engaged with.

As already stressed, this thesis is not primarily concerned with whether or how the naturalistic project can be justified. However, after raising all these problems with it, it is appropriate to end with a few positive remarks. First, one could note that the preceding discussion was perhaps overly simplistic and dichotomous. There are obvious tensions between naturalistic aspirations and scientific practice. But this does not mean that the two could not influence each other and have a fruitful coexistence. What we might have to conclude, however, is that we need to abandon – at least for the time being – the more radical naturalistic idea that philosophy should be replaced with empirical science. We might need to settle for the more conventional view that stresses only the continuity between the two. What this view must also acknowledge – and this is often overlooked – is that it is not only philosophy that needs to be informed by the sciences. The empirical sciences must also be sensitive to philosophical analysis, whenever the need for such appears. If the naturalist is serious about insisting that the two must be on a par with each other, then she must acknowledge that the continuum runs both ways.

Causal explanation is a topic that exemplifies well this more relaxed and interactive view of naturalism. Causal explanation is something that is clearly a characteristic of science, as already stressed. But it is also something that is in need of philosophical clarification. This does not have to be a stalemate. What we can, and should, go after is a
scientifically informed philosophical account of causation, an account that is fundamentally philosophical, but also sensitive to actual scientific practice and views of reality. There can be no a priori guidelines for formulating such an account. It can only arise out of a sincere dialogue between philosophy and science.

It is worth pointing out that in fact Russell (1948) was an excellent representative of this attitude towards causal explanation. Although he is known for his criticism of the notion of cause, he also admitted that there is a “primitive notion of cause” that is both fundamentally philosophical and scientifically relevant. He writes:

The concept “cause”, as it occurs in the works of most philosophers, is one which is apparently not used in any advanced science. But the concepts that are used have been developed from the primitive concept (which is that prevalent among philosophers), and the primitive concept [...] still has importance as the source of approximate generalisations and pre-scientific inductions, and as a concept which is valid when suitably limited. (1948, p. 471.)

Russell goes on to define the notion of a “causal line”, which in essence is a series of events that manifests “quasi-permanence” of some quantity or quality. Neither the details of Russell’s account, nor the final verdict on it, are important. What is notable is the characteristically naturalist attitude that shifts the emphasis from an eliminativistic conclusion to one that is sensitive to both sides of the story, the philosophical and the scientific.

What is fundamentally meant by causation is of course an interesting and relevant question, but not the question on which the present study is focused. The aim is not to develop a naturalistically adequate account of causal explanation. This study does not even rely on some particular view of causation. The reasons for such agnosticism were laid out in the reasoning just presented. Since neither pure philosophy nor scientific practice on its own seems able to provide a complete account of causation, we need to rely on advice from both. That advice is scattered. There are many different but mutually inconsistent philosophical intuitions, and many different scientific disciplines relying on different explanation practices. Fitting all of this together to form, ultimately, a naturalistically solid account of causation that is both philosophically and scientifically useful will depend on an iterative process of give and take. We need to contrast philosophical analysis with scientific practice, piece by piece, back and forth. Sudden revolutionary insights are not in sight, the steps are inevitably small. What we can and must do is continue to present different causal scenarios and test them both philosophically and scientifically. Then, ultimately, we may reach a better understanding of causal explanation and its scope. We need to look at the issue from different angles. And this, in essence, is what the present thesis sets out to do.
3. **WHAT IS THE PROBLEM OF MENTAL CAUSATION?**

As already stated, this study treats the problem of mental causation as a general problem for philosophy of science at an abstract metaphysical level. However, it is instructive to begin by enumerating some of the issues that make up the problem of mental causation. Often in philosophical discussion the “mental” and the “physical” are stated to be difficult to reconcile. But why is that? What are the features in one that make it so ill-suited for interacting with the other? Note that even though it is the mental realm that typically gets represented as some sort of an anomaly, on a closer inspection it is not at all straightforward how the physical realm should be defined (as became apparent in the preceding discussion). However, there are features in these two notions that are clearly in tension with each other. Thus, we can recognize the problems we are facing in trying to combine these two realms, even though both lack clear-cut definitions.

There are at least five different problems that need to be separated. Although they are connected, each has its own characteristics and an industry of philosophical analysis built around it. None of them is negligible, but not all of them are relevant to the current discussion. Hence they are mentioned here mainly to sharpen the focus of this study.

Let us start with a problem that was already outlined in the introduction, *the problem of causal exclusion*. As this is the problem that this study is mainly concerned with, a short description of the issue will now suffice. Suppose that mental states, our conscious decisions, can have an impact on the physical world. We can control the movements of our limbs and alter the course of physical events around us, for example. But suppose also that the changes in the material world can be fully accounted for (to the extent that they can be accounted for) by physical means. That is, suppose that every physical effect has a sufficient physical cause. But then it appears that every physical effect that is caused by mental states is caused twice over: once by mental means and once by physical means. This seems like an overuse of causes. If that is the case, then one of the two will have to go. And if we are not prepared to allow the mental to have a direct, independent influence on the material world, then it seems that we have no choice but to exclude it altogether: mental states have no real effect on the material world but only piggyback on a single continuous and self-contained chain of physical causes and effects.

Second, there is *the problem of representational* or *intentional content*. Our mental states, at least those that we deem “cognitive”, have semantic content; our thoughts, beliefs and desires are about something. I can think of something, and you can think of the same thing, and the fact that our mental states have these particular contents can explain our behaviour. A simple example: that we both utter “Julius Caesar” when asked which great Roman statesman was murdered on the Ides of March 44 BC, is explained by the fact
that we both believe that Julius Caesar was murdered on that day. We seem to share the same belief, i.e. be in the same mental state with respect to this particular issue, and this explains our responses.

Moreover, it seems that this intentional content of our mental states is something that is constitutive of them. For suppose that these states are necessarily physically grounded, but not uniquely so. That is, although mental states do not exist on their own, independently of any physical basis, they are not type-identical with some such basis. What makes a mental state the particular mental state it is, is its semantic content, not the fact that it happens to be physically realized in a certain way. In consequence, different physical states could realize one and the same mental state, as is arguably the case with our beliefs about Julius Caesar: our beliefs are identical, but the physical (neural) states grounding these beliefs presumably are not.

However, representational content is by definition something relational: representation is a relation between the symbol of representation and the object of representation. Given that the object of representation stands outside the person and her mental states, it becomes puzzling how that object, or the relation between a given mental state and that object, can have an effect on the person and her behaviour. Now, although the issue of how terms come to have their references fixed is one of the central problems of philosophy of language, this is not the concern here. The problem is that whatever way references become fixed, and whatever meaning is, it is something that would not appear to make a causal difference in our thinking or our behaviour. It would seem that the syntactic features of our mental states and processes are all that is required to account for our behaviour. It might be convenient for us to describe our behaviour from an intentional perspective, but what really matters in how a certain output results from given input are the computational inner workings of the system.

Note that couched in these terms, the problem of representational content is a sort of a special case of the more general problem of causal exclusion. There seem to be two possible explanations for our behaviour, the semantic and the syntactic. But if both are equally capable of accounting for our behaviour, it becomes relevant to ask if we are postulating unnecessarily many explanations. And given that the syntactic account seems to be both necessary (in the sense that there is no semantic content without a computational platform) and sufficient to produce the given output, the syntactic description of the behaviour can be seen to pre-empt the semantic description of it.

Third, there is the problem of consciousness. “Consciousness” is a tricky word. The whole problem of mental causation can be said to be about consciousness: unless we are conscious of what we are doing, our behaviour can hardly be attributed to us. But there are different notions of consciousness. If we are not conscious of our actions, we are not aware of them. However, when we are discussing the philosophical problem of con-
sciousness, we are concerned with a different issue: the nature of subjective, qualitative consciousness, and its relation to other mental features and overt behaviour.

