PANPSYCHISM AND CAUSATION: A NEW ARGUMENT AND A SOLUTION TO THE COMBINATION PROBLEM

Hedda Hassel Mørch

Ph.D. thesis

Department of Philosophy, Classics, History of Art and Ideas
Faculty of Humanities
University of Oslo

January
2014

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor, Camilla Serck-Hanssen, for all her advice, support, trust and encouragement during my time at the University of Oslo, first as a B.A. student, then as her M.A. student and finally as her Ph.D. student. She has been an insightful critic, an open-minded discussion partner, an inspiring teacher and, not least, a good friend. I could not have wished for more in a supervisor.

I am also very grateful to David Chalmers, who kindly accepted to be my secondary, external supervisor and has given very helpful comments on this thesis as well as on many talks on the same material. I would like to warmly thank him also for welcoming me to a long visit to ANU and a shorter visit to NYU, for inviting me to a workshop on panpsychism at the Great Barrier Reef (2012), for organizing and including me in the panpsychism reading group at ANU, and for his participation at the workshop "Panpsychism, Russellian Monism and the Nature of the Physical" in Oslo (2013). These have all been invaluable experiences.

Thanks to the Faculty of Humanities for their generous funding, and the Department of Philosophy, Classics and History of Arts and Ideas for having been a wonderful place to work.

Very special thanks to Philip Goff for extensive and thoughtful comments on this thesis as well as many talks and drafts on the same material, for many discussions from which I have learned and benefitted immensely, and for inviting me to visit ANU. Special thanks also to Sam Coleman, for valuable comments and discussions, and for giving me the opportunity to present a paper at the conference "The Metaphysics and Ontology of Phenomenal Qualities", which was very helpful for developing the ideas of this thesis.

Many thanks and my greatest appreciation go to my fellow Ph.D. students (one a post-doc by now), the three Kantians, Toni Kannisto, Jonas Jervell Indregard and Jacob Lautrup Kristensen, for extensive and patient comments on the first draft of this thesis, for countless fun and interesting discussions and many good times over the course of the last three years – and for teaching me so much about the great Kant.

Many of my other fellow Ph.D. students also deserve thanks. In particular, I would like to mention Ole Martin Moen, Lars Christie, Veli-Pekka Parkkinen, Monica Roland, Jorid Moen and Jola Feix, who have contributed with valuable comments on my work as well as much appreciated chats and coffee breaks. Among other colleagues, I would

especially like to thank Anders Strand, for many thoughtful pieces of advice, and Carsten Hansen, for his discerning comments on my work on many occasions. I would also like to thank all of the other students or researchers with an interest in panpsychism whom I've had the privilege to meet and discuss with, in particular, Brentyn Ramm, Benjamin Andrae, Jonathan Simon and Luke Roelofs, for sharing their insights with me.

I'm very grateful to all the speakers and participants at the workshop "Panpsychism, Russellian Monism and the Nature of the Physical", of which I was the organizer, for creating an incredibly interesting and inspiring event, and for providing helpful and encouraging feedback on some of the material from this thesis, which I presented there. Let me also thank all the participants at the workshop at the Great Barrier Reef, on panpsychism and the combination problem, for their comments and suggestions on what have become chapters 5 and 6 of this thesis.

Unfortunately, I can't mention all the kind and brilliant people whom I've met and learned from along the way. I'm grateful to every one of you, and happy to have found that the world of academic philosophy is such a friendly place.

Finally, I gratefully acknowledge the invaluable support of my friends and family, a few of whom I would like to mention especially: my friend Victoria Kielland, for always understanding and being on my side; my aunt Grete Hassel, for providing a fantastic writing retreat at her farm in Sigdal, where large parts of this thesis were written; and my parents, Åse Hassel and Morten Mørch, for always supporting me in every way.

Hedda Hassel Mørch, December 2013

TABLE OF CONTENTS

1

Introduction

Panpsychism is the view that everything is mental – or more precisely, that every concrete and properly unified thing has at least a fundamental form of phenomenal consciousness or experience. It is an age-old doctrine, which, to the surprise of many, has recently taken on new life in philosophy of mind and metaphysics. In philosophy of mind, it has been put forth as a simple and radical solution to the mind–body problem, or the problem of finding a place for consciousness in the physical world. In metaphysics and philosophy of science, it has been put forth as a solution to the problem of accounting for the nature of the physical itself. Philosophers who have recently defended[1] panpsychism in one or both of these ways notably include Galen Strawson (2006b), David Chalmers (1996, 2003, forthcoming-b) and Thomas Nagel (1979, 2012).

In this thesis, I aim to show that panpsychism is also importantly related to the metaphysics of causation, in two main ways. Firstly, panpsychism is a consequence of an argument from our (arguable) acquaintance with the nature of causation in agency. This is an argument that has appeared within many traditions in the history of philosophy, but nowadays is mostly forgotten, and I think unjustifiably so. Secondly, causation is important insofar as mental combination, the manner in which complex, i.e., human and animal-type, consciousness results from simple, i.e., fundamental particle-type, consciousness according to panpsychism can be construed as a causal process. This is opposed to accounts of combination as a matter of either constitution or radical emergence, which have been held to exhaust the alternatives, and I think incorrectly so.

Furthermore, I aim to show that this would help panpsychism avoid two serious challenges. Firstly, it avoids a challenge to the line of argument from metaphysics and philosophy of science, according to which a metaphysics of primitive causal relations, powers or dispositions can account for the nature of the physical at least as well as, and perhaps much better than, panpsychism. Secondly, it would help avoid a challenge to the line of argument from philosophy of mind, known as the combination problem. This is

[1] But not necessarily endorsed.

the problem of accounting for complex consciousness within panpsychism without facing problems analogous to those of physicalism, dualism and other competing views – problems which according the line of argument from philosophy of mind panpsychism was supposed to dissolve all traces of.

In this introduction, I will elaborate and discuss all of these claims and aims. I will begin with the definition of panpsychism. Then I will present and analyze the two main arguments for the view, followed by their respective problems. Finally, I will give an overview of the structure of the thesis, with an outline of the two new arguments and how they are able to solve the problems of the two old ones.

1 THE DEFINITION OF PANPSYCHISM

What does it mean to say that everything, or every concrete and proper thing, is mental or involves at least a fundamental form of experience? There are some aspects of the doctrine which most of its contemporary defenders agree about, and some which are more contested. In this thesis, I will define panpsychism so as to include most of what I take to be the generally agreed upon features, and to leave most of the contested questions open. I will take the definition word by word,[2] and explain what I think goes where.

1.1 Everything

What do defenders of panpsychism normally mean when they say that *everything* is mental? It seems generally agreed upon that the "pan" of "panpsychism" requires that mentality is to be attributed to at least every fundamental and concrete thing, in addition to humans and other animals. Being concrete means being non-abstract, perhaps in virtue of being spatiotemporal, so numbers and other abstract objects are excluded from the thesis. The fundamental concrete entities are often taken to include at least the ultimate particles of physics,[3] but to exclude most ordinary objects like tables, chairs and rocks. Therefore, panpsychism does not require that such ordinary objects have mentality (as

[2] Here I follow the example of Daniel Stoljar (2009), who discusses the definition of physicalism by a similar procedure.

[3] Some, however, regard the whole universe as fundamental, and in combination with panpsychism this would yield the thesis of *cosmopsychism*, i.e., the thesis that mentality should be attributed to the universe as a whole. This variety of panpsychism is much less discussed than varieties that regard non-cosmic objects as fundamental. Why is this? Perhaps it is because science seems to treat particles and other non-cosmic objects as fundamental (although one sometimes hears about quantum holism), or perhaps it is because of cosmopsychism's theistic associations, together with the fact that it does not seem to offer any theoretical advantages over other varieties of panpsychism. In any case, I will mostly set cosmopsychism aside.

emphasized by Strawson 2006b: 26). The same goes for more esoteric objects sometimes considered by philosophers, such as undetached rabbit parts or the set of my nose and the planet Venus (however, see Goff (forthcoming) for an argument to the contrary). Such presumably non-fundamental things can be regarded as mental only in virtue of having mental parts or constituents, i.e., in the same indirect way that we ordinary think of a society of people as having mentality.[4]

However, panpsychism does not prohibit attributing mentality directly to some non-fundamental objects either. Many hold the view that organisms or brains, that we know (sometimes) have mentality, are not fundamental things, and if so, it could well be that some other non-fundamental things, such as cells, molecules, and so on, have mentality is well. It could also be argued that these kinds of things should be regarded as fundamental, and as mental in virtue of that. Whether any other things, in addition to fundamental particles and animals (including humans), have mentality, and if so which things and by what principle they come to possess it, is an open question. Accordingly, I take panpsychism to entail that mentality is to be attributed to all things that are, firstly, concrete, and secondly, either fundamental or otherwise properly unified according to a principle yet to be determined.

1.2 Mental

What does it mean to say that everything is *mental*? What is the "psyche" of "panpsychism"? As I will explain in section 2.1 below, much of the appeal of panpsychism is that it seems to straightforwardly dissolve the mind–body problem (or at least some of its most important aspects) by positing as fundamental (and ubiquitous) those mental features which physicalists have most difficulty with reducing or explaining, namely phenomenal consciousness or experience. This feature of mentality is what gives rise to what Chalmers has called the hard problem of consciousness, as opposed to the easy problems concerning the behavior and cognitive functions associated with consciousness.[5] If something has phenomenal consciousness or experience it means that

[4] Cosmopsychists could regard tables, chairs and rocks as mental merely in virtue of being part of a larger mental whole, which is how we could perhaps think that a neuron of a waking brain involves mentality.

[5] "The hard problem of consciousness […] is that of explaining how and why physical processes give rise to *phenomenal* consciousness" (Chalmers 2010: 105, my emphasis). The (comparatively) easy problems of consciousness as the problems of explaining the behavior and cognitive functions associated with consciousness, such as "the ability to discriminate stimuli, or to report information, or to monitor internal states, or to control behavior" (Chalmers 2003: 103). Concerning the easy problems of consciousness in the

there is something that it is like to be that thing – in the phrase introduced by Nagel (1974). One could also describe it as having states with a qualitative feel, or qualia (Chalmers 1996: 4), or as having an inner life or feeling broadly construed. I will use the terms phenomenal consciousness, phenomenal qualities, qualia, experience and phenomenology, and the associated adjectives conscious, experiential and phenomenal, interchangeably to point to this aspect of mentality, and use mentality such as to include this aspect unless otherwise implied. I will take panpsychism to entail that everything has mentality in this sense.

Some philosophers think that in order to dissolve the mind–body problem one does not have to say that there is something that it is like to be every (fundamental or unified) thing; rather, it may suffice to attribute protomental (Chalmers 1996, forthcoming-b) or neutral (Nagel 1986; Coleman 2013b) properties to them. However, as will be discussed below, it is controversial whether such properties can serve the same explanatory purpose as full-blown experientiality. What is less controversial is that it is useful to reserve the term panpsychism for views that posit full-blown experience as ubiquitous, while panprotopsychism and neutral monism should denote views that posit protomental or neutral properties as ubiquitous.

A substantive question related to protomentality is whether phenomenal qualities[6] always come with a subject of experience, or whether they can exist unexperienced or without a subject. Strawson argues that it is a necessary truth that there is always a subject of experience wherever there is experience (Strawson 2006b: 26; 2008b). Others, such as Sam Coleman (2012) and Gregg Rosenberg (2004), posit unexperienced qualities.[7] I will count unexperienced phenomenal qualities, insofar as they are possible, as protomental or

behavioral and functional sense, he argues that: "There is no real issue about whether these phenomena can be explained scientifically. All of them are straightforwardly vulnerable to explanation in terms of computational or neural mechanisms" (Chalmers 2010: 4). The hard problem of phenomenal consciousness or experience, is different: "To explain experience, we need a new approach. The usual explanatory methods of cognitive science and neuroscience do not suffice. These methods have been developed precisely to explain the performance of cognitive functions, and they do a good job of it. Still, as these methods stand, they are equipped to explain only the performance of functions. When it comes to the hard problem, the standard approach has nothing to say" (Chalmers 2010: 9).

[6] I said I would use phenomenal qualities interchangeably with experience, but experiences without a subject of experience sounds like a contradiction in terms, while phenomenal qualities without a subject does not (at least not as clearly), so there is a difference in meaning which I acknowledge. However, since the phenomenal qualities posited by panpsychism as I define it will always be experienced phenomenal qualities, the terms will be extensionally equivalent in this context, so this will not be a problem.

[7] Rosenberg posits them as belonging only to the realm of mere possibilia. Coleman posits them as actual, but categorizes them as neutral, not mental.

neutral, and take panpsychism to entail that everything involves subjective experience. The definition will leave entirely open the nature of subjects and whether phenomenal qualities are necessarily experienced and if so why (but I will argue for a position on both points in the course of the thesis). In saying that experiences are always had by subjects, we can elaborate on what it means to *have* mentality – directly, as opposed to indirectly in virtue of mental constituents. Having mentality means being a subject of experience. Tables and chairs, if indeed they do not have mentality except indirectly, are constituted by things which are subjects of experience, but without being subjects of experience themselves.

Panpsychists say that there is *something* that it is like to be a fundamental physical entity such as an electron, but do they also say something about *what* it is like? Historically, many panpsychists have had definite views about this. Arthur Schopenhauer claimed that groundless will constituted fundamental reality, and Empedocles claimed that everything was governed by love and strife – they seem to respectively posit volitional phenomenology and emotional phenomenology as fundamental.[8] In recent times, it seems most defenders of panpsychism have been non-committal on this issue,[9] although sensory phenomenology, such as color qualia, often figures in thought examples. I will not take panpsychism to entail anything about the character of fundamental experience – although, in the course of the thesis, I will argue for a view about this.

An objection which is frequently heard is that to say that fundamental entities have fundamental experiences is a borderline category mistake, because, for example, the kind of experience we can positively conceive of is of a kind which would be incoherent to attribute to things like electrons, because they are too physically and behaviorally simple, or have none of the features we ordinarily take to be reasons for attributing mentality.[10]

One response could be to say that we can vaguely, but positively, conceive of experience which is maximally simple and perhaps minimally rich compared to ours. Why would it be incoherent to attribute simple experience to simple things? Perhaps because it seems explanatorily redundant. For example, humans and animals perceive and

[8] There is always controversy about whether historical figures who seem to be panpsychists are really panpsychists. In chapter 2 (section 2), I will point to evidence that Schopenhauer was in fact a panpsychist. David Skrbina (2005: 23–33) claims that Empedocles was.

[9] Lewtas (2013) is an exception: he argues that quark experiences are simple and homogenous fields of the basic types of qualities or raw feels that we ourselves can experience.

[10] Thanks to Camilla Serck-Hanssen and Einar Duenger Bøhn for this objection.

react to their environment in accordance with purposes, while particles do not, and as Catherine Wilson (2006a: 181) objects to panpsychism, why would consciousness be of any use to them? In response, one could say that we cannot really complain that consciousness does not play a causal or explanatory role in the behavior of particles, because we cannot say, on the basis of the debate in philosophy of mind, what causal or explanatory role consciousness plays in our own behavior either – this is the problem of mental causation, and some think epiphenomenalism is an adequate response to it. We can assume that the simple behavior of particles could in principle be explained with reference to their simple mental states, if we could know them, in the same way that complex human behavior can be explained with reference to our complex mental states.

One might also doubt whether we can really positively conceive of simple experiences or whether we are just stipulating them without any positive conception. Another response, then, is to grant that perhaps there are no experiences that we can positively conceive of that are appropriate to attribute to an electron. Still, one should be able to claim that there is something that it is like to be an electron without having any idea about what it is like. Some might doubt whether it is meaningful at all to attribute experiences without any determinate specific character. In response, it can be pointed out how this worry it at least not widely shared in philosophy of mind. This is clearly illustrated by Nagel's canonical question about what it is like to be a bat, part of the point of which is precisely that we *can* meaningfully consider that there is something that it is like to be a bat, without having any positive conception of what it is like. Someone might reply that in fact we do implicitly have some idea about what it is like to be a bat in general, in the sense that bats must perhaps have desires, beliefs, pain or such things; we only lack a conception of the particular qualia associated with sonar navigation. My response would be that, firstly, we have no determinate idea about what it is like for a bat to, say, have desires, or what bat pain is like. Furthermore, even if in the bat case we do have some indeterminate, but positive, idea about what it is like to be a bat, the fact that we can, relatively unproblematically, meaningfully consider very indeterminately conceivable experiences is a starting point for clarifying or introducing the concept of a wholly unspecific or indeterminate (for us, not for its subject) experience. It is that which lies at the limit of the kind of abstraction process we perform when considering bat experience. I think there are many concepts in philosophy which are specified in an equally abstract way, and which are not thereby dismissed as meaningless.

1.3 Is

Finally, what does it mean to say that everything *is* mental? Is it an "is" of predication or identity? And what remains of the physical when everything is mental?

The kind of panpsychism which, as I will explain below, is compatible with the important argument from physical causal closure in philosophy of mind, and is able to answer questions about the nature of the physical according to the arguments from metaphysics and philosophy of science (these arguments will be discussed in sections 2.1 and 2.2) is a form of Russellian monism. Russellian monist panpsychism is the view that the phenomenal somehow "fills a gap in the picture of nature painted by physics" (Alter and Nagasawa 2012: 67). Russellian monism may also be non-panpsychist and assert that protophenomenal, neutral or unknown properties fill the gap. What is this gap? According to Bertrand Russell, in *The Analysis of Matter* (1927), physics only tells us about the abstract, purely relational structure of reality, and nothing about what realizes the structure or about the intrinsic nature of the entities that stand in the relations. Russell also takes it that realizers of the structure or relata with intrinsic properties nevertheless must exist, and argues that nothing rules out that these are mental properties:

> Physics is mathematical, not because we know so much about the physical world, but because we know so little: it is only its mathematical properties we can discover. For the rest, our knowledge is negative […] The physical world is only known as regards certain abstract facts about its space-time structure – features which, because of their abstractness, do not suffice to show whether the physical world is, or is not, different in intrinsic character from the world of mind. (Russell 1948/1992: 240)

Russellian monism is not a kind of property dualism where everything has both mental and physical properties contingently; rather, the mental and the physical are complementary properties or aspects and, therefore, to some extent necessarily or intelligibly connected. There are many, not necessarily mutually exclusive, alternatives for how the physical and the mental may complement each other – for how to specifically understand character of the gap and how the mental fills it, and for how the physical is to be understood. Mental properties have been claimed to be the categorical grounds of physical dispositions, the concrete realizers of physical structure and the intrinsic relata of physical relations (Alter and Nagasawa 2012: 72 mention all these alternatives).

Sometimes the physical and the mental are described as two aspects of the same reality.[11] Chalmers has suggested that "information (in the actual world) has two aspects, a physical and a phenomenal aspect" (Chalmers 1996: 286) and that perhaps "wherever there is information, there is experience" (Chalmers 1996: 297).[12] Strawson suggests that all experiences have an inside as well as an outside, where the outside is a matter of how the experience interacts with other experiences (Strawson 2006a: 255–256).

Most of these relations suggest that the mental is more fundamental than the physical. Grounds and realizers are prior to what they ground or realize and relata are often taken to be prior to relations. It is arguable, especially for some kinds of relations, that they are not only posterior to but really nothing over and above their relata. This suggests that panpsychism might entail or at least be compatible with idealism. Strawson defines *pure* panpsychism as the view that "all being is experiential being" (Strawson 2006a: 222), which entails that there is no physical, in the sense of non-experiential, being. He claims that this is not idealism in the conventional sense:

> I avoid the word "idealism" because conventional idealism – the claim that reality consists entirely of ideas or experiences – is blatantly incoherent […] in assuming [a] that the subject of experience is in some way ontologically over and above its experiences and [b] that the subject of experience is not itself a mere idea. (Strawson 2006a: 229, footnote 95).

The experiences of pure panpsychism "[…] must not for a moment be conceived as some sort of 'mere experiential content', where this is in some way passively conceived" (Strawson 2006a: 243). I will take panpsychism to be compatible with pure panpsychism, according to which the physical is nothing over and above the mental, but incompatible with conventional idealism regarded as the view that the physical is nothing over and above mere mental *content*, in a Berkeleyan sense (where this refers to the popular conception of Berkeley). Panpsychism mainly requires that everything physical *has*

[11] The term aspect sometimes indicates properties which are somehow observer or point-of-view dependent. However, since it seems it will also depend on the properties of a thing what aspects that are available to an observer, the distinction can be taken as a difference in connotation, not literal meaning.

[12] It is not obvious that this view is a kind of Russellian monism, and Chalmers does not present it as such, but it is not obvious that a dual-aspect view of information could not be Russellian either.

mental content – everything is not mental in sense of being the contents of someone else's experience, but in the sense of itself having experience or being subjects of experience.[13]

Is panpsychism also compatible with saying that the physical irreducibly exists and is co-fundamental with the mental – i.e., can it also be impure? Strawson argues that this is hard to make sense of, because it seems like a contradiction in terms to say that the same thing is both experiential and non-experiential (Strawson 2006a: 234–238). That is, it sounds like one is asserting that the same thing both has and does not have the property of experientiality. This argument, however, depends on Strawson's metaphysics of properties (Strawson 2006a: 239) and is not supposed to follow directly from the definition of panpsychism. Ordinarily, we see no problem in one thing being both, say, blue and cold, even though the fact that coldness is not the same property as blueness entails that the thing would thereby be both blue and non-blue. As long as the negative property non-blueness is constituted by a positive property, coldness, as opposed to the pure absence of blueness, a thing being both blue and non-blue is not contradictory. Correspondingly, as long as the non-experiential and physical is not defined wholly negatively as absence of fundamental mentality, but is rather constituted by a positive property, it seems logically possible that everything has both fundamentally mental and fundamentally physical properties.

But what could these positive, non-experiential, physical properties be? The physical cannot be regarded simply as the kind of properties physical sciences talk about, because according to Russellian monism, the physical sciences tell us only about structure, relations or dispositions, which seem dependent on realizers, relata or grounds and therefore not fundamental. However, we might understand structure not merely abstractly. Some think that dispositions are not merely relations between categorical relata, but are somehow irreducibly causal (a view I will discuss in section 3.2 below). Many would also hold that science describes an irreducibly spatiotemporal structure. If causation and space-time are regarded as having concrete reality, then it is not obvious that this reality must be construed as reducible to or grounded in the fundamentally mental. Therefore, it seems that causation, spatiality and temporality are candidates for fundamental physical properties of an impure panpsychism. There might be other candidates too. Accordingly, I

[13] According to Berkeleyan idealism, inanimate objects only exist when perceived (by us or by God). For panpsychists, objects keep existing when not perceived from the outside, but in a sense, this is because being subjects, they will continue to perceive (or experience) themselves.

will not take panpsychism to entail that the physical is reducible to or nothing over and above the mental. However, in the course of the thesis, I will argue that causation can and should actually be understood as being grounded in mental properties.

According to Russellian monism, either the physical consist of the structural or relational aspects or properties of the mental, or the mental and the physical are two aspects or properties of the same structure. In either case, the mental and the physical have structure in common, and this is part of what explains their connection. The structure of reality as revealed via the physical "from the outside" aspect, or by observing its physical properties via the physical sciences, must therefore match the structure as revealed from the mental "from the inside" aspect, or by observing its mental properties via introspection or phenomenological investigation. Russellian monist panpsychism therefore entails that the physical and the mental are either structurally isomorphic, i.e., perfectly matching, or at least fully compatible, as the full structure of reality might not be discernible from both aspects.

2 THE ARGUMENTS FOR PANPSYCHISM

I will now present and discuss the most important and influential arguments for panpsychism in the recent literature and identify some of their central presuppositions.