A way to sharpen these distinctions is to separate two different problems, the easy and the hard (Chalmers 1995, 1996). The “easy” problem concerns those aspects of consciousness that can be analyzed by means of a functional role. For example, awareness is something that we can all appreciate as being useful to the subject who is being aware: by being aware we can observe our environment and guide our actions in accordance with our intentions. Awareness is something that helps us to attain what we want and avoid what we do not want. This is a capability that is clearly advantageous and, as such, something that we have no trouble assigning a functional role to. However, what is really peculiar are the subjective qualitative aspects that accompany these experiences and feelings through which we become aware of our surroundings. We are not merely aware of things around us, or of our own mental states, but there is a specific qualitative aspect to these experiences. Seeing a sunset on an African savannah looks like something very specific, and being stung by a bee feels like something very specific. These qualitative aspects of our experiences are characteristically mental, but can we assign a functional role to them? It seems not. We could do very well without the nasty feeling of pain, as long as the experience conveyed the information that we are in a situation we had better avoid. But if these qualitative aspects of our experiences do not have a role to play in how these experiences affect us, it seems that they cannot figure in explaining our behaviour.

Fourth, there is the problem of free will. This problem is often associated with the problem of consciousness, especially in scientific discussions. Here it is not consciousness in the sense of subjective qualitative mental content that is at issue, but rather consciousness in the sense of awareness. In this context awareness appears as a necessary condition of free will: only agents capable of being aware of their actions can have free will. However, although awareness can constitute an interesting and relevant subject for psychological research, the proper problem of free will lies elsewhere.

On closer analysis it becomes apparent that it is not at all clear what “free will” is supposed to mean. At first sight the problem seems to be primarily about our apparent inability to influence the physical course of events: if events, including our mental states that are dependent on physical states of our body, are determined by prior physical events, then there seems to be no room for our conscious decisions to have an effect on the world. That is, free will is in contradiction with determinism. However, free will is also in contradiction with indeterminism. If the physical course of events were totally random, we would be equally incapable of influencing it. In fact, it seems that the notion of free will is not in contradiction with determinism tout court, but actually presumes a sort of determinism, namely psychological determinism: the ability of our mental states to determine courses of events.
One aspect of the problem of free will is clearly reminiscent of the problem of causal exclusion. If we think that every physical effect has a sufficient physical cause, then it seems that our conscious decisions and our will for the given effect to occur are completely superfluous. What free will seems to require is that we can cut into these physical chains of events and change them – at will. If that is the case, then the idea is in direct contradiction to the physicalistic thesis that all physical causes have a physical effect. But there are also other aspects of the problem that touch on the very image of ourselves as conscious and morally responsible agents. The fear is that unless there is genuine mental causation in the sense of free will, we are stripped of moral responsibility. No doubt this is a worry that largely motivates the whole discussion on mental causation. However, the verdict on this issue does not depend solely on our account of causation and the role of mental states in it, but clearly also on what it takes to be a conscious person and a morally responsible agent. Such questions are outside the scope of this study.

Fifth, there is the problem of normativity. This is a characteristically philosophical problem. The issue is particularly thorny, incorporating themes not only from the philosophy of mind, but also from philosophy of language and from the general discussions on the role of normativity within the naturalistic scheme.

The basic issue could perhaps be stated in the following way. There seem to be rationality constraints on our behaving on the grounds of mental states. That is, to act in a way that is essentially dependent on our mental states, our actions must conform to some predisposed rules for what it takes to act in that particular way. In other words: completely irrational behaviour is typically considered not to be due to mental causes, for such behaviour does not match our expectations on how one should act based on such causes. But then it appears that whether we have a particular mental state or not, and whether this mental state is causally responsible for our behaviour, is dependent on how we ought to behave. Such normative facts are in turn difficult to incorporate into the purely descriptive, naturalistic way of explaining.

There are many different ways to state this problem more precisely. One way stems from the Wittgensteinian (1953) idea that meaning itself is essentially a social and normative phenomenon: to know the meaning of a term or to have an understanding of a concept is constitutively dependent on the subject’s ability to use that concept in appropriate situations in appropriate ways. This is what meaning is. But since concepts, and grasping their meaning, seem to be an essential part of explaining behaviour in terms of intentional content, this content, and the explanation that rests upon it, seems to become dependent on the way these concepts ought to be used in social contexts. This, in turn, seems to render explanations in terms of intentional content essentially non-naturalistic.

Another way is to draw a distinction between reasons and causes and claim that it is the former rather than the latter that are essential in explaining behaviour. This idea has
attractive intuitive backing: when asking why a person behaved in the way that she did, we are typically not interested in hearing a neuro-muscular lecture on her behaviour, but an account that gives us understandable reasons for her acting the way that she did. What this seems to require is that in order to act in a manner due to mental causes, we are required to possess some means of rationally justifying our behaviour. But again, justification seems to be a thoroughly social and normative phenomenon. Moreover, giving reasons for actions seems to be an essentially teleological project: acting in a certain way only makes sense in a context where desires and beliefs are appropriately linked to specific goals that the given agent has in mind. But teleological explanation is in direct conflict with the naturalistic view that is committed to the idea that there is no purpose or primitively goal-directed behaviour in nature, and that all proper explanation is ultimately causal and mechanistic.

These five problems, at least, are at issue when “the” problem of mental causation is evoked. They are separate problems, each with their own characteristics. It is clear that answering some of these issues may not help to solve the others. But the problems are also clearly linked, some more closely than others. One recurring theme seems to be the problem of causal exclusion. It appears to lie behind many of these separate problems, at least partly. Whether this diagnosis can be shown to hold, and whether unlocking the causal exclusion problem would result in the unlocking of some of these other problems as well, is certainly an interesting and relevant question. However, this is not the question this thesis is primarily concerned with. What is of prime concern is the causal exclusion problem itself. In the following, the problem will be presented in more detail together with some responses and elaborations. The upshot of this analysis will be that the problem is in need of clarifications, and that once these clarifications are made, the simple reductionist solution may not be the only response at the naturalist’s disposal.

4. CAUSAL EXCLUSION

Let us start by stating the causal exclusion argument in a more rigorous manner (Kim 1998, 2005). Suppose that a mental state \( M \) is realized by a physical state \( P \). \( M \) and \( P \) are supposed to be distinct from each other, that is, \( M \) is neither identical nor reducible to \( P \). However, \( M \) is dependent on \( P \) in the sense that \( M \) could not have occurred on its own, without a realizing physical state, and given that \( P \) occurred, it was necessary for \( M \) to occur. Suppose now that there is a further physical state \( P^* \), and suppose that \( P \) is causally sufficient for \( P^* \). What role does \( M \) have to play in this? It seems none. Since \( P \) is already sufficient for \( P^* \) to occur, \( M \) is left only a role as an overdeterminer. However, such causal overdetermination where mental states would systematically accompany
physical states as distinct causes is unacceptable. Hence, M must be excluded as a cause of P*.

The problem can be seen to grow out of distinct elements. Consider the following set of claims (Bennett 2008):

<table>
<thead>
<tr>
<th>Claim</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinctness</td>
<td>Mental states are distinct from physical states.</td>
</tr>
<tr>
<td>Completeness</td>
<td>Every physical effect has a sufficient physical cause.</td>
</tr>
<tr>
<td>Efficacy</td>
<td>Mental states cause physical effects (in virtue of their mental features).</td>
</tr>
<tr>
<td>Non-overdetermination</td>
<td>Systematic overdetermination is impossible.</td>
</tr>
<tr>
<td>Exclusion</td>
<td>No effect has more than one sufficient cause unless overdetermined.</td>
</tr>
</tbody>
</table>

These claims seem to be inconsistent with each other. One of them has to go, but which one? That is the problem of causal exclusion. The argument of causal exclusion is the argument to the effect that the first claim, the distinctness of the mental from the physical, is the claim that should be dispensed with.

The following will present the three key elements of the issue. These are the distinctness assumption, the completeness assumption and the claim that overdetermining causes should be excluded.

4.1 Distinctness

Mental states are supposed to be distinct from physical states, yet they are also supposed to be dependent on them. This is not very informative. In what sense are the two distinct from, yet dependent on, each other? This notion clearly calls for clarification.

According to the received view, M and P can have this sort of relationship, when the former supervenes on the latter. Supervenience is a notion with a well-defined meaning: a set of S-properties (supervening properties) can be said to supervise on a set of R-properties (realizing or subvening properties) when entities cannot differ with respect to their S-properties without differing with respect to their R-properties. In other words, if two entities are identical with respect to their R-properties, they are necessarily identical with respect to their S-properties.