First a note about terminology: I will use the term physicalism in roughly the way Strawson uses physicSalism (2006b: 4), Chalmers uses narrow physicalism (forthcoming-b: 9–10) and Daniel Stoljar uses t-physicalism, based on what he calls the theory-based conception of the physical (2001: 256). According to Strawson, physicSalism is "the view – the faith – that the nature or essence of all concrete reality can in principle be fully captured in the terms of physics." (Strawson 2006b: 4). However, I will make one modification: I will use the term physicalism to mean the view that everything, or at least the mental, can in principle be accounted for in terms of a completed physical theory, either physics or physics together with a set of physical sciences. This is because, as I will discuss in chapter 5 (section 3), it may turn out that not all sciences, such as chemistry or biology, are in principle reducible to physics, and if so, it seems they should still be regarded as physical sciences.

Strawson uses physicalism or real physicalism (2006b: 4), Chalmers uses broad physicalism (forthcoming-b: 9–10), and Stoljar uses o-physicalism, based on what he calls the object-based conception of the physical (2001: 257), for the view that everything is physical, but where the meaning of the term physical is not defined in terms of the

properties revealed by physics or physical theory. Rather, the physical is defined by pointing to a paradigmatic physical object and referring rigidly to whatever stuff it is fundamentally made of, whose nature might not be fully captured by physical science. As Strawson puts it: "I simply fix the reference of the term 'physical' by pointing at certain items and invoking the notion of a general kind of stuff" (Strawson 2006b: 8). Real, broad or o-physicalism is compatible with panpsychism, and according to Strawson (2006b), it entails it.

Physicalism, as I will use the term (unless otherwise noted), is incompatible with panpsychism. This follows from (1) taking panpsychism as a view according to which mental properties are fundamental, (2) defining "psyche" or the mental in terms of *what it is like*-ness or experientiality, as I have done here,[14] and (3) the thesis that *what it is like*-ness or experientiality will not be fundamental terms of the completed set of physical sciences.

The latter thesis is widely accepted by both physicalists and non-physicalists, but it is justified in different ways. Some stipulate that physical properties are fundamentally non-mental (Wilson 2006b; Papineau 2001), partly in order to capture what is fundamentally at stake in debates between those who tend to call themselves physicalist and non-physicalists about consciousness, namely whether or not the mental is fundamental. It would follow from this stipulation that any science that ends up positing mental properties as fundamental will not be a physical science. Others reject the stipulation, but still support the thesis on the basis that it seems highly unlikely, based on that present science looks like, that future physical science will end up positing mental properties as fundamental (Dowell 2006). Others again stipulate that (narrowly) physical properties are structural (Chalmers forthcoming-b: 10), where structural properties may include dispositional and spatiotemporal properties as well as abstract logico-mathematical properties (as per the Russellian view of the physical). If mental properties are non-structural (as is supported by many argument soon to be discussed below), then it follows that physical properties are non-mental, and subsequently that any science that posits

[14] Russellian monist panpsychism entails (1), that mental properties are fundamental, since they are the grounds of the physical. If one rejects (1) or (2), (narrow/t-) physic(S)alism can be compatible with panpsychism. For example, if one has a reductive functionalist and computational conception of the mental (which is incompatible with both (1) and (2)) and if one thinks, e.g., that computational functions are abstract and observer-relative so that everything can be interpreted to realize any kind of computational function (Searle 1984), then physicalist (t-/narrow/physicSalist) panpsychism would be the result.

mental properties as fundamental will not be a physical science. One might justify this stipulation by arguing that the methods of science are in principle unable to reveal non-structural properties, or that they are unlikely to do so. Arguments to this effect will be discussed in section 2.2 below. I will assume the thesis that *what it is like*-ness or experientiality will not be fundamental terms of the completed set of physical sciences, but remain neutral for now on the justification.

2.1 The Argument from Philosophy of Mind

Chalmers claims the best reason for accepting panpsychism is that it is compatible with both the conceivability argument against physicalism and the argument from physical causal closure against dualism, two of the most influential and powerful arguments in philosophy of mind:

> I call my argument the Hegelian argument for panpsychism [because it] takes the dialectical form often attributed to Hegel: the form of thesis, antithesis, synthesis. […]
>
> In my Hegelian argument, the thesis is materialism, the antithesis is dualism, and the synthesis is panpsychism. Or at the level of arguments: the thesis is the causal argument for materialism (and against dualism), the antithesis is the conceivability argument for dualism (and against materialism), and the synthesis is the Hegelian argument for panpsychism. In effect, the argument presents the two most powerful arguments for and against materialism and dualism, and presents a certain sort of panpsychism as a view that captures the virtues of both and the vices of neither. (Chalmers forthcoming-b: 2)

Torin Alter and Yujin Nagasawa make the same kind of point: "It can be argued that Russellian monism retains strengths of traditional versions of physicalism and dualism, while overcoming their weaknesses" (2012: 87). Note that Russellian monism is not necessarily panpsychism, and I will return to the advantages panpsychism arguably has over other kinds of Russellian monism later.

What are the main weaknesses of physicalism and dualism respectively? I will begin with physicalism. Chalmers mentions the conceivability argument, varieties of which have been put forth by himself (e.g., Chalmers 1996) and Kripke (1980). Together with the knowledge argument (Jackson 1986) and the explanatory argument (from the presence of an explanatory gap (Levine 1983) or a hard problem (Chalmers 1995)) they constitute what is often regarded as the main anti-physicalist arguments. According to Chalmers, these arguments all have in common that they point out an *epistemic gap* between the mental and the physical (2003: 107). They claim that the epistemic gap is not closable in

principle – no amount of physical information would make zombies inconceivable, enable Mary to deduce what it is like to see red, or close the explanatory gap or solve the hard problem. Finally, they suppose that epistemic gaps which are not closable in principle must be explained by an *ontological gap*; by the properties or entities in question in fact being distinct. This refutes physicalism, understood as the view that asserts that mental properties are identical to, realized by or constituted by the physical or in some other way nothing over and above it.

Physicalists who accept that epistemic gaps that are not closable in principle entail ontological gaps could deny that there is an epistemic gap between the mental and the physical. Chalmers calls this view type-A materialism or physicalism.[15] He claims that "the obvious problem with type-A materialism is that it appears to deny the manifest" (2003: 109), and that this is highly counterintuitive. Strawson similarly argues (in effect, he does not use the term epistemic gap) that the idea that:

> […] all characteristics of what is going on, in the case of experience, can be described by physics and neurophysiology or any non-revolutionary extensions of them […] amounts to radical 'eliminativism' with respect to experience […] (Strawson 2006b: 7)

Eliminativism he regards as "the strangest thing that has ever happened in the whole history of human thought, not just the whole history of philosophy" (Strawson 2006b: 5). Type-A physicalism includes views which are not explicitly eliminativist, such as analytic functionalism. However, in saying that the mental exists, but is exhausted by functional or behavioral properties, the phenomenal aspect of the mental is arguably implicitly eliminated (in section 2.2 below I will discuss arguments that the phenomenal is not purely functional or otherwise relational).

Type-A physicalists could acknowledge that phenomenal properties, or mental properties that are not exhausted by a functional or behavioral description, certainly *seem* to exist, but claim that this is an illusion. Against this, Strawson makes an appeal to Cartesian certainty, claiming that the existence of the phenomenal is more certain than anything else (2006b: 3). Similarly, John Searle has argued that dismissing the phenomenal as mere illusory appearance is self-defeating, because "where consciousness is concerned, the existence of the appearance is the reality" (Searle, Dennett, and Chalmers 1997: 112). Chalmers is not as absolute – he claims that highly counterintuitive

[15] Chalmers uses materialism and physicalism interchangeably.

claims must be supported by extremely strong arguments (2003: 110), and argues that the arguments for the denial of the phenomenal fail to reach this standard.

Physicalists who accept that epistemic gaps that are not closable in principle entail ontological gaps could also accept the epistemic gap, but claim that it is closable in principle, for example, when physics and neuroscience have developed further. However, if physical properties are stipulated to be structural, then we can already tell that they cannot give a sufficient basis for consciousness. This is because, as Chalmers argues (2003: 120–122), the anti-physicalist arguments mentioned above can all be seen to demonstrate that there will always be an epistemic gap between *any* structural properties and phenomenal properties.

Some physicalists would not agree that physical properties are necessarily structural. But it could still be argued that the epistemic gap would remain as long as our understanding of what it means to be physical, however the notion is to be precisely explicated, does not undergo a complete revolution. It seems no conceivable configuration of properties that are continuous with properties that are fundamental in any current physical theory is such that it cannot conceivably exist in the absence of phenomenal properties, or such that phenomenal properties would be deducible from them.

If physicalists claim that the mental is physical only in a sense of "physical" that we can grasp after a complete revolution in science, physicalism will end up no longer distinguishable from competing views. We cannot rule out that after a revolution, mental properties will themselves end up as fundamental, which would make physicalism compatible with panpsychism and dualism. Neither can we rule out that the kind of non-mental properties that non-panpsychist Russellian monists currently posit (to be discussed below), but which are currently not part of any physical scientific theory, end up as fundamental, and so physicalism (i.e., narrow/t-/physicSalism) also becomes compatible with all kinds of Russellian monism (i.e., broad/o-/real physicalism). In other words, physicalism becomes an almost trivial thesis. To avoid this, physicalists could stipulate that fundamentally mental (as Papineau and Wilson do) and Russellian properties will not qualify as physical. This makes physicalism less trivial, but in return, less justified, because what reason would there be for thinking that physicalism in this sense will survive a revolution which is radical enough to let the epistemic gap be closed?

Finally, physicalists can deny that an epistemic gap that is not closable in principle entails an ontological gap. This is the view of type-B physicalists (in Chalmers'

terminology) or *a posteriori* physicalists, such as Brian Loar (1997), Christopher Hill (1997) and David Papineau (2002). On this view, mental properties are identical[16] with physical properties, but this identity can neither be established nor made sense of on the basis of *a priori* philosophical reasoning (hence the epistemic gap). Rather, the identity must be accepted as a brute empirical fact, much like the identity between water and H_2O.

Chalmers points out some important disanalogies between the alleged mind–brain identity and any identities found elsewhere in nature, such as between water and H_2O, and DNA and genes. Other identities are deducible from the complete physical truth about the world, while the mind–brain identity is not. Furthermore, he points out that:

> Elsewhere, the only sort of place that one finds this sort of primitive principle is in the fundamental laws of physics. Indeed, it is often held that this sort of primitiveness – the inability to be deduced from more basic principles – is the mark of a fundamental law of nature. (Chalmers 2003: 113)

Because laws of nature relate distinct entities or properties, if the primitiveness associated with them also characterizes the relation between the mental and the physical, then this would, by standard scientific reasoning, indicate that the mental and the physical are distinct, which in turn indicates the falsity of physicalism, Chalmers concludes that the primitive and inexplicable identities needed in type-B physicalism seem to be suggested "largely in order to preserve a prior commitment to materialism. Unless there is an independent case for primitive identities, the suggestion will seem at best ad hoc and mysterious, and at worst incoherent" (Chalmers 2003: 113). Additionally, Chalmers gives an elaborate argument based on the theory of two-dimensional semantics (2003: 115–119; 2009), which is too complicated to go into here.

One popular response to this kind of criticism is the phenomenal concept strategy, pursued by Papineau (2002) and Loar (1997). This strategy consists in arguing that the fact that epistemic gap between the mental and the physical is not closable in principle can be explained by theories according to which our phenomenal concepts work in a peculiar way: they do not actually reveal any information about the nature of their referents, even though they robustly appear to do so. As an objection to the this strategy, Philip Goff, who has also defended panpsychism, shows how it entails that thinking of

[16] Type-B physicalists normally claim that the physical and the mental are identical. One could appeal to other kinds of necessitation, but most of the arguments discussed here seem adaptable to such versions.

phenomenal experiences such as pain in terms of how it feels reveals nothing about what it is to feel pain – which he argues is highly implausible (Goff 2011). Similarly, Strawson puts forth what he calls The Partial Revelation Thesis: "In the case of any particular experience, I am acquainted with the essential nature of the experience in certain respects, at least, just in having it" (2006a: 252). According to this thesis, not only can we know with certainty that experience, or phenomenal properties, exist, in the sense we know that phenomenal concepts succeed in referring to *something;* we also know something about the nature of what we refer to through phenomenal concepts. The nature of phenomenal properties appears to consist in how they feel, and hence be such that they cannot be even *a posteriori* identified with any physical properties. The Partial Revelation Thesis, which justifies trusting this appearance, would therefore seem to be a sufficient, and perhaps also necessary, basis for rejecting type-B physicalism, at least the kind that depends on the phenomenal concepts strategy.