Supervenience seems to be well-suited for analyzing the relationship of the mental and the physical in naturalistic terms. What the notion entails is that once all the physical facts in the world are fixed, all the mental facts are fixed as well: “Any world which is a minimal physical duplicate of our world is a duplicate *simpliciter*” (Jackson 1998, 12).
Mental states or properties do not hover in their own non-material realm of existence, but are completely determined and dependent on the physical nature of the world. In this sense mental states are nothing “over and above” the physical states they supervene on. This is perfectly in line with what the monistic and physicalistic metaphysics of naturalistic philosophy requires. Moreover, the notion also entails that in order to manipulate a mental state you need to manipulate the physical state it supervenes on. There is no direct access to mental states, but all their changes must proceed via changes in the subvensing physical states. This also accords well with our physicalist intuitions that all mental changes are necessarily accompanied by changes in the neural system grounding those mental states. Otherwise neuroscience would hardly make much sense: it is based on the idea that we can study mental processes by studying the neural processes underlying them.

But now the obvious question arises: if the mental and the physical are so tightly bound together, is there any reason to hold them as distinct? If the mental is not identical or reducible to the physical, must it not be something “over and above” the physical? This is the question that continues to burden the non-reductive physicalist. The conventional reply has been that the supervenience thesis leaves room for the mental states to be multiply realizable: it is consistent with the notion that two entities might differ with respect to their $R$-properties without differing with respect to their $S$-properties. Such cases would satisfy the physicalist intuition that no mental state occurs without some physical state that realizes it. However, these two states would not have to be identical, for the very same mental state could have occurred without this particular physical state. At the most, the two states are token-identical: this particular event brings the two together, by means of physical realization, which explains how the mental could have a role to play in the physical courses of events.

However, a moment’s reflection will show that even if the supervenience thesis is made more concrete in this manner, there are still fundamental problems left unaddressed. Ever since Davidson (1970) brought the supervenience notion to the contemporary discussion on the philosophy of mind, it has played the role of a naturalistically respectable way of explaining the mind-body relationship. But it is clear that it provides no such thing. By itself the notion is compatible with a variety of accounts of the mind-body problem, such as epiphenomenalism, emergent dualism and identity theory. All of these accounts respect the idea that there cannot be differences in mental states without differences in the underlying physical states. More will need to be said in order for the notion to become metaphysically useful. Note also that although the notion is typically presented as a particularly physicalistic way of accounting for the mind-body relationship, there is nothing physical about it, quite the contrary. This becomes clear if we look at the history of the notion. Most notably, Moore (1903) used the notion (although not the term) to argue for a non-naturalistic account of the normative realm. Moreover,
the British emergentists used it to elucidate the relationship of basic physical properties and higher-level emergent properties. This relationship was inexplicable in naturalistic terms, and often the emergent properties possessed novel causal powers. As such, the use of the term seems merely to add mystery to mystery. If, then, the naturalist seeks to use the term to explicate the relationship of the mental and the physical, it needs to be made thicker and more naturalistically viable.

A further thing to note is that it is doubtful whether appealing to the multiple realizability and token-identity theses is able to amplify the notion in a satisfactory way. The basic problem is that once the mental and the physical are deemed distinct, it urges us to seek distinct causal roles for the two. But the mental cannot have such a role, if the physical realm is causally complete. At most, it is left with an epiphenomenal or redundant role. What we should try to find is a role for the mental features of the given mental state to play: such that the mental state, *qua* mental state, would make a causal difference. However, the token-identity thesis seems to aspire to just the opposite: to make the mental state causally efficacious by identifying it with a physical state. Moreover, sometimes the supervening properties can be clearly epiphenomenal. Consider this example:

Suppose ricketiness, in a car, is defined as the property of having some loose part. Then ricketiness will supervene on physical properties. In a given car, it may be realized by a disconnected wire between ignition and starter motor. This disconnected wire will cause this car not to start. But it doesn’t follow that this car’s then not starting will be caused by its property of ricketiness. Most rickety cars start perfectly well. (Papineau 2007.)

It may well be that ricketiness, in the way described, could in some circumstances constitute an acceptable explanation for the fact that the car does not start. But it should be incontestable that an explanation in terms how the ricketiness happens to be realized will be better and more informative. More importantly, if we start to inquire after the causes of this state of affairs, and if we want to make the car start again, the effective strategy is not to intervene on its “ricketiness” but to intervene on the physical state behind the ricketiness. Ricketiness may supervene on this physical mechanism, and be token-identical with it, but it does not appear to play the causal role played by the underlying physical mechanism.

4.2 The causal completeness of the physical

The causal completeness of the physical can be taken to be one of the defining features of physicalism. The basic idea is easy to grasp: the physicalistic conviction is that changes in the physical world can be perfectly accounted for by physical means. There are no gaps,
and there are no miracles. However, there are different ways to define this idea in more detail.

Let us make a distinction between completeness and closure. The causal completeness thesis could be stated thus:

Every physical effect has a sufficient physical cause.

The causal closure thesis could be stated thus:

Every physical effect has only physical causes.

Note first a feature that connects these two theses. They both refer to physical events that have causes, namely: physical effects. Neither of the theses should thus be confused with the “principle of sufficient reason”, which states that everything has a cause. The completeness and closure theses require only that those events that have causes have physical causes.

The difference between the two theses is obvious: the first one requires only that physical events that have causes have physical causes, while the second one excludes the possibility of any other causes. These theses, and their titles, are often used interchangeably, but it is important to maintain the distinction, for the latter is clearly stronger than the former. It commits us to the idea that we must not go outside the physical realm in our search for the causes of physical effects, whereas the former only commits us to the idea that we need not go beyond the physical realm.

It is also clear that the latter is unsuitable for discussing the causal exclusion problem: there is hardly a problem if one of its key elements states outright that physical effects have only physical causes. There would simply be no room for mental causation, at least not with respect to the physical realm. The completeness thesis does leave such room, albeit in a redundant sense, but enough to keep the discussion alive. This difference should not, of course, be allowed to count in favour of either of the theses. It may well be that the physical realm is more adequately characterized as being closed than as being merely complete. If that turns out to be the case, then the last half century of philosophy of mind has been in vain. However, this is not the place to decide between these two theses. Since the discussion here concentrates on the causal exclusion problem, it is simply assumed that the former formulation is adequate.

The thesis has obvious connections to physics, or at least it reminds us of some very fundamental physical principles, namely the conservation principles, and the conservation of energy in particular. The textbook definition of conservation of energy, or the 1st law of thermodynamics, states:

The total energy of an isolated system is constant.
Supposing that effects can be construed as some sort of energetic changes in a given system, the principle of conservation of energy seems to imply that for any given effect in a system there will be an another corresponding energetic change in the system – a cause – and that the two changes will balance each other out. This seems prima facie to be an appealing train of thought. A further interesting (and neglected) historical fact is that Julius Robert von Mayer, who was the first to state the modern principle of conservation of energy, originally derived it through steps of metaphysical reasoning concerning causal interactions. Based on the then widely accepted metaphysical principle that causes and effects in a causal chain must be equal, he derived two general properties of causal interactions: the first is that causes are covertly immutable; the second is that causes must be overtly transformable. This in turn led him to conclude that “taking both properties together, we may say, causes are (quantitatively) indestructible and (qualitatively) convertible objects” (Mayer 1842, p. 252). Given that he had already assumed that “forces are causes”, this conclusion led him to be the first to formulate the idea that we now know as the 1st law of thermodynamics.

How, or whether, this connection between the causal completeness of physics and the principle of conservation of energy can be made to hold, is something that deserves more detailed study. The intuitive kinship is obvious. However, there are also important differences. Most notably, it seems clear that if the connection between the causal completeness of physics and the principle of conservation of energy is very tight, the completeness thesis will collapse into the closure thesis. If causes and effects, or whatever is transferred between them, should remain invariant across causal interactions, there is no room for any additional supervening causes to make their contribution. Any such causes would simply constitute a violation of the conservation of energy.