Physicalism can also be understood to include the view that the mental is distinct from, but metaphysically necessitated by, the physical. If so, it is in principle compatible with an ontological gap. One problem with this view, however, is that it might seem compatible with dualism: necessitation is arguably not sufficient for physicalism (Stoljar 2010: ch. 8). I will refer to this view simply as emergentism, and leave it open whether it should be regarded as a kind of physicalism, a kind of dualism, or as compatible with both.

Strawson argues, against emergentism and in defense of panpsychism, that there is no intelligible way in which the mental can depend on the physical, given that we fully accept the reality of experience and also hold that the physical is fundamentally non-experiential. He argues that the experiential cannot emerge from the non-experiential in the intelligible way in which phenomena like liquidity emerge from a configuration of non-liquid constituents. An analogy of the right size would be the idea that the extended might emerge from non-extended points (Strawson 2006b: 16), but this is a destructive analogy, because this kind of emergence is either impossible or not really emergence at all:

> If it really is true that Y is emergent from X then it must be the case that Y is in some sense wholly dependent on X and X alone, so that all features of Y trace intelligibly back to X (where 'intelligible' is a metaphysical rather than an epistemic notion). *Emergence can't be brute.* It is built into the heart of the notion of emergence that emergence cannot be brute in the sense of there being absolutely no reason in the nature of things why the

> emerging thing is as it is (so that it is unintelligible even to God). For any feature Y of anything that is correctly considered to be emergent from X, there must be something about X and X alone in virtue of which Y emerges, and which is sufficient for Y. (Strawson 2006b: 18, emphasis original)

To further emphasize the unacceptability of brute emergence, Strawson claims that:

> One problem is that brute emergence is by definition a miracle every time it occurs, for it is true by hypothesis that in brute emergence there is absolutely nothing about X, the emerged-from, in virtue of which Y, the emerger, emerges from it. (Strawson 2006b: 18)

> If emergence can be brute, then it is fully intelligible to suppose that non-physical soul-stuff can arise out of physical stuff – in which case we can't rule out the possibility of Cartesian egos even if we are physicalists. I'm not even sure we can rule out the possibility of a negative number emerging from the addition of certain positive numbers. (Strawson 2006b: 19)

Similarly, Nagel argues, in a tentative defense of panpsychism, that "there are no truly emergent properties of complex systems" (Nagel 1979: 182).

These problems of physicalism can all be classified as *problems of intelligibility*. According to the arguments, physicalism is committed to the world in some way being fundamentally unintelligible – by containing miraculous brute emergence (insofar as this is compatible with physicalism at all), by containing primitive, inexplicable identities or constitutive relations between properties that robustly appear distinct (i.e., separated by an epistemic gap), or by being such that we are completely ignorant about the nature of our own consciousness and possibly wrong in thinking that it even exists, a proposal many judge to be *a priori* self-defeating.

The main problem of dualism, on the other hand, can be classified as an *empirical problem*. By this I mean a problem of fitting the mental into the scientific picture of the world, especially as regards its causal structure, while at the same time preserving theoretical virtues such as elegance and parsimony. These theoretical virtues seem to be part of scientific methodology, and therefore I count having to go against them as an empirical problem.

The empirical problem of dualism is the argument from the causal closure of the physical. It seems intuitively and theoretically reasonable (which is perhaps to put it mildly) that the mental often has physical effects, mainly in agency. However, it is widely held that the physical is causally closed; that is to say, that every physical event (that has a cause) has a sufficient physical cause (Stoljar 2009). I will go into the reasons for accepting this principle in chapter 6 – they are largely based on empirical observations and features of scientific methodology. To say that some types of physical events systematically have a mental cause in addition to a sufficient physical cause, i.e., that some physical events are systematically causally overdetermined, is an inelegant, unparsimonious and *ad hoc* hypothesis. The denial of epiphenomenalism and overdetermination together with acceptance of physical causal closure seem to entail that the mental is physical, which refutes dualism.

Standardly, dualism is regarded as a view according to which mental properties and physical properties (and perhaps also substances) are distinct, but causally related (either mutually, or just one-way, as with epiphenomenalism). Many hold that causal relations are not metaphysically necessary, but depend on contingent laws of nature that could vary between different possible worlds. Others hold that the laws of nature hold with metaphysical necessity; that there are no possible worlds with the same things, but where the laws on nature are different (a view that will be discussed extensively in many chapters below). Given this so-called necessitarian view of causation, there is no clear distinction between emergentism and dualism: both entail that mental properties are metaphysically necessitated by physical ones (and vice versa for non-epiphenomenalism). This would give dualism an intelligibility problem as well as an empirical one, the problem of brute emergence or brute necessitation. But dualism arguably has an intelligibility problem anyway. A traditional objection to dualism, first posed to Descartes by his correspondent Princess Elizabeth, is that interaction between mental and physical substances is unintelligible. The unintelligibility of physical–mental emergence, pressed by Strawson, could perhaps be seen as a further aspect of this problem. Causation and emergence are both productive relations, and it seems that productive relations in general between the mental and the physical can be argued to involve an intelligibility problem.

It should be noted that physicalism also has both kinds of problems. An empirical problem of causally integrating mental properties arguably arises for physicalism even as mental properties are identified with physical properties. This is the exclusion problem, which was introduced by Jaegwon Kim (1989). It results from the fact that physicalism

identifies the mental with higher-level macrophysical properties such as properties of brains or relevant brain areas, properties which are not fundamental in physics. According to the principle of microphysical causal closure, there is a sense in which all causation fundamentally goes on at the microphysical level. It may seem that if microphysical causal closure is true, microphysical constituents will causally exclude the macrophysical wholes they constitute, including mental wholes, rendering them epiphenomenal. The principle of microphysical causal closure is stronger than the principle of mere physical causal closure that the argument against dualism depends on. The principle of microphysical causal closure is widely held, but some philosophers argue that our best evidence only supports the weaker principle of physical causal closure – I will discuss this in chapter 6, section 3.1.

Many philosophers dismiss dualism's intelligibility problem, of the intelligibility of interaction, and physicalism's empirical problem, the exclusion problem, as pseudo-problems. I will come back to their relevance for panpsychism. For now, I will focus on physicalism's intelligibility problem and dualism's empirical problem, which are more broadly recognized and can be regarded as their respective main problems.

Physicalism avoids dualism's empirical problem by identifying the mental with the physical, and dualism avoids physicalism's intelligibility problems by not identifying them. How does panpsychism, as Chalmers, Alter and Nagasawa, Strawson (implicitly[17]) and Nagel (tentatively and implicitly[18]) claim, avoid both?

---

[17] Strawson argues that physicalism finds itself in a dilemma between eliminativism and radical emergentism, where the former is obviously false and the latter is deeply unintelligible, and that panpsychism is the only monist view which lets us escape this dilemma. He dismisses substance dualism as a view for which there has never been a good argument (Strawson 2006b: 25–26), and argues that property dualism is either incoherent or collapses into substance dualism (Strawson 2006b: 28).

[18] Nagel argues that panpsychism *seems* to follow from four plausible premises: "1. *Material composition:* Any living organism, including a human being, is a complex material system. It consists of a huge number of particles combined in a special way. […] No constituents besides matter are needed. 2. *Nonreductionism*: Ordinary mental states like thought, feeling, emotion, sensation, or desire are not physical properties of the organism – behavioral, physiological, or otherwise – and they are not implied by physical properties alone.[footnote omitted] 3. *Realism:* Nevertheless they are properties of the organism, since there is no soul, and they are not properties of nothing at all. 4. *Nonemergence:* There are no truly emergent properties of complex systems […]" (Nagel 1979: 181–182).

Although Nagel does not present it this way, the argument can be made to fit into Chalmers' Hegelian structure. The disadvantage of physicalism is that it seems incompatible with either *Nonreductionism* or *Nonemergence* (as Strawson also argues). The disadvantage of dualism is its incompatibility with *Material composition*. Panpsychism seems compatible with all the premises.

Panpsychism avoids the intelligibility problems of physicalism because it does not identify the mental with, reduce it to, or have it emerge from, properties that can be exhaustively described by physics or the physical sciences. Rather, panpsychism takes the mental to ground these properties in a Russellian way. Panpsychists can accept that there *would* be an ontological gap from the physical to the mental if the physical is taken to be *prior* to the mental, and panpsychism is therefore compatible with all the anti-physicalist arguments that aim to demonstrate such a gap. However, when the mental is conceived of as prior to the physical instead of vice versa, as the categorical grounds of physical dispositions, the realizer of physical structure or the intrinsic relata of physical relations, there will be no epistemic gap, at least not one that seems not closable in principle, and therefore no ontological gap. In a way, this inverts the idea central to functionalism about the mental. A panpsychist would, as discussed above, reject functionalism about the mental because the existence of phenomenal properties that are not functionally exhaustible is obvious or self-evident (I will say more about how this seems obvious in section 2.2 below). But it is not equally obvious that we should resist functionalization of the physical, and as Russell and many others have argued, there is good reason to accept it.

Panpsychism avoids the empirical problem of dualism, the problem of mental causation within causal closure, because it does not secure a causal role for the mental by having it generate its own extra causal relations that will either violate causal closure or redundantly mirror existing causal relations. Rather, the mental fills in physical causal structure, by grounding or realizing it or constituting its relata. By filling in the structure with non-structural properties only, panpsychism leaves the structure itself unchanged. It does not introduce overdetermination, because mental properties are essential to the causal structure and the opposite of redundant. But does the claim that mental properties are essential to all causal relations amount to denying that every physical event has a sufficient physical cause after all? Not in the any sense that is supported by empirical evidence. The evidence for causal closure (which I will discuss in chapter 6, section 3.1) only supports the claim that physical causal structure is independent of any other causal structure, or alternatively, that the physical in the sense that does not exclude a hidden inner nature, i.e., the o-physical or really/broadly physical, is causally closed. In a way, it

should be obvious that empirical evidence[19] cannot rule out that physical events systematically depend on inner natures that are by hypothesis hidden from empirical view, i.e., there is no way to empirically detect their presence or absence.

Now, as mentioned, only pure panpsychism wholly reduces the physical to purely relational or structural properties, properties that can be fully grounded in mental properties so as to completely avoid the problems of physicalism and dualism in these ways. Impure panpsychists, who posit, say, non-mental spatial properties or some non-mental aspect of causation, will not have a purely relational or abstractly functional account of the physical. But although this makes it impossible that mental properties wholly constitute or realize physical properties, one can still see how the two sets of properties systematically complement and depend on each other. Clearly, space and things in space seem to intelligibly depend on each other, and so do causal powers and the things they belong to and affect. Therefore, the epistemic gap between the mental and the physical is still lessened, and the two kinds of properties are not kinds that can causally compete and render each other redundant, so the problem of causal closure is still avoided.

In this way, it seems that panpsychism is able to intelligibly locate and causally integrate the mental in nature, while avoiding the main problems of physicalism and dualism. But the argument for the view is not complete until it is also shown that panpsychism is the only view that can do so. Alter and Nagasawa claim that other kinds of Russellian monism avoid the problems, and Chalmers argues that at least panprotopsychism does (Chalmers forthcoming-b: 2). Varieties of non-panpsychist Russellian monism include micropsychism, neutral monism, panprotopsychism and a view I will call mysterianism.