Whatever the details of these two principles, all of these considerations seem to point to the basic physicalistic idea that there have to be underlying physical processes in every causal interaction, and that these processes are necessarily complete. There cannot be independent, totally autonomous, mental to physical causation, but the mental has to work via, or in tandem with, the underlying physical processes. To suppose that the mental could influence the physical realm directly, without any physical cause as a partner, would amount to the mental appearing as a physical anomaly: physical events caused by mental states would appear to us as “miracles” and the physical description of the world would become “gappy”, with physical processes vanishing and appearing from nowhere. This is something that is unacceptable to every naturalist – no matter how non-reductivist. The basic physicalistic conviction is that there cannot be gaps in our description of the physical world. This is the conviction that the thesis of causal completeness of physics tries to capture.

---

Pernu: Interactions and Exclusions
4.3 Overdetermination and exclusion

Although the previous two elements of the causal exclusion problem have thorny issues related to them, they are nevertheless concerned with very basic commitments of naturalism. Mental states simply cannot exist on their own, but need a physical basis on which to depend. And mental states cannot exert causal influence in the physical world on their own, but need the company of physical states. But the last element in the exclusion problem is more problematic, or at least it is something that does not follow directly from the basic commitments of naturalism, but rather from more general metaphysical presuppositions.

Let us first make it clear why this last element is necessary to complete the problem. One might just maintain that while \( M \) supervenes on \( P \), and \( P \) causes \( P^* \), \( M \) also causes \( P^* \). Since \( M \) and \( P \) are supposed to be distinct, \( P^* \) would now have two causes. And this would happen systematically: always, when a suitable physical state causes some other physical state, there is a mental state supervening on the first physical state that also causes the latter physical state. One might claim that this is exactly as it should be, and that there is nothing wrong with the picture. If one is not uncomfortable in postulating such redundant causes, there seems to be no logical contradiction in such a move.

However, consider now the following “Principle of Causal Exclusion”:

If an event \( E \) has a sufficient cause \( C \) (at \( t \)), no event (at \( t \)) distinct from \( C \) can be a cause of \( E \) – unless this is a genuine case of causal overdetermination.

With this addition, the depicted solution does not work anymore. It has already been granted that \( M \) and \( P \) are distinct, hence the principle of causal exclusion seems to apply. The only option left to defend the depicted solution would be to bite the bullet and claim that every case of mental causation is a genuine case of causal overdetermination. But genuine cases of causal overdetermination are typically thought to be rare and coincidental. Mental causation is thought to be something quite different: mundane and intentional. Claiming that mental causes are always overdetermining causes does not seem plausible.

Note first that the principle of causal exclusion, and the whole argumentation concerning overdetermination, is *prima facie* independent of any physicalist commitments. The principle seems to address the issue at a general metaphysical level. As such it is something that could easily be acceptable. However, general metaphysical intuitions tend to be murky and contradictory, and the principle of causal exclusion is no exception to this rule.

The notion of causal overdetermination is especially problematic. What is causal overdetermination? The notion is often used but rarely defined precisely. The typical
definition (for two causes) goes something like this: causes $C_1$ and $C_2$ are overdetermining causes of an effect $E$, iff $C_1$ and $C_2$ are distinct and independently sufficient for $E$ to occur. Two rocks thrown at a window constitute over-determining causes of the window’s breaking if it is the case that each of the rocks would have broken the window on its own.

But now it becomes apparent that this notion is difficult to apply to the case at hand. First, are $M$ and $P$ distinct? That is what the non-reductive physicalist is committed to believe. But she is also committed to believe that the two are connected. Whatever the distinctness claim amounts to, it is clear is that the two are not distinct in a similar way as two rocks thrown at a window: first, you cannot intervene on just one of the events and prevent it from occurring, while leaving the other event in place, the way that you can on either of the rock throwings. Second, are both $M$ and $P$ “independently sufficient” for $P^*$ to occur? Much of the preceding discussion revolved around the physicalistic conviction that $M$ alone was not capable of doing any causing but had to be assisted by $P$. It seems that $M$ alone is not “independently sufficient” for $P^*$ to occur, at least not in the way that each of the rocks is independently sufficient to break the window.

What this suggests, at minimum, is that there are at least two different notions of causal overdetermination at play. Overdetermining causes are typically construed as spatiotemporally distinct. This type of causal overdetermination can be called horizontal causal overdetermination; $C_1$ and $C_2$ occur at the same level of reality and are thus “horizontally” distinct. But sometimes, as is the case with the causal exclusion problem, the focus is on a type of overdetermination where the overdetermining causes are attributed to different hierarchical levels of reality. This type of causal overdetermination can be called vertical causal overdetermination; $C_1$ and $C_2$ occur at different levels of reality and are thus “vertically” distinct. (Cf. Funkhouser (2002) who discusses “incorporating” and “iterative overdetermination”, and Paul (2007) who discusses “constitutive overdetermination”.)

The way might now be open to argue that although horizontal cases of causal overdetermination are bad, or at least should not be systematically postulated, there is nothing wrong with systematic vertical causal overdetermination (e.g. Schaffer 2003; Sider 2003). Although this reply might work with some other cases of vertical overdetermination, it does not seem very promising in this case. A lot hinges on how the “distinctness” of the levels of reality is understood. Note that the relationship between overdetermination and distinctness can be construed as a trade-off. The more you pull two events apart by stressing their distinctness, the more exposed you become to the threat of overdetermination. The more you bring the two closer to each other, the less you may need to worry about overdetermination, but the more you are in danger of making the two identical. Surely there is no problem with $M$ being an overdetermining cause of $P^*$ if $M = P$? But that is exactly the claim that the non-reductive physicalist is committed to avoid. So you need to stress the distinctness of the two. But once you do that, it is not
clear why $M$ should not become equipped to function as an overdeterminer, albeit as a vertical one. Many cases that might be construed as involving vertical overdetermination are different in this respect. For example, you might claim that a single rock that flies through a window constitutes a vertically overdetermining cause of the breaking of the window, because the shattering is caused both by the rock and by its parts (cf. Merricks 2001; Sider 2003). Here, hopefully, you are not at pains to identify the rock with its parts. Or, if you are, then you should hold that the rock and its parts are *distinct* physical objects, with different masses and other physical properties, which in turn accounts for their separate causal roles. Mental states and physical states that realize them are not attributed to distinct physical objects, however: they are token-identical states of one and the same entity. Hence the issue of vertical causal overdetermination with respect to the case at hand might not be on a par with other such more concrete cases.

Although applying the notion of causal overdetermination to the relationship between the mental and the physical is clearly complicated, it is nevertheless a widely shared assumption in the discussion that worries about overdetermination are in this case justified. However, construing mental states as overdetermining causes is not in itself problematic. The real conflict arises when the principle of causal exclusion is brought on board.

The exclusion principle is simply a rejection of systematic causal overdetermination. It states that events cannot be systematically double-caused, and if this appears to be the case, one of the causes has to go. Since the principle is stated as a general metaphysical thesis, it does not refer to mental or physical states, or even to higher and lower-level states. It assumes that these are simply special cases of a more general metaphysical principle grounding causal interactions. But as such, it does not in itself dictate that it is the mental cause that has to be excluded. Supposing that both mental and physical states are individually sufficient for the given effect, and supposing that this is systematically so, the principle merely urges us to exclude either. The conclusion that the mental causes have to go results from joining the principle with the completeness thesis: since the physical causes are always there, the mental causes appear redundant and eliminable.

Although the principle is quite generally thought to be plausible, or at least intuitively appealing, it is practically never argued for. It is simply assumed to be true. As such, it is simply a metaphysical credo, a conviction that things are not typically double-caused. However, even if this is so, it would be instructive to learn why such a conviction is so appealing. At least two arguments can be made in support of the principle.

First, it is fairly obvious that overdetermination is closely related to epiphenomenalism. Suppose that one accepts everything else that has been said so far, but rejects the exclusion principle. That would result in a view where mental states would always accompany physical states as causes. But why would we postulate such causes? What difference would such causes make? Apparently none. Since we have already on inde-
Pendent grounds decided that physical causes are all we need, what purpose – other than soothing our worries – would it serve to postulate these additional overdetermining mental causes? Overdetermining causes are by definition redundant or superfluous causes, they are something unnecessary and extra; they are something we can well do without. That is why the principle sounds so appealing.