Micropsychism is the view that physical structure is partially realized by mental properties and partially by other kinds of properties. This view is distinct from impure panpsychism, which says that physical structure overall has aspects or properties that are not grounded in the mental, for example an irreducibly causal or spatiotemporal aspect. Rather, it is the view that only some of the realizers or grounds of physical structure (or those aspects of physical structure that needs grounds) are mental. For example, one might think that some types of particles, such as neutrinos, that have not been observed to

[19] By empirical evidence I mean observational evidence, but excluding evidence from introspection and phenomenological investigation.

play any part in constituting brains, have non-mental intrinsic properties, but other particles, the kinds that are found in brains, have mental intrinsic properties. Micropsychism is suggested and quickly rejected by Strawson (2006b: 24–25), mainly on the basis of parsimony: "I would bet a lot against there being such radical heterogeneity at the very bottom of things" (2006b: 25). He also regards it as a kind of dualism and it would perhaps therefore not really belong with the other Russellian monist views. However, as a form of Russellian dualism, as one may call it, micropsychism could have a chance of avoiding the argument from causal closure in a similar way as Russellian monism, so this important reason for rejecting ordinary dualism does not necessarily apply to it (I will discuss a form of Russellian dualism briefly in chapter 5, p. 202). But the argument from parsimony should be sufficient on its own, if one accepts the principle that other things being equal one should prefer the more parsimonious theory. Micropsychism does not seem to offer any significant advantages over panpsychism and is therefore at best equal in other respects – for example, they are both equally revisionary and counterintuitive in positing mentality at the fundamental level.[20]

Mysterian Russellian monism is the view that physical structure is realized by unknowable properties, which are not mental (otherwise they would not be unknown), but would explain consciousness if known. That consciousness must be explained by properties that are unknowable for creatures like us, properties of matter that are not physical in the ordinary sense, is a view that has been defended by Colin McGinn (1989) and that has subsequently been labeled mysterianism. McGinn does not commit to the Russellian view about the physical, but his mysterianism seems compatible with it and the way it avoids the problem of causal closure.[21] McGinn rules out that the unknown properties might really just be mental, dismissing panpsychism as "extravagant" (1989: 350, footnote 2).

[20] Micropsychism and panpsychism are very close views, and have all of their most controversial aspects in common, so refuting micropsychism is not a high priority for panpsychism. Chalmers defines panpsychism so as to be compatible with micropsychism: "[…] we can understand panpsychism as the thesis that some fundamental physical entities have mental states […] we can read the definition as requiring that all members of some fundamental physical types (all photons, for example) have mental states" (Chalmers forthcoming-b: 1). I find it useful to distinguish them, however, in particular in view of the argument from metaphysics and philosophy of science, to be discussed in the next section, which rules out micropsychism.

[21] If the causal relevance of the mysterian properties can in some (mysterious) way be transferred to the mental properties they explain.

Mysterian Russellian monism does not clearly avoid physicalism's and emergentism's intelligibility problems. Mysterian properties are non-mental, and the view thereby entails that the mental is constituted by or emerges from the non-mental, which Strawson argues is impossible. However, even if one agrees that the mental cannot be constituted by or emerge from the physical, one might not be equally convinced that the mental cannot be constituted by or emerge from other, non-physical, non-mental properties. Chalmers makes the following argument:

> The epistemic arguments all turn on a more specific gap between the physical and the phenomenal, ultimately arising from a gap between the structural (or the structural/dynamical) and the phenomenal. We have principled reasons to think that phenomenal truths cannot be grounded in structural truths. But we have no correspondingly good reason to think that phenomenal truths cannot be grounded in nonphenomenal (and nonstructural) truths […]. (Chalmers forthcoming-b: 14)

Given our *positive* grasp of the physical as structural, according to Chalmers, we can see how it is incapable of grounding the mental. One might also add that we can see how the mental cannot be (non-constitutively) necessitated by the physical. Given our *lack* of a positive grasp of mysterian non-mental properties, then, it follows by this reasoning that we *cannot* claim to see how they are also incapable of this. In contrast, Strawson thinks that merely based on our *negative* grasp of the physical as non-mental, we can rule out that it can ground or necessitate the mental. By this reasoning, since mysterian properties are also non-mental, the same can be ruled out for them.

However, even if one sides with Chalmers on this, and grants that mysterianism can therefore avoid the main problems of dualism and physicalism, the view faces another complaint. Mysterianism does not actually intelligibly locate and causally integrate consciousness in nature; it merely states that this would be possible if we knew the unknown or unknowable properties. It provides a schema for an explanation, without the explanation itself. Now, mysterianism can of course be true even though it is not very explanatory, and it is hard to think of an argument that would show that mysterianism could *not* be true. Strawson remarks, when confronted with this kind of view (he calls it radical Ignorantism), that it is "consistent, indefeasible, safe, but it opts out of the real difficulty" (Strawson 2006a: 273). It seems the argument for panpsychism, unless the strong principle that the mental cannot be necessitated by anything non-mental is accepted, will have to rely on a pragmatic or methodological premise that we should not

opt out of difficulties in this way, that we should not declare ignorance before we have considered and found good reason to reject all positive theories, or that we should always prefer theories with that give explanations as opposed to mere schemata for explanations. Perhaps this could be justified on the same grounds as general anti-skepticism about knowledge or justification: if we are not justified in believing a hypothesis just because an alternative hypothesis that nobody so far has been able to conceive of could be true instead, then it seems there is not much that we are justified in believing.

Neutral monism posits properties that are neither physical nor mental as realizers of physical structure and explanatory grounds of the mental. Neutral monist views can be divided into positive and negative versions, according to whether they also positively characterize the neutral properties, or only leave it at the "neither/nor". An example of negative neutral monism is panprotopsychism, as Chalmers defines it.[22] Protophenomenal properties are characterized only negatively (and relationally/dispositionally) as follows:

> […] protophenomenal properties are special properties that are not phenomenal (there is nothing it is like to have a single protophenomenal property), but that can collectively constitute phenomenal properties, perhaps when arranged in the right structure. (Chalmers forthcoming-b: 13)

Negative neutral monism is hardly distinguishable from mysterian Russellian monism, and it would therefore face the same objections: either, that it involves the mental being grounded in or necessitated by the non-mental, or that it fails to provide an actual positive explanation.

Positive neutral monism could also be quickly dismissed on the former basis, that the mental cannot be grounded in or necessitated by the non-mental, because neutral properties are in any case not mental. But what if one rejects, or is uncertain about, this strong principle? Positive neutral monist views still face another problem as to precisely how to characterize the neutral properties. Neutral properties seem to either end up looking much more mental than neutral with the result that the view collapses into panpsychism, or they will leave an epistemic gap between the neutral and the mental yielding problems of intelligibility similar to those of physicalism.

[22] Chalmers explicitly likens panprotopsychism to neutral monism (forthcoming-b: 17).

An example of a view that is susceptible to the former problem is the neutral monism of William James, who defended it at one point but later returned to panpsychism.[23] He describes the neutral as *pure experience*, and claims that consciousness arises when pure experiences relate to one another in a certain way (James 1912a). This might sound more like an explanation of self-consciousness or reflective experience in terms of unreflective experiences[24] than an explanation of the mental in terms of the non-mental, and accordingly more like panpsychism than neutral monism.

An example of a view susceptible to the latter problem is Coleman's neutral monism. Coleman holds that neutral *unexperienced qualities* are ubiquitous in matter, and whenever they relate appropriately, along the lines of higher-order thought theory, they constitute experienced qualities (Coleman 2013b). One problem with this view is that there seems to be an epistemic gap between unexperienced qualities and subjective experience of them. Chalmers (forthcoming-b: 25–27) constructs a conceivability argument against Coleman's view, according to which no matter how we conceive of unexperienced qualities as being related to each other does it seem necessary that they become experienced.[25]

Since panpsychism posits neither unknown nor any novel non-mental properties, it thus avoids the problems of non-panpsychist Russellian monism, as well as those of physicalism and dualism. Now, this claim naturally presupposes that panpsychism really does succeed in avoiding every one of these problems or closely similar ones, i.e., that the outline given so far of how it avoids all these problems survives closer scrutiny. As will

[23] According to Marcus P. Ford, "most of James' interpreters agree that in 1902 and 1903 James was a panpsychist, however, most contend that in 1904 and 1905, with the publication of his articles in the *Journal of Philosophy*, James turned away from panpsychism. According to Ralph Barton Perry, in 1904 James gave up his panpsychic notion that every instance of actuality is inherently psychical (i.e., an experience for itself) and endorsed instead the notion that every actuality is inherently neither psychical nor physical. In these articles, which were later published as *Essays in Radical Empiricism*, James did indeed speak of 'pure experiences' which are inherently neither psychical nor physical. However, there are good reasons to suspect that this was not his final position. One can find panpsychic notions in *Essays in Radical Empiricism* and one can find James endorsing panpsychic beliefs after 1905" (Ford 1981: 163–164).

[24] Together with a deflationary view of the subject, which is not incompatible with panpsychism (as explained in Strawson 2006a). I will discuss this further in many of the chapters below.

[25] In the passage quoted above (p. 23), Chalmers claimed that we have good reason to deny only that the gap between the physical (*qua* structural) and the mental is closable in principle, the reason being that we have a positive grasp of the physical as structural. Unexperienced qualities, are not (purely) structural, so the conclusion that there is an epistemic gap between unexperienced qualities and mental *qua* experiential properties that is not closable in principle cannot be deduced from this former conclusion. Rather, this gap must be independently demonstrable on the basis of our positive grasp of what unexperienced qualities are supposed to be.

be discussed in section 3.1 below, this is something that the so-called combination problem casts into doubt. For now, I will note that the argument for panpsychism from philosophy of mind presupposes that panpsychism can be defended against any such challenges that may come up.

In summary, it seems panpsychism, from the point of view of philosophy of mind, is motivated by a set of principles which are jointly incompatible with any other (so far suggested) view. These can be roughly listed as follows:

(i) There is an epistemic gap between mental and physical properties that is not closable in principle.
(ii) Epistemic gaps that are not closable in principle entail ontological gaps.
(iii) There is nothing about the physical in virtue of which the mental can non-brutely emerge.
(iv) Brute emergence is impossible.
(v) The mental is not epiphenomenal.
(vi) There is no systematic overdetermination.
(vii) The physical is causally closed.
(viii) The mental is not grounded in or emergent from properties whose nature is unknown or not positively conceivable.

The problems of the various views discussed so far can be put in terms of incompatibility with these principles as follows. Physicalism's intelligibility problem is its conflict with principles (i) or (ii). Emergentism's (which some regard as compatible with physicalism and others regard as entailing dualism) intelligibility problem is its conflict with principles (iii) or (iv). Dualism's empirical problem is its conflict with (v), (vi) or (vii). Physicalism's empirical problem is its conflict with principles (v), (vi) or (vii), but where (vii) is given a strong reading, as affirming not only physical but also microphysical causal closure. Non-panpsychist Russellian monism mainly runs into conflict with principle (viii). However, some versions of positive neutral monism are in conflict with (ii) insofar as an epistemic gap can be demonstrated between their Russellian properties and mental properties, or with (iv) insofar as it can be shown that there is nothing about these properties in virtue of which the mental can intelligibly emerge. Micropsychism is ruled out mainly on the basis of parsimony, which is not on the list, but is among the underlying reasons for accepting, e.g., (vi), that there is no overdetermination.

If principle (i) is strengthened to say that there is an epistemic gap between mental properties and any non-mental properties, including but not limited to physical properties, and principle (iii) is strengthened, correspondingly, to say that there is nothing about the non-mental in virtue of which the mental can non-brutely emerge, principle (viii) will be redundant, and non-panpsychist forms of Russellian monism can all be rejected on this basis. However, it seems principle (viii) is more widely endorsed than these strong principles.

2.2 The Argument from Metaphysics and Philosophy of Science

The other main argument for panpsychism starts from a problem in metaphysics and philosophy of science. This argument is wholly independent of the problem of consciousness, but it shares with the argument from philosophy of mind the underlying premise that science only tells us about the structure of the world, or that all physical properties are structural. But instead of taking this gap as something that *can* be filled by mental properties in order to solve problems in philosophy of mind, this argument takes the gap as something that *must* be filled. Then it asserts that mental properties are the only properties that could do so.

William Seager explains how G. W. Leibniz arrived at his panpsychism by combining the structuralist view of the physical with what he calls a strong reducibility principle: "All extrinsic properties are determined by intrinsic properties" (Seager 2006: 131). Seager goes on to describe the general structure of the Leibnizian argument:

> According to the reducibility principle, matter must have an intrinsic nature to ground the relational or structural features revealed to us by physical science. We are aware of but one intrinsic property of things, and that is consciousness. It is plausible to assert physicalism – we are physical beings and our consciousness is a feature of certain physical structures.[footnote omitted] Therefore, consciousness is an intrinsic property of matter. (Seager 2006: 136)

Alter and Nagasawa present a similar argument for Russellian monism:

> In the philosophy of science, there is a problem of a lack of metaphysical grounding. All fundamental physics gives us is nomic spatio-temporal structure. That is, it gives us little more than structure without any underlying non-structural properties. Some believe that what we should conclude from this is that nature consists in nothing but structure (Ladyman and Ross 2007). But Russell and others think that we must look outside of

> physics for properties that ground the network of causes and effects that physics describes. (Alter and Nagasawa 2012: 88–89)

According to both arguments, Russellian monism follows from structuralism about physics or physical theory together with a reducibility or grounding principle for the structural. Adding the principle that consciousness is the only categorical, intrinsic or otherwise suitably non-structural property we know, and that we should not posit wholly unknown properties, we get panpsychism.

I suggest that this argument for panpsychism can be formulated as follows:

> 1) *Structuralism about physical theory:* Science only tells us about the structure of the physical.
>
> 2) *Denial of ontological structuralism*: All physical structure must be instantiated by (individuals with) categorical properties.
>
> 3) *Mental categoricity:* The only categorical properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> 4) *Anti-mysterianism:* The (nature of the) properties that instantiate physical structure are knowable or positively conceivable.
>
> Therefore,
>
> 5) *Panpsychism:* Physical structure is instantiated by (individuals with) mental properties.

This formulation leaves open various ways of interpreting the premises, while at the same time entailing the conclusion. It is presupposed that none of the premises are vacuously true, i.e., that we know some categorical mental properties, and that there is some physical structure. Strictly speaking, premise 1 is superfluous – the conclusion follows from premises 2, 3 and 4 alone.[26] Its role is rather, as will be discussed, to support premise 3. I still include it because it matches the way in which the argument is usually presented and defended, and makes it read more intuitively. A preliminary note on

[26] Insofar as the non-vacuous truth of premises 2 and 4 requires that there is some physical structure.

categorical properties: I take them to be a kind of intrinsic properties, but this will be elaborated further below. I will now discuss all the premises one by one.

Premise 1 claims that science only tells us about the structure of the physical. What kind of structure? Russell claimed that science only reveals logico-mathematical structure. Alter and Nagasawa claim it reveals spatiotemporal, nomic structure. Simon Blackburn has argued that science only reveals dispositional structure:

> When we think of categorical grounds, we are apt to think of spatial configurations of things – hard, massy, shaped things, resisting penetration and displacement by others of their kind. But the categorical credentials of any item in this list are poor. Resistance is *par excellance* dispositional; extension is only of use, as Leibniz insisted, if there is some other property whose instancing defines the boundaries; hardness goes with resistance, and mass is knowable only by its dynamical effects. Turn up the magnification and we find things like an electrical charge at a point, or rather varying over a region, but the magnitude of a field at a region is known only through its effects on other things in spatial relations to that region. A region with charge is very different from a region without [...] It differs precisely in its dispositions or powers. But science finds only dispositions all the way down. (Blackburn 1990: 62–63)

As long as dispositional, nomic and spatiotemporal relations are not taken to include any categorical properties, it can be left open whether science reveals, in some sense, a structure of dispositional, nomic or spatiotemporal relations, or just pure mathematico-logical structure, because the conclusion will still follow.

What are the arguments for premise 1? Blackburn can be taken to give an inductive argument from the lack of counterexamples. If even paradigmatic candidates for physical categorical properties such as mass and extension can be analyzed as really dispositional, it is plausible that they all can. One can also note the how the theories of physics are formulated in mathematics. Mathematics is a language of relations – mathematical objects are exhaustively defined in terms of how they relate to each other. The number 2 seems not to have any intrinsic properties, only a position within the system of numbers. If we stipulated that from now on the number 2 will have no relationship to the other numbers – it is neither smaller nor greater than 3, it is not half as much as 4, and so on – then we would no longer in any meaningful sense be talking about the actual number 2, or any number at all – at least this would be so according to structuralism in philosophy of mathematics, as defended by, e.g., Stewart Shapiro (1997). If mathematics describes relational structures only, physics is formulated solely in terms of mathematics, and all

other physical theories are reducible to physics or themselves mathematical, then it follows that science only tells us about relational structure.

There are also arguments based on accounts of our epistemic situation. One might think that empirical knowledge is gained by way of causally relating to things, and that from this it follows that the relational properties of things are all we can know empirically. Kant can be understood as making an argument based on our epistemic situation that we could not know intrinsic and categorical properties. Such properties pertain to things as they are irrespective of their relations to us, and this would place them in the category of unknowable things-in-themselves. Rae Langton (1998) defends such an interpretation of Kant, as well as the plausibility of the view on its own terms.[27]

Premise 2 – all physical structure must be instantiated by individual with categorical properties – replaces Seager's Leibnizian reducibility principle, according to which all extrinsic properties reduce to intrinsic properties. Reducibility premises can be formulated in terms of all the Russellian distinctions. It is often said that dispositions need categorical grounds, or that relations need relata with intrinsic properties. One could regard all of these dependencies as ways in which structure can be instantiated.

As mentioned, categorical properties are a kind of intrinsic property. Intrinsic properties, according to David Lewis and Langton, are properties that are compatible with both the loneliness and the accompaniment of its bearer (1998: 334). As Lewis and Langton specify, some additional qualifications must be made in order to rule out all counterexamples, but it still makes the general idea reasonably clear. Intrinsic properties contrast with relational properties that make up structures. These are properties that *are,* characterize or are partly constituted by a relation between two or more properties or things. They are not independent of loneliness or accompaniment; they require the existence of at least one other property or thing.

Some think that categorical properties are not just a kind of intrinsic property, but that they are the only kind. The terms are very often used interchangeably. However, the categorical is typically contrasted with the dispositional and not the relational or extrinsic. Categorical properties can be defined as properties that are always manifest, or manifest

[27] How could a panpsychist accept this kind of argument, but claim that we can still know mental intrinsic properties? Is must be argued that mental intrinsic properties escape it, because they are not properties of other things, but our own properties, and we do not know them by empirically, or by causally relating them, but rather via direct acquaintance. This is in the spirit of Schopenhauer's criticism of Kant, which I will discuss in chapter 2 (section 2).

whenever they are actual. This means that they are manifest independently of what goes on around them, and are therefore intrinsic. Dispositional properties can be defined as properties that are manifest only in certain circumstances, or that can be actual without manifesting, which gives them a relational aspect. Since all categorical properties are intrinsic, and in order to capture the contrast with dispositionality, I put the argument in terms of categorical properties only. Whether the terms are in fact fully interchangeable, i.e., whether all intrinsic properties are also categorical, I leave open for now (but it will be discussed in section 3.2 below).

It is very intuitive, and often taken for granted, that all relations must have relata with intrinsic and categorical properties, that all dispositions have categorical grounds, and that structure cannot exist without categorical realizers. However, it is hard to back up the intuition once challenged. As Alter and Nagasawa mention in the quote above, acceptance of premise 1, the view that physical sciences only tell us about dispositional or relational structure and nothing about categorical grounds or properties of its relata, has led some philosophers, such as James Ladyman and Don Ross (2007), to conclude that no such properties are really needed, and that they should be eliminated from our ontology.

This view is known as ontic structural realism. It contrasts with epistemic structural realism, the view defended by Russell, according to which intrinsic, categorical, non-structural properties exist, but are either unknowable or accessible by non-scientific modes of inquiry. According to ontic structural realism, structure and relations can subsist on their own, or at least *prior* to their relata, such that the relata are constituted by their position in the relational structure and would have no reality outside of it. On this view, physical objects will be just like nodes in a graph, entities that have no properties except their position in the graph, or like numbers when conceived of according to the structuralist view about mathematical objects.

Premise 2 is often defended by appeal to arguments against ontic structural realism. One such argument starts out precisely from the manner in which physical objects become comparable to mathematical objects on ontic structural realism. On this basis, the view can charged of collapsing the distinction between the physical and the mathematical. It has been argued that "the difference between mathematical (uninstantiated) structure and physical (instantiated) structure cannot itself be explained in purely structural terms" (Bas van Fraassen as paraphrased in Ladyman and Ross 2007: 158). Ladyman and Ross

respond that the question of what makes structure physical and not mathematical is a question they refuse to answer (Ladyman and Ross 2007: 158).[28] If the physical ends up being indistinguishable from the mathematical, it seems the result is a kind of Pythagoreanism. If ontic structural realism entails Pythagoreanism, or it cannot explain why it does not, it can be regarded as a *reductio ad absurdum.* There seems to be only one current philosopher, Max Tegmark, who argues for an explicitly Pythagorean view.

Seager cites Newman's problem (Newman 1928) as one of the most powerful arguments against ontic structural realism:

> The conclusion of Newman's argument, when interpreted to bear on relationalism [i.e., structural realism], is that the existence of a system of relations is trivially true of a set of objects, so unless there is something, as Newman says, "qualitative" (I read this as involving intrinsic properties) about the relata, relationalism says exactly nothing about the world, beyond an assertion of cardinality. This is because, assuming there are enough entities it follows from pure logic that any system of relations over those entities is instantiated. How can that be? Because, conceived apart from considerations of the intrinsic properties of the relata, relations are simply sets of ordered sequences of entities (e.g. a two-place relation is a set of ordered pairs) and, given the entities, those sets and sequences will automatically exist. Newman puts it thus: "any collection of things can be organized so as to have the structure W, provided there are the right number of them." (Seager 2006: 142–143).

Now, Ladyman claims the objection from Newman's problem does not apply to ontic, as opposed to epistemic, structural realism because the ontic structural realist "eschews an extensional understanding of relations" (Ladyman 2013). This sounds *prima facie*

[28] When Ladyman and Ross refuse to answer the question, it is not because they think there is no distinction between the physical and the mathematical; rather, it is because they think there is nothing we can meaningfully say about the distinction. They appeal to a primitive notion of ineffable reality: physical structure is just real in a way mathematical structure is not. As they themselves point out, this sounds suspiciously Kantian (Ladyman and Ross 2007: 158). They pose themselves the question: "We say these structures describe real patterns, but since we can only represent the real patterns in question in terms of mathematical relationships, in what sense are these real patterns 'real' other that in which, according to Kant, noumena are real?" (Ladyman and Ross 2007: 299) – and again, do not give an answer (they merely point out other ways in which their view differs from Kant's). One might regard the reply as question-begging: merely stating that physical structure is "real" will not do when we are asking for an account of what makes it real. One might also claim that primitive "realness" should count as a categorical property, which means that ontic structural realism is compatible with premise 2 after all instead of threatening to refute it. Further down the line, the view would be eliminated on the basis that primitive realness (modeled on Kantian noumenality) is in tension with premise 4 of the argument: "*Anti-mysterianism*: The properties that instantiate physical structure are knowable or positively conceivable."

reasonable – if there are no entities prior to relata, how can all relations follow trivially from them? Still, it seems Newman's worry about indeterminacy can be generalized to ontic structural realism. The core of Newman's worry can be regarded as being that every abstract logico-mathematical structure or relational system *exists*, in the same way that we say that every number abstractly exists (and the existence of a number as a relational entity seems to entail the existence of the entire relational system of numbers). In contrast, the concrete structure of the world, of which there is only one, must have some non-mathematical and non-structural qualities. If there are no objects and no intrinsic, categorical properties, as ontic structural realism holds, then what makes it the case that the world has the mathematical structure that it actually has and not some other mathematical structure?