Second, although the principle is supposed to be a general metaphysical thesis, independent of physicalism or any other metaphysical accounts, it has obvious connections to the previous element of the exclusion problem. There are reasons to suspect that it is a derivative of the closure formulation of the completeness thesis, or of the basic ideas concerning conservation of energy. For, as already noted, if our physical notion of causation is tied to conservation principles, and to conservation of energy in particular, then something of an exclusion principle will become a built-in feature of causation: once a physical cause for an effect is designated, it fills the only place vacant for a cause to fill and no other cause can make its contribution. Note that on this account the exclusion results through elimination of causal overdetermination altogether: it is not due to the mental cause being mental that results in it being excluded, rather it is it being an overdeterminer that makes it unsuitable as a cause. All overdetermining causes would need to be eliminated and the whole notion of causal overdetermination would need to be waived. Being such a problematic notion as it is, this result might just be greeted as a welcome bonus.

5. RESPONSES TO CAUSAL EXCLUSION

Although the causal exclusion problem is put together from elements that seem fairly incontestable to the naturalist, discussions are rife with responses to it. Each step in the causal exclusion argument has been questioned. Here, however, only responses directly relevant to this study will be reviewed, namely those that concern the account of causation on which the exclusion issue relies.

5.1 Mental causation and physical causation

If you go to a layperson with some philosophical sense and ask her whether there is such a thing as mental causation, what would she reply? Most likely: “It depends on what you mean by ‘causation.’” In this case the reply is perhaps more appropriate than ever. It is by no means obvious what causation is. There are many different accounts, all with their own characteristics and problems. It is therefore by no means quibbling to hold one’s judgement and await further information. After all, different accounts of causation might result in different verdicts.
However, there is also an intuitive uneasiness with the notion of mental causation. Many feel that there is something wrong, or at least problematic, with it. It seems that we associate things with mental causation that are hard to reconcile with some of our other commitments. The preceding discussion has already detailed many of the concrete issues that might act as sources of conflict between the mental and the physical. But let us now look at things from the perspective of the causal relation itself.

Modern reactions to the mind-body problem often start by discussing Descartes. According to his dualistic thesis, minds and bodies are distinct substances: my body is one entity and my mind is another entity, distinct from my body. This idea has been endlessly criticized. However, it is largely a matter of taste whether the modern form of non-reductive physicalism is really that far from Descartes’ view. Substance dualism is explicitly rejected, of course, but holding on to the distinctness of the two realms seems to amount to some sort property dualism. The connections of these views also become apparent when we turn to look at the problematic causal relationship of the two realms.

Although postulating these two different substances may help us to explain some of the differences we attribute to these two realms, it immediately raises others. The most notable of these is the problem of interaction: if the mind and the body are separate substances, each with their own characteristic – and conflicting – features, how can the two be in a causal interaction with each other? A much cited formulation of this problem was made by Princess Elizabeth of Bohemia in 1643:

> [H]ow the human soul can determine the movement of the animal spirits in the body so as to perform voluntary acts – being as it is merely a conscious substance. For the determination of movement seems always to come about from the moving body’s being propelled – to depend on the kind of impulse it gets from what sets it in motion, or again, on the nature and shape of this latter thing’s surface. Now the first two conditions involve contact, and the third involves that the impelling thing has extension; but you utterly exclude extension from your notion of soul, and contact seems to me incompatible with a thing’s being immaterial.”

(Anscombe & Geach 1954, p. 274–275.)

Supposing, as Descartes did, that the two substances differ radically from each other, it becomes relevant to ask how the two can causally interact. One of the defining features of mental substances is their non-extension, whereas one of the defining features of physical substances is their extension. But if that is the case, how can a non-extend ed entity have a causal impact on an extended entity? Moreover, if causation (at least physical causation) is something that involves contact by extended bodies – as Princess Elizabeth in the quotation assumes – then the mental substances seem to be stripped of causal power.
The easy and often used reply to the problems raised here is that not only is Descartes’ view on the relationship of the mental and the physical invalid, but the notion of causation that he and many of his contemporaries relied on is now defunct. The line of thought is that if we only replace this naively mechanistic and outdated view of causation with a more sophisticated and modern one, the problem will evaporate. How the notion of causation is understood is at the heart of the matter.

Is the view presented in the quotation above naïve and outdated? It may be premature to announce it as such. First, the problems annoying the modern-day non-reductive physicalists are not that far from the problems that Descartes was facing. Mental and physical features are now attributed to one and the same entity, but the features themselves are supposed to be distinct. And that is, as before, the real source of the problem. How can mental states, *qua mental*, have an effect on the physical world that is causally complete – that is the question. The issue is still with something that is separate and fundamentally different having an influence on the physical. The only difference, it seems, is that now we are admitting that the physical realm is causally complete and that the mental is always physically realized.

Second, it is not far-fetched to say that the problems that we are now facing with the notion of mental causation stem from essentially similar commitments about fundamental physical causal interactions. Although the problem in the quotation is couched in terms of mechanical contact – of which a non-extended entity is incapable – and although such a notion of causation may be obsolete, there are still strong intuitions that causal interactions fundamentally involve physical contiguity and local interaction. This intuition is often shared even by those who are prone to analyze the notion of causation in terms other than physical processes. The idea that genuine causation essentially has something to do with physical production, generation, transference or “bringing about” is strong within our physicalist world view. This intuition may be wrong, of course, but it is certainly something that lies behind our difficulties in comprehending the notion of mental causation. And it is something that Princess Elizabeth would no doubt have found easy to accept.

Recent philosophy has also witnessed a rise of well-developed physical analyses of causation (e.g. Aronson 1971; Bigelow *et al.* 1988; Bigelow & Pargetter 1990; Castañeda 1980; Dowe 1995, 2000; Fair 1979; Heathcote 1989; Salmon 1984, 1994, 1997, 1998). Typically causation is understood to consist in energy/momentum transfer. Most notably Dowe (2000) defines a *causal process* as a world line of an object that possesses a conserved quantity, and a *causal interaction* as an intersection of world lines that involves exchange of a conserved quantity. Such accounts should have no trouble accepting the completeness thesis, exclusion arguments and – in essence – the view of causation that Princess Elizabeth endorsed (cf. Kim 2007). Although these accounts of causation are modern and elaborate, what the critics would still insist, however, is that they are based
on intuitions that should be given up – or, at least, that there are other stronger causal intuitions that we should respect at the cost of these physical intuitions. These intuitions, in turn, might amount to a different notion of causation. Maybe such a notion might also provide a way out of the exclusion worries? Let us turn to these accounts.

5.2 The counterfactual response

The most notable alternative to the physical notion of causation is the counterfactual notion of causation. This notion was introduced by Lewis (1973, 1986) and has since gained a large following. Typically this view is said to hold that an event \( C \) causes (a logically distinct event) \( E \) only in case it is true that if \( C \) had not occurred, \( E \) would not have occurred either. Again, this sounds like a prima facie intuitively plausible idea.

Let us first be more precise in defining the notion, for this will turn out to be important later on. According to the counterfactual account of causation, an event \( C \) causes the occurrence of another distinct event \( E \) (in a world \( w \)), iff the following pair of counterfactuals hold (in \( w \)):

\[
\begin{align*}
(a) & \quad C \rightarrow E \\
(b) & \quad \neg C \rightarrow \neg E
\end{align*}
\]

The first condition is usually omitted. Whether such an omission is justified depends on the details of the semantic interpretation imposed on these counterfactuals. If it is assumed, as it typically is, that the occurrence of \( C \) and \( E \) (in \( w \)) makes the first counterfactual vacuously true (in \( w \)), then the omission is understandable.