If no answer can be given to this question, the ontic structural realist is committed to the physical existence of all mathematically possible structures. Tegmark embraces also this consequence, making precisely this claim: "all structures that exist mathematically exist also physically" (Tegmark 1998: 1). For this reason, I think the main argument against ontic structural realism should be regarded as the *reductio* of Pythagoreanism, and the argument from Newman's problem (or something in its spirit) as a sub-argument spelling out why Pythagoreanism is to be taken a *reductio*, for those who do not find it problematic in and of itself.

Can the answer to what makes the structure physical or concrete be that it is spatiotemporal, nomic or dispositional? As discussed, the notion of structure can be understood so as to leave these options open. If one is a substantivalist, as opposed to a reductive relationalist, about space or space-time, one could certainly say that it can concretely and physically exist without there being things with some non-spatiotemporal properties to fill it. One could then claim that the properties of particles are reducible to their spatiotemporal extension and their spatiotemporal relations to other particles. However, it seems that entities that only have spatiotemporal properties are not distinguishable from mere spatiotemporal points (i.e., unoccupied locations). The difference between a mere spatiotemporal point or area and a real particle cannot be accounted for in purely spatiotemporal terms. This is a problem, because granted that space-time exists, it follows trivially that every spatiotemporal point, i.e., location in it, exists, and every spatiotemporal point is spatiotemporally related to every other spatiotemporal point. The result would be that all possible spatiotemporal structures exist

at once, or that the structure is radially indeterminate between all possibilities. Therefore, spatiotemporal structuralism ends up with the same problem as Pythagoreanism.

How about causal, i.e., nomic or dispositional, structure? Reductionism or Humeanism about causation entails that causation is either wholly grounded in relations between categorical properties, or that relations between categorical properties constitute all that we can know about causation (according to Hume, if there is anything more to causation, it cannot be experienced nor positively conceived of). Hence, there could be no causal structure without categorical properties on any such view. David Armstrong's influential non-reductive realism about the laws of nature construes them as relations between categorical universals, so there can be no nomic structure without categorical properties on this view either. However, another common view is that causation is grounded in causal powers or dispositions, which are not reducible to anything non-causal or non-dispositional, i.e., categorical. Ordinarily, it is supposed that objects with irreducible causal powers must still have categorical properties in addition to their causal powers. However, some philosophers deny this, claiming that objects can have purely dispositional essences. This potential problem for the argument from categorical properties is sometimes hinted at in the literature on panpsychism and Russellian monism; for example, Alter and Nagasawa remarks that Sydney Shoemaker's dispositionalist view of properties complicates the articulation of Russellian monism (Alter and Nagasawa 2012: 73), and Seager notes how some philosophers have taken causation to be among the qualitative properties needed to evade Newman's problem (Seager 2006: 143, footnote 9). But there has not been much discussion of what would be the best response to this. Below, in section 3.2, I will look at the problem in more detail. For now, I will just note that premise 2 rules out that things can have purely dispositional essences.

Premise 3 says that the only categorical properties (whose nature) we can know, or have a positive conception of, are mental properties. Mental properties such as the redness of red, or a total field of experience, could conceivably exist all alone in the universe, and could conceivably keep existing in the very same way if other things were introduced, i.e., independently of loneliness and accompaniment. Seager thinks Descartes is the first philosopher to argue for this thesis: "The philosophical problem of the external world and the coherence of solipsism entail that consciousness is an intrinsic property of things. We do not have to embrace Descartes's dualism to share this insight" (Seager 2006: 136). The fact that Cartesian solipsism is at least in some sense a conceivable and coherent scenario

shows that total experiential-cognitive fields could exist independently of whatever may or may not be going on outside of them.

The inverted spectrum thought experiment can also be taken to show that color phenomenology is not exhausted by the relations in which the colors stand to physical properties they represent, nor by their relations of similarity or difference to the other colors in the spectrum. If the redness of red remains even if all its relational properties were to be changed, as the inverted spectrum thought experiment aims to demonstrate, redness could not be a relational property.

This is not to say that mental properties are purely intrinsic or categorical, or that they have no relational and dispositional aspects. Alter and Nagasawa specify that:

> […] when [proponents of Russellian monism] characterize inscrutables intrinsic, they likely mean not that inscrutables lack extrinsic aspects altogether but only that such properties have intrinsic aspects. […] it is sometimes noted that phenomenal properties have structure and dynamics. The series of auditory phenomenal properties typically caused by hearing a musical scale plausibly has a structure corresponding to the scale. And your headache might become more intense over time. At first glance, such simple observations might seem to create problems for Russellian monism (Stoljar, 2006, pp. 144–9). But […] this concern is unfounded. The relevant claim is not that phenomenal properties lack structure or dynamics, but only that phenomenal properties are not merely structural or dynamic (Alter, 2009). (Alter and Nagasawa 2012: 74)

In addition, mental properties can be relational insofar as they represent, and dispositional in the sense that it seems the feeling of pain essentially disposes one toward avoidance, and pleasure does the opposite (this will be central to the argument I will defend in chapter 3). But not only does this not show that they are *purely* relational or dispositional; reflection on these properties seems to reveal the opposite. Perceptual phenomenology seems to represent in virtue of its intrinsic character, and pain seems to dispose in virtue of how it intrinsically feels, meaning that the relational or dispositional aspect does not exhaust these properties.

Are mental properties the *only* categorical properties we can know, or of which we have a positive conception? If science does not tell us about any such properties, there are not many places to look for them – this is how premise 1 supports premise 3. Alter and Nagasawa mention protophenomenal and neutral properties as alternatives. A panpsychist could in the context of this argument not reject these properties because they leave an epistemic gap with respect to mental properties, because the argument is supposed to

motivate panpsychism independently of the problem of consciousness. They would have to argue that we have no positive conception of them, that all we can specify is the role they are supposed to play (which is relational/dispositional) or their negative properties, and that they are therefore ruled out by premise 4.

But what about Coleman's unexperienced qualities? Qualitativeness is an intrinsic and positively specified property. Panpsychists would therefore have to argue that unexperienced qualities are metaphysically impossible, or that our positive conception of them is incoherent. Perhaps what we are really conceiving of, in the case of, say, unexperienced redness, is something that has the *disposition* to produce the experience of redness, in certain observers.[29] Accounting for how secondary qualities would be able to exist mind-independently is a classic philosophical problem, which indicates that entirely ruling out such concepts and related ones might not be a hopeless case. Perhaps one could claim that the burden of proof is on the proponents of unexperienced qualities to show that they are coherent and possible.

Alter and Nagasawa mention Derk Pereboom's suggestions for other positively conceivable intrinsic properties, which Pereboom claims to be physical, not neutral:

> Pereboom considers two candidates for what specific sorts of physical properties the inscrutables [i.e., Russellian intrinsic properties] might be: Aristotelian prime materiality; and absolute (or perfect) solidity, the notion of which he attributes to Locke and Newton. The former is notoriously obscure, but Pereboom implies that the latter should be regarded as a serious option. If absolute solidity is to qualify as an inscrutable, then it would have to differ from ordinary solidity, which seems manifestly dispositional. Whether we can make sense of this idea is not entirely clear. (Alter and Nagasawa 2012: 79)

As with unexperienced qualities, absolute solidity is arguably either dispositional or not clearly coherent, and panpsychists will have to assume that the burden of proof is on those who claim otherwise.

Stoljar has argued, partly on the basis of the problem of consciousness, but also on the basis that dispositions must have categorical grounds, that we should posit intrinsic or categorical properties which are physical according to the object-based conception of the physical but are inaccessible via science as we know it. These properties will either be

[29] Coleman seems to hold that the qualities are *in* the things and not *on* them like secondary qualities, but this does not clearly make them less mysterious or less mind-like.

made accessible to us through a scientific revolution, or it could be that: "[…] our contingent psychological nature is such that […] we cannot develop concepts of the categorical properties" (Stoljar 2001: 274). The view that these properties will be accessible after a scientific revolution, i.e., that they are unknown but not unknowable, is implausible in view of some of the arguments for premise 1, in particular the Kantian or Langtonian argument that the intrinsic properties of other things can in principle not be known empirically. If the properties are not only unknown but unknowable, they will be ruled out by premise 4 – along with Kantian (or Langtonian) things-in-themselves and other merely negatively defined properties.

Premise 4, which states that the properties that instantiate physical structure are knowable or positively conceivable, is similar to the anti-mysterian principle (viii) (p. 26 above) which forms part of the motivation for panpsychism from philosophy of mind. Given that it seems hard to prove that unknowable or inconceivable realizers cannot possibly exist, this premise must, like principle (viii), also be defended on pragmatic, methodological or anti-skeptical grounds. A reasonable pragmatic or methodological premise is that we should not declare ignorance before we have considered and found good reason to reject all positive theories, i.e., that we should, if possible, avoid positing properties that (or whose nature) are unknowable or not positively conceivable. This does not in itself entail premise 4, but in combination with premise 3, according to which there *are* some suitable knowable properties, and hence not positing unknown ones *is* possible, it comes close to. One could make all this explicit, by putting the methodological principle as a premise instead of premise 4. However, then the argument would not be formally valid, unless further modifications are made: one must either add premises which license moving from a methodological or pragmatic recommendation to a factual statement (i.e., premise 4 as it stands), or put the conclusion in terms of a pragmatic or methodological recommendation as well (e.g., "we should accept panpsychism" instead of "panpsychism is true").

Perhaps premise 4 would also be defensible on the basis of premise 3 in combination with a principle of parsimony: why multiply our ontology with unknown properties, when all that is needed are the mental properties that are part of it already? However, if one takes the view that mental properties can be grounded in the posited unknown properties, as, e.g., neutral monists would think, introducing them lets us exchange one fundamental property for another, e.g., the mental for the neutral, and our fundamental ontology will

not be multiplied after all. Therefore, considerations of parsimony cannot be invoked against unknown properties if they are regarded as suitable to ground consciousness.

How does the argument from categorical properties relate to the Hegelian argument from philosophy of mind? Alter and Nagasawa point out how the two arguments – or the two arguments they consider, which are equivalent except that they leave out the premises and presuppositions that rule out non-panpsychist Russellian monism – together constitute a "solving two problems at once"-argument:

> In the philosophy of science, there is a problem of a lack of metaphysical grounding. […] In the philosophy of mind, there is a problem about integrating consciousness into nature. […]
>
> At first glance, these problems may seem to have nothing to do with each other. But on reflection, they might be related. The philosophy of science problem could be described as a help-wanted problem. Physics wants to hire help: it wants to employ something outside its purview to ground the structure it so elegantly describes. The philosophy of mind problem could likewise be described as a job-seeking problem. Consciousness wants a job: it wants to be integrated into nature by playing a role in the causal nexus known as the cosmos. Seen in this way, a unified solution suggests itself: consciousness can be employed to ground fundamental physical relations – which is what Russellian monism says, with the one qualification that on some versions of Russellian monism it is not consciousness itself but its components (protophenomenal properties) that ground the properties found in physics. So, Russellian monism provides what seems on reflection to be a natural solution to two significant philosophical problems. This speaks in favour of the view. (Alter and Nagasawa 2012: 89)

Nagel is one who finds the argument from philosophy of mind alone compelling, but regards panpsychism as so implausible that he concludes that there must be an error somewhere in the argument – even though he cannot precisely identify it.[30] Adding in the argument from categorical properties, and seeing the "solving two problems at once" or "help-wanted/job-needed" meta-argument, would give panpsychism more plausibility and

[30] Nagel claims that panpsychism "appears to follow from a few simple premises, each of which is more plausible than its denial, though not perhaps more plausible than the denial of panpsychism." (Nagel 1979: 181). He concludes, accordingly, that: "[…] panpsychism should be added to the current list of mutually incompatible and hopelessly unacceptable solutions to the mind–body problem. It can be avoided by denying any of the premises of the argument" (Nagel 1979: 193). He also considers whether there might be an alternative, yet to be conceived of, position which is also compatible with the premises: "There is no reason to think that all possibilities have been thought of so there is no reason to assume that a view is correct if all currently conceivable alternatives are even more unacceptable" (Nagel 1979: 193). In *Mind and Cosmos* (2012) he is more sympathetic to the view.

therefore perhaps lead some philosophers to apply *modus ponens* instead of *modus tollens* to the individual arguments.