What is noticeable about this account is that it is the counterfactual dependence of \( E \) on \( C \) that makes the latter the cause of the former. That is, there is no other connection or process between the two. If we connect this to the preceding discussion we can draw a distinction between two radically different types of causal notions (Hall 2004). The first we can call the production notion. This is the idea that causation is essentially a matter of spatiotemporally local transference, exchange or production of a physical quantity. The second can be called the dependence notion. This is the idea that causation is simply a matter of counterfactual dependence between distinct events. Now two questions arise: should such a dependence notion be preferred at the cost of the production notion, and if so, is the notion able to provide a non-reductivist answer to the exclusion problem? Let us tackle the latter question first.

The basic idea, advocated numerous times and ways in the discussion, how the counterfactual notion of causation can help us to escape the exclusion problem is as follows. Suppose again that we have a mental state \( M \) supervening on a physical state \( P \) and that there is a physical state \( P^* \) for which \( P \) is causally sufficient. Now it appears that the fol-
lowing counterfactual is true: if $M$ had not occurred, then $P^*$ would not have occurred either. Given that this is sufficient for holding that $P^*$ is counterfactually dependent on $M$, and given that counterfactual dependence is all there is to causation, then it appears that there is such a thing as genuine mental causation after all.

Could it really be this easy? There are reasons to be doubtful. The following objection should immediately spring to mind: in virtue of what does this counterfactual hold? The reductive physicalist is ready to reply: the counterfactual holds simply in virtue of the underlying physical causal interaction between $P$ and $P^*$. Even an epiphenomenalist might not object to the apparent veracity of the counterfactual. She would just think of it as a reflection of the genuine underlying causal relationship. Thus, merely noting that the described counterfactual dependence exists is not enough to soothe the worries of those who are moved by the exclusion problem. It would need to be supplemented with an argument to the effect that counterfactual dependency is really all there is to causation.

The physical notion of causation – whatever its details – is definitely problematic, as already discussed. One strategy, then, is to criticise the very idea that there are any deeper physical causal processes of which the counterfactual dependencies are derivative (cf. Loewer 2007). That is, there are not two notions of causation, production and dependence, but only one, the latter. However, the counterfactual notion of causation, of course, is not free of problems. First, the account relies on the notion of possible worlds, which is notoriously problematic. They are ontologically problematic: what – and where – are all these possible worlds? But more importantly, they are semantically problematic: how to assess which worlds are closer than others to the world that we happen to be interested in? We can easily find ourselves on a slippery slope to all sorts of goofy scenarios that might not be that attractive to a physicalist. Second, it may be necessary for the counterfactual view to resort to elements of physical causation in trying to find a way to get rid of annoying counterexamples. Redundant causation (overdetermination and preemption) are problematic, for in these cases the described counterfactual dependence seems to break down: the occurrence of $E$ is no longer counterfactually dependent on $C$, since there is an alternative to $C$ that would bring about $E$ in the absence of $C$. Consider preemption. The theory needs to be supplemented by a clause that explains why the preempting cause rather than the preempted alternative brings about the effect. A natural, and widely accepted, solution is to appeal to the fact that there is a “complete process” between the preempting cause and the effect, which the preempted cause and the effect lack (e.g. Lewis 2000; McDermott 1995; Menzies 1989, 1996, 1999; Noordhof 1999; Ramachandran 1997; Rueger 2006; Schaffer 2001, 2003). How to cash out the notion of “complete process”? It turns out that these accounts typically build on the view developed by Dowe (2000) and Salmon (1984, 1994, 1997, 1998). It appears that we are back to square one.
5.3 Interventionism and difference-making

The discussions and notions reviewed above have been hugely influential and are still in many ways relevant. However, a more elaborate way of dealing with the exclusion issue has arisen in recent years. This interventionist or difference-making account is a descendant of the counterfactual view, which makes it a species of the dependency account of causation. However, it has been carefully constructed to avoid the problems the more traditional counterfactual account is bound to face.

The interventionist account of causation draws from the theory of causal explanation developed by Woodward (2003). Causation is defined thus:

A necessary and sufficient condition for $C$ to be a direct cause of $E$ with respect to some variable set $V$ is that there be a possible intervention on $C$ that will change $E$ (or the probability distribution of $E$) when all other variables in $V$ besides $C$ and $E$ are held fixed at some value by interventions. (Woodward 2003, p. 55.)

It is already clear from this short characterization that this account changes some of the basic elements in the discussion in a dramatic way. First, the preceding discussion has been overly vague about the exact nature of the causal relata. They have been referred to as properties, states and events. However, the interventionist account has a clear view on how the causal relata should be understood: they are variables (type-causation) or values of variables (token-causation). It may be that property instantiations or the occurrence of states or of events can be understood as a given variable assuming a specific value, but whether such analogies or reductions can be made to hold is not of concern here. What is more important is that this approach allows us to analyse claims about causal relationships in a much more precise way. Construing the causal relata as variables should also appeal to the naturalist, for this is the way causal relationships are represented throughout the sciences.

Second, causal claims are always assessed with respect to a specific variable set $V$. Again, this is a welcome clarification, for now you have to make it explicit right from the start what sort of variables (and relationships) the discussion pertains to. However, this also makes the results malleable, for changing the variable set might have an effect on whether or how a given causal relationship holds.

Third, “intervention” is clearly a causal notion. Causality is thus defined in terms that are causal themselves. In other words, the interventionist account of causation relies on a non-reductive analysis of causation (in contrast to the accounts discussed previously). Of course, it is pertinent to ask whether such an analysis is ultimately illuminating. However, this approach has the advantage that it sits in well with the explanatory practices prevalent in the special sciences. From this perspective it matters very little
whether causality can be analysed in physical or counterfactual terms. What is important is how causal inferences are actually made in the sciences. This is again something that should appeal to the naturalist.

Armed with these reforms, does the interventionist account offer a solution to the causal exclusion problem? It has certainly been argued that it does (e.g. Campbell 2008, 2010; List & Menzies 2009, Menzies 2008; Menzies & List 2010; Raatikainen 2010; Shapiro 2010, 2012; Shapiro & Sober 2007; Woodward 2008a, 2008b). However, what becomes immediately clear is that there are many interpretative difficulties in applying the notion to the exclusion issue. This is something that has been largely overlooked.

Suppose first that the variable set we are interested in is simply $V = \{M, P, P^*\}$ (throughout the discussion these are assumed to be binary variables). But this is not enough. What the exclusion problem supposes is that these variables are connected in subtle ways – that is the whole source of the problem. What the interventionist account assumes, however, is that the variables can be intervened on independently of each other (except for the cause and effect variables). Suppose that the states to which the variables in $V$ refer all occur, i.e., the variables assume the values $M = 1, P = 1, P^* = 1$. Suppose that you now intervene on $M$ and change its value to 0, to see whether it is the cause of $P^*$. To honour the definition of direct cause, you would need to hold $P$ fixed. But that cannot be done because $M$ supervenes on $P$ and holding $P$ fixed would in this case amount to $M$ being held fixed, which contradicts the idea of testing whether $M$ is a cause by changing its value. Since it appears that there is no intervention on $M$ that would change the value of $P^*$ while other variables in $V$ (i.e. $P$) are held fixed, the definition leads us to conclude that $M$ is not a cause of $P^*$ (Baumgartner 2009). This is epiphenomenalism, and not the result we were hoping for.

Suppose we relax the definition and assume that in this case – or more generally in cases where supervenient variables are intervened on – we are not compelled to hold the subvenient variable fixed. Now one could argue in the following way:

Because changing $M$ is impossible without simultaneously changing $M$’s supervenience base $P$, and because $P$ is a cause of $P^*$, a change in $M$ does result in a change in $P^*$. This is evidence that $M$ is a cause of $P^*$. (Shapiro 2010, p. 601.)

Is this conclusion plausible? Note that this is exactly the same inference that the traditional counterfactualist account attempted to offer as a solution to the exclusion problem. If it didn’t convince you then, it is difficult to see why it should convince you now. In fact, in this context the inference is even more clearly unfortunate, for now the result is that the values of $M$ and $P$ vary entirely in phase: setting $M$ to 1 results in setting $P$ to 1 and setting $M$ to 0 results in setting $P$ to 0, and vice versa. That is, $M$ and $P$ don’t seem
to differ at all with respect to their causal profiles. This certainly solves the exclusion problem: this is simply reductive physicalism. Again, not the result we were hoping for.

However, if we fiddle with V things get more interesting. Suppose we take into account the idea that M is multiply realizable, thus V = \{M, P_1, P_2, P^*\}. Now P_1 and P_2 are mutually exclusive supervenience bases of M: when M assumes the value 1, either of them also assumes the value 1; when M assumes the value 0, both of them assume the value 0. This arrangement allows us to intervene on these two subvenient variables while holding the supervenient variable fixed. Or, to be more precise, if the value of M is 1, then we can hold it fixed while intervening on its base, provided only that either P_1 or P_2 is assuming the value 1. Given this setting, intervening on either P_1 or P_2 does not result in a change in P^*, which means that neither of them is a cause of P^*. However, as was already established, intervention on M does result in such a change. Hence, M is a cause of P^* and the idea of mental causation is finally vindicated.

One way of putting this result in more concrete terms is to say that construing M as the cause of P^* captures the relationship at the right level. Although either P_1 or P_2 is necessarily there as well, each of them is too specific to function as a genuine cause of P^*. M seems to act as a “control switch” or “control variable” for P^* (cf. Campbell 2010): the values of P^* seem to change in accordance with the values of M, not in accordance with the values of P_1 or P_2.

Yet another way to formulate this idea is to say that M is a difference-making cause of P^*, whereas P_1 or P_2 are not. To say that M is a difference-making cause of P^* means simply that the following pair of counterfactuals hold (cf. List & Menzies 2009):

\begin{align*}
(1a) & \quad M \rightarrow P^* \\
(1b) & \quad \sim M \rightarrow \sim P^*
\end{align*}

It seems intuitively uncontroversial to accept these counterfactuals (as before, even a reductive physicalist or epiphenomenalist might accept them). But consider now the following two pairs of counterfactuals:

\begin{align*}
(2a) & \quad P_1 \rightarrow P^* \\
(2b) & \quad \sim P_1 \rightarrow \sim P^* \\
(3a) & \quad P_2 \rightarrow P^* \\
(3b) & \quad \sim P_2 \rightarrow \sim P^*
\end{align*}

Those attracted by this approach will claim that both (2b) and (3b) are actually false by arguing, roughly, in the following way. Given that M occurs and we hold it fixed, then in the case of (2b) the fact that P_1 does not occur must imply that P_2 occurs, and vice versa, in case of (3b) the fact that P_2 does not occur must imply that P_1 occurs. So the occurrence
of $M$ implies that either of the supervenience bases occurs, and hence intervening on these bases while holding $M$ fixed does not result in a change in $P^*$. To present this argument in detail would of course entail a lot of semantic jiggling, but the basic idea seems intuitively clear enough. No doubt this argumentation is something that the original counterfactual account was seeking all along.

The key assumption behind these results is that causes and effects must be proportional to each other (Yablo 1992). Proportionality can be understood as a thesis that causes must not contain anything more than what is necessary to bring about the effect. $M$ seems to be proportional to $P^*$, for the occurrence of $M$ brings about the occurrence of $P^*$ (1a) and the absence of $M$ brings about the absence of $P^*$ (1b). However, the supervenience bases of $M$ are not proportional to $P^*$, for the latter condition is not met.

This approach sounds attractive, but there are at least three problematic issues that it has to address.

First, the underlying proportionality assumption is by no means innocent. It has to face at least two different critical questions. First, are causal relationships really governed by such a requirement? Is it really an essential mark for genuine causal relations? There are intuitively clear cases of causal relationship that are not proportional (Shapiro & Sober 2012). Second, are physical effects really proportional to mental causes? Are they always such? The argumentation has simply assumed that such a proportionality relationship holds, but this is clearly an issue that needs more thorough analysis.

Second, the other key assumption in the argumentation is that mental states are multiply realizable. But again, two critical questions arise. First, is there such a thing as multiple realization? The standard assumption in the discussion has been that there is indeed such a thing. But this assumption has started to receive criticism (e.g. Bechtel & Mundale 1999; Bickle 1998, 2003, 2010; Couch 2004; Shapiro 2000, 2004, 2008). Perhaps the use of the thesis is justified because it has had such a prestigious position and there doesn’t seem to be much non-reductive physicalism left if the thesis is rejected. However, as the discussion progresses, this thesis cannot simply be taken for granted. Second, what role exactly does the multiple realizability thesis play in the argumentation? More specifically: how to interpret the idea that given that $M$ is held fixed and either $P_1$ or $P_2$ is intervened on, suddenly the other ”back-up” realizer pops up and assumes the position that corresponds to the occurrence of $M$ and $P^*$? It seems that this picture requires more detailed explanation and justification.

Third, the argumentation seems to run into trouble with the thesis of causal completeness. If it is claimed, as it typically is, that the argumentation not only shows how $M$ is the proper cause of $P^*$, but also shows how $P_1$ or $P_2$ is not, then the conclusion seems to amount to the rejection of the completeness thesis. But the thesis has typically been taken to be non-negotiable. Moreover, if one is prepared to reject the completeness thesis, why bother with the exclusion problem in the first place?
6. THE AUTONOMY SOLUTION AND INTERVENTIONIST PARALLELISM

There is still one remaining response for the causal exclusion problem to face. The *autonomy solution* stresses that there is really no competition between the mental and physical states over the status of the “real cause”. Instead, the mental is autonomous with respect to the physical, and vice versa. This response has different variations and is by no means a new invention. It emanates from Leibniz’s response to the problems evoked by Descartes’ interactionism, namely the idea of pre-established harmony. Although Leibniz is practically never credited in the current discussion, the view has many modern incarnations. The following will not review all of them, but rather focus on the one that is elaborated in the present thesis, namely the idea of *interventionist parallelism*.

The autonomy response has been presented in the current discussion in a number of different forms (e.g. Crisp & Warfield 2001; Gibbons 2006; Jackson 1996; Marras 1998; Schlosser 2009; Thomasson 1998). Again, merely saying that “the mental is autonomous with respect to the physical” does not amount to much. An epiphenomenalist and a reductive physicalist can accept such a statement. They do not necessarily have difficulties with the idea that the special sciences are autonomous with respect to fundamental physics. They do not necessarily advocate eliminativism and they can accept the idea that these higher-level descriptions and explanations may be useful to us and hence something worth holding on to. What they do resist is leaping from these pragmatic notions to the full-blown metaphysical notion that mental causation is a genuine and autonomous phenomenon. What the defender of this solution must do, then, is offer more than just pragmatic or epistemic considerations. The argument has to sink into the very notion of causal explanation.

One way of being an autonomist is to note that there is no problem with claiming that one and the same thing can have several different causal explanations that do not necessarily compete. Explanation is a highly context and interests sensitive matter. Surely, then, one thing can have several causal explanations, each appropriate in its own way? Causation itself might also be divided into different types. Already Aristotle distinguished four different types of causal relationships. These have not been considered to be in competition with each other. Yet, the reductive physicalist would insist that we talk about physical explanations and physical (or “efficient”) causation. Only causal explanations in these terms amount to fundamentally real explanations. Once this is acknowledged, other explanations seem to compete with them.

However, the typical autonomist response amounts to a stronger claim. According to it, the effect to which we are trying to pin-point the correct cause is distinct from the effect that has an unequivocally physical explanation. Thus, the mental and the physical are not in interaction with each other, but both mental causes and mental effects are distinct from physical causes and effects. Hence there is clearly no competition be-
tween these explanations. Again, both the reductive physicalist and the epiphenomenal-
ist might concede this view, but insist that now both the causes and effects are merely
reflections of the real, underlying physical causal processes. However, the force of this
view, when both causal relata are separated in this way, is that now both of the domains
become closed. Most notably: now mental effects – or higher-level effects in general –
can be shown to be due to exclusively mental causes.

Before introducing an argument to this effect in the interventionist framework, let
us review a general argument purporting to show that these sorts of autonomy solutions
cannot work. This is the so-called supervenience argument, an extension of the exclusion
argument (in fact, the latter is often formulated in terms of the former, but it is impor-
tant to keep these arguments apart). Suppose that instead of the question of whether the
mental state $M$ causes the physical state $P^*$, we are interested in the question of whether
the mental state $M$ causes another mental state $M^*$. According to Kim (1998, 2005), this
type of mental-to-mental causation will also succumb to the exclusion argument in the
following way. Given physicalism, $M^*$ has to have a physical realizer on which it super-
venes, namely $P^*$. Now, according to Kim (1998, 2005), there are two competing explana-
tions for the occurrence of $M^*$, namely $M$ and $P^*$. The relationship between $P^*$ and $M^*$ is
not causal, of course, but one can still see it as a competing explanation, or at least creat-
ing tension between these two explanations. But once $P^*$ is there, $M^*$ cannot help being
there, and hence this relationship seems like a better explanation for the occurrence of
$M^*$. So in order for $M$ to cause $M^*$ it has to cause the supervenience base of $M^*$, namely
$P^*$. But now we are back at the original exclusion problem, and the argument says that
$M$ is not a cause of $P^*$. Hence, $M$ is not a cause of $M^*$ either, and the autonomy solution
is refuted.

Based on the interventionist notion of causation, however, it can be shown that this
train of thought is not valid. That is, the supervenience argument can be shown to be
invalid although the exclusion argument is valid. This scenario also has the advantage of
putting many other problematic issues in place.

Note first that in order for any substantial argument to get off the ground, we need to
expand the supervenience bases. Suppose that simply $V = \{M, P, M^*, P^*\}$. Now, as before,
the supervenient and subvenient variables would vary in phase with each other. That
would amount to identifying $M$ with $P$ and $M^*$ with $P^*$, and that would not be a welcome
result. However, suppose that we now expand the supervenience bases and take into
account that $M^*$ is also supposed to be multiply realizable. So now let $V = \{M, P_1, P_2, M^*,
\}$, and suppose that $M$ happens to be realized by $P_1$ and $M^*$ by $P^*$, and thus the vari-
ables assume the following values: $M = 1, P_1 = 1, P_2 = 0, M^* = 1, P^* = 1, P^* = 0$. If we now
test different causal claims with respect to this variable set (and the relevant constraints
imposed on it) there are three results worth highlighting: $M$ is a cause of neither $P^*$ nor
$P^*$, and neither $P_1$ nor $P_2$ is a cause of $M^*$, but $M$ is a cause of $M^*$.
First, and contrary to what the previous interventionist argument showed, there is no such thing as mental-to-physical causation. In other words, the exclusion argument holds. For suppose that we now intervene on $M$ and change its value, and suppose we are interested in testing whether $M$ is a cause of $P^*$. Given the distribution of values of the variables, this particular intervention (that sets $M$ to 0) does indeed result in a change in $P^*$. However, suppose that we start with a different distribution, with one that corresponds to a situation where none of the states that these variables refer to occur, thus $M = o, P_1 = o, P_2 = o, M^* = o, P^*_1 = o, P^*_2 = o$. Intervene now on $M$ and change its value. Does the value of $P^*$ change accordingly? Not necessarily, because $P^*$ was an alternative supervenience base of $M^*$ and each of these bases could equally well occur. Setting $M$ to 1 could result, for example, in the following distribution of values: $M = 1, P_1 = o, P_2 = 1, M^* = 1, P^*_1 = o, P^*_2 = 1$. Hence, setting $M$ to 1 would result in setting $P^*_1 = 0$. This in turn would mean that $M$ would fail to be a “control variable” for $P^*$, and, consequently, a cause of $P^*$. The same holds, mutatis mutandis, with respect to $P^*_2$. Thus, $M$ is a cause of neither $P^*_1$ nor $P^*_2$.

Second, there is no such thing as physical-to-mental causation either. Suppose that we start with the original distribution of values and intervene on $P_1$, but hold $M$ fixed. But we are supposing that in such a case the alternative supervenience base of $M$ assumes the value of 1. This in turn leads to either $P^*_1$ or $P^*_2$, assuming the value of 1 (supposing that the bases of $M$ are causally sufficient for the bases of $M^*$). But that means that the value of $M^*$ remains unchanged. Hence $P_1$ would fail to be a “control variable” for $M^*$, and, consequently, a cause of $P^*_1$. The same holds, mutatis mutandis, with respect to $P^*_2$. Thus, neither $P_1$ nor $P_2$ is a cause of $M^*$.

Finally, there is such a thing as mental-to-mental causation. In other words, the supervenience argument fails to hold. Suppose that we start with the original distribution of values and intervene on $M$, but set both $P_1$ and $P_2$ fixed. This in turn means that all of the other variables are also set to 0. Hence, setting $M$ to 0 results in setting $M^*$ to 0. The original distribution of the values of the variables already shows that setting $M$ to 1 can result in setting $M^*$ to 1. However, supposing that the bases of $M$ are causally sufficient for the bases of $M^*$, setting $M$ to 1 will result in setting $M^*$ to 1. Hence $M$ would seem to be a “control variable” for $M^*$, and, consequently, a cause of $M^*$.

Since this argument is a developed version of the interventionist response to the causal exclusion argument it is expected to share its pros and cons. It relies on the proportionality requirement and multiple realizability thesis. As before, these are not unproblematic assumptions. However, interventionist parallelism has certain advantages over the interactionist version. Most notably, the physical domain is now safely closed: it was established that mental states never cause the occurrence of physical states. Moreover, the worries about overdetermination evaporate: physical and mental states cause separate effects, and thus there is no single effect that could become overdetermined.
However, rejecting interactionism works both ways. It was not only established that the physical domain is closed. It was also established that the mental domain is closed, i.e. there is no such thing as physical-to-mental causation either. In other words, the pin prick on your finger does not cause your pain. But isn't this unacceptably counter-intuitive? Not necessarily. We can interpret the causal processes thus: as the pin pushes through your skin this leads to a particular state in your nervous system that in turn realizes your pain experience. This sounds both scientifically acceptable and intuitively plausible. What we need to realize is that letting go of interactionism – which was a deeply problematic position to begin with – eliminates not only mental-to-physical causation, but also physical-to-mental causation. The best we can do is to embrace parallelism.

7. CONCLUSIONS

The notion of causal explanation is an essential element of the naturalistic world view. However, it is not something that can just be taken from a toolbox and used to solve any outstanding philosophical problem. Views vary on what the notion amounts to exactly. Fields of science vary with respect to their causal commitments. There is no single monolithic, unanimously accepted notion of causation that the naturalist can simply just put into use.

One of the metaphilosophical messages of this study is that the relationship between naturalistic philosophy and the notion of causal explanation requires clarification. What is the causal criterion for reality? What kind of notion of causation does that presuppose? Is there such a thing as physical causation? If there is, what is its relationship to causal explanation in other sciences? Naturalism and physicalism are undoubtedly close companions. That physical facts determine all the facts there are is the central conviction of naturalism. However, whether this entails a particular physical view of causation, and whether such a view is the only naturalistically acceptable one, is an open philosophical question.

The exclusion argument purports to show that physical causal processes indeed constitute all of the real causal processes that there are. However, this conclusion relies on the assumption that all genuine causes must have physical effects: the argument invites us to ask whether a given mental state that supervenes on a particular physical state can have an effect on another physical state. But isn’t this begging the question? Was it not the very idea of wholly physical causation that was called into question? No wonder the reductive view has the upper hand if the game is constantly played on its home ground. This game has been allowed to continue because the non-reductivist protests have been sparse. It has been taken for granted, by all parties, that genuine causal interaction results in physical effects. But is that really something that a naturalist must be committed to?
The non-reductive physicalist is ready to grant more to reality than basic physical properties and physical causal processes. Apparently that could, and should, include also effects other than the physical kind. Then, apparently, the view could, and should, include causal interactions that are completely distinct from underlying physical processes. Such parallelism should constitute the default non-reductivist position. But for some reason, this option has not been on the table.

The only reasonable explanation for this situation seems to be that some kind of physical notion of causation has been tacitly taken for granted. That is why the exclusion argument has seemed so insurmountable. But the non-reductivist is typically explicitly sceptical – on naturalistic grounds – of the physical notion of causation, and instead advocates a notion that relies solely on counterfactual dependence. What the present study shows is that, in this context, the parallelistic view is particularly enticing.

REFERENCES