3 PROBLEMS FOR THE ARGUMENTS

Each line of argument faces (at least) one serious objection, both of which I have already indicated. The overall argument from philosophy of mind, constituted mainly by the Hegelian and anti-mysterian arguments, is possibly undermined by the combination problem. It appears to show that panpsychism is not, after all, compatible with the principles listed above (p. 26), as the overall argument from philosophy of mind can be construed as claiming. The argument from metaphysics and philosophy of science, on the other hand, i.e., the argument from categorical properties, is challenged by metaphysical positions such as dispositionalism and dispositional essentialism, because it relies on a reducibility principle which these views deny.

3.1 The Combination Problem

According to the Hegelian argument for panpsychism, mental properties cannot arise, at least not intelligibly, from any (known) non-mental properties, which supports the view that mentality must itself be fundamental. However, it is generally assumed that panpsychism allows that *our* mentality is not fundamental; rather, it results from the appropriate combination of the simple or primitive mental properties belonging to the ultimate particles that constitute our brain. This is what gives rise to the combination problem.

Before explaining the different aspects the problem I will introduce some terminology. Micromentality, microexperience, and so on, is the kind of experience that would belong to ultimate particles, and microsubjects are the particles considered as subjects of microexperience. Macromentality, macroexperience, and so on, is the kind of experience had by humans and animals and any other properly unified complex physical things, and macrosubjects are subjects of macroexperience.

I will use the terms constitutive, non-constitutive and emergent panpsychism as Chalmers defines them:

> Constitutive panpsychism is the thesis that macrophenomenal truths are (wholly or partially) grounded in microphenomenal truths. Nonconstitutive panpsychism is the thesis that macrophenomenal truths are not grounded in microphenomenal truths. The most important form of nonconstitutive panpsychism is emergent panpsychism, on which macrophenomenal properties are strongly emergent from microphenomenal or

> microphysical properties, perhaps in virtue of fundamental laws connecting microphenomenal to macrophenomenal. (Chalmers forthcoming-a: 3)

Grounding is a constitutive, as opposed to causal, determination relation. The grounded obtains *in virtue of* the ground and the ground necessitates the grounded in a way that leaves no explanatory gap (Fine 2012: 39). It has been suggested that reduction in philosophy of mind should be understood in terms of grounding, and the supervenience correlation should be regarded as a symptom of a grounding relation (Schaffer 2009). Strong emergence is when the properties of a whole are not predictable in principle from complete knowledge of the properties and configuration of its parts – that is, the properties of the parts as they manifest in isolation or in different kinds of wholes (Broad 1925: 61).[31] There are various accounts of what kind of determination relation underlies strong emergence, but it cannot be constitutive.

The combination problem is the problem of explaining how macroexperience results from microexperience. But, as discussion has showed, this general problem really comes down to a set of many different problems. In one way, it is natural to regard the combination problem as the set of all problems having to do with mental combination. From the dialectical point of view, however, it makes sense to regard the combination problem as a problem whose solution requires showing two things: (1) that mental combination is not demonstratively impossible, and (2) that when accounting for mental combination panpsychism does not face any problems that are strongly analogous to the main problems of physicalism, dualism and non-panpsychist Russellian monism.[32] In other words, it would involve showing that panpsychism cannot be accused of merely

[31] C. D. Broad is one of the classics of British emergentism. His definition of strong emergence is endorsed (with some qualifications) by contemporary philosophers such as Brian McLaughlin (1992) and F. C. Boogerd, F. J. Bruggeman, Robert C. Richardson, Achim Stephan & H. Westerhoff (2005).

[32] What about problems that are non-analogous, or weakly analogous, but just as hard? If "just as hard" means "just as unlikely to have a solution", then problems that are non-analogous but just as hard would perhaps be as significant as strongly analogous problems. But how can you tell that two problems are just as hard, unless they are analogous? If there are two problems, both of which philosophers have worked on for a long time with equal lack of progress, that would be an indication that they are just as hard. But the combination problem has not received much attention compared to the problems of physicalism and dualism, so one cannot compare their difficulty in this way. Therefore, it seems analogous structure is the only way (at the moment) to tell whether he combination problem is as hard as the problems of other views, and the combination problem can (for now) be understood in these terms.

An additional consideration is that even if two problems are non-analogous, but still as unlikely to have a solution, as long as neither is wholly unlikely to have a solution, there is still a chance that one of the problem can eventually be solved and the other cannot. If two problem are strongly analogous, there is more reason to believe that if one is insoluble then so is the other, and vice versa.

moving and repeating the problems that it purports to solve, given that the fact that it appears to solve these problems is the only reason, from the point of view of philosophy of mind, for considering the view in the first place. If these two demands cannot be fulfilled then there would be no motivation for accepting panpsychism, at least none deriving from philosophy of mind. However, if panpsychism *replaces* the problems of dualism, physicalism and non-panpsychist Russellian monism with different problems, problems which are not demonstrably insoluble, then it could be reasonably regarded as making some progress on the mind–body problem, even if some of these new problems have to do with mental combination. I do not think this way of limiting the problem would exclude many problems that have been called combination problems in the literature, and many seem to think of the problem in the same way (see, e.g., Goff (2009), Coleman (2012: 138) and Seager (2010: 170)).

Therefore, I will take it that an adequate solution to combination problem would consist in showing that mental combination can occur in a way which does not conflict with any of the premises and presuppositions of the main arguments against dualism, physicalism and non-panpsychist Russellian monism, the most important of which I regard as being the principles identified and listed at the end of section 2.1 (p. 26) above.

The combination problem was first posed by James:

> Take a sentence of a dozen words, and take twelve men and tell to each one word. Then stand the men in a row or jam them in a bunch, and let each think of his word as intently as he will; nowhere will there be a consciousness of the whole sentence. (James 1890/1981: 162)

> Where the elemental units are supposed to be feelings, the case is in no wise altered. Take a hundred of them, shuffle them and pack them as close together as you can (whatever that might mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feeling were set up, a consciousness *belonging to the group as such* should emerge. And this 101st feeling would be a totally new fact; the 100 original feelings might, by a curious physical law, be a signal for its *creation*, when they came together; but they would have no substantial identity with it, nor it with them, and one could never deduce the one from the others, or (in any intelligible sense) say that they *evolved* it. (James 1890/1981: 162)

James points out an apparent epistemic gap between microexperience and macroexperience: macroexperience seems not deducible from microexperiences and does not in any intelligible sense evolve from it.

Constitutive panpsychism seems to have the same type of options with respect to the epistemic gap pointed out by James as (non-emergent) physicalism has with respect to the epistemic gap from the physical to the mental: denying it, arguing that it is closable in principle, or arguing that epistemic gaps that are not closable in principle do not entail ontological gap when the gap exists between mental properties. Emergent panpsychism needs an argument that the strong emergence of macroexperience from macroexperience is not brute or wholly unintelligible.

Panpsychism thus seems to face something that looks very much like analogues of physicalism's and emergentism's intelligibility problems. If it turns out that they are strongly analogous and equally difficult, then the argument from philosophy of mind is undermined – panpsychism would not avoid the problems of physicalism and emergentism after all, it would just have relocated them. However, it is far from clear that panpsychism's intelligibility problems are analogous. I will discuss some of the reasons for and against thinking this is so, for constitutive and emergent panpsychism respectively. I will also point out how each also faces analogues of physicalism's and dualism's respective empirical problems in view of the principle of microphysical causal closure.

3.1.1 Combination Problems for Constitutive Panpsychism

Is it plausible to deny that there is an epistemic gap from micro- to macroexperience? The epistemic gap between micro- and macroexperience can be demonstrated in a way which looks perfectly analogous to the way in which the physicalist epistemic gap is demonstrated. One can construct a panpsychist conceivability argument, a panpsychist knowledge argument and a panpsychist explanatory gap.

Goff (2009) argues that we can conceive of a physical and microexperiential duplicate of our world – where every particle has the same physical properties and the same microexperiential properties that panpsychists think they have in our world – but where there is no macroexperience. The individual particles in human bodies are all microconscious, but our kind of experience is absent. This seems conceivable, and conceivable distinctness is an epistemic gap.

Chalmers constructs a version of the knowledge argument for panpsychism:

> We can suppose that inside her black-and-white room, Mary is told all the microphysical facts, and also learns all the microphenomenal facts: she learns what it is like to be a quark, a photon, and so on. Perhaps this is accomplished by giving her versions of those experiences, or by somehow enabling her to imagine them. One might think that in this situation, Mary would still be unable to know what it is like to see red, even given arbitrary a priori reasoning. (Chalmers forthcoming-a: 11)

As for an explanatory gap, one could take the general formulation of the combination problem to constitute a hard problem of macroexperience: why and how does microexperience along with its physical aspects give rise to macroexperience? It seems no amount of physical and microexperiential information would answer this question.

Because the panpsychist epistemic gap can be demonstrated in a way analogous to the way in which the physicalist epistemic gap is demonstrated, it is implausible for panpsychists who are motivated by the main anti-physicalist arguments to deny that it exists. Is it more plausible to claim that it is closable in principle?

Strawson suggests, in effect, that the epistemic gap results from ignorance of the nature of *macro*experience (2006a: 252–253). The argument from philosophy of mind as I have presented it presupposes that we are acquainted with the nature of our own experience. However, it is precisely in view of the combination problem that Strawson makes clear that he does not endorse a Full Revelation Thesis but only the Partial Revelation Thesis (see p. 16 above), according to which we are acquainted with the essential nature of experience only in certain respects. This allows that there may be unknown properties or aspects of our experience that account for how it can result from combination of microexperiences. If they were revealed to us, the epistemic gap would be closed.

Goff has criticized the partial revelation thesis on the grounds that the notion of partial or "semi-acquaintance" that it presupposes has not been properly spelled out, and furthermore, that it seems that the combination problem arises from features of experience that belong to the part of our experience that we *are* in fact acquainted with (Goff manuscript). Even if there are unknown parts of our experience that avoid these problems, it is hard to see how that changes the fact that the *known* parts seem unable to be constituted by microexperience.

Goff proposes instead a solution in terms of our ignorance about the nature of *micro*experience. Chalmers claims, similarly, that we know so little about what microphenomenal properties are like, and that "perhaps once we grasped them, we would

understand their connection to experiences of red and to other experiences" (Chalmers forthcoming-a: 11). But how can it be that we are fully acquainted with the nature of macroexperience, in relevant respects (as per Goff's argument against Strawson's semi-acquaintance), while at the same time ignorant about the nature of its constituents? Goff thinks this is possible on the basis of an objection he credits to Chalmers: "it seems that we can completely understand the nature of a property without understanding the nature of the properties that constitute it." Goff proposes that there could be a special *phenomenal bonding* relation, which is such that when microphenomenal properties stand in it, they would constitute a macrophenomenal property. He speculates that it might be the intrinsic nature of the spatiotemporal relation, and that if we were acquainted with phenomenal bonding, we could see how microexperiences thus related transparently entail macroexperience (Goff forthcoming).

This is, as Goff admits, a form of mysterianism about combination. We have no positive grasp of the nature of the phenomenal bonding relation. One could therefore ask whether phenomenal bonding ends up explaining macroexperience much better than mysterian non-panpsychist Russellian monism. Neither of these views give us an actual explanation of consciousness; they rather explain how an explanation is possible. Panpsychist mysterianism does give us some more detail on how the explanation is possible; there are fewer blanks to fill in. But it is not clear whether this is sufficient reason to prefer panpsychist mysterianism over non-panpsychist mysterianism.

Additionally, the claim that "we can completely understand the nature of a property without understanding the nature of the properties that constitute it" – I will refer to this as *Opaque Constitution* – is not obviously true for all cases of actual constitution, at least not the ones that would be most similar to constitutive mental combination. For functional or abstract properties, *Opaque Constitution* might hold true – it seems, for example, that we can completely understand the nature of computation without understanding the nature of silicon and other constituents of actual computers. But macroconsciousness is not a functional property, according to Russellian panpsychism, so this example would not show anything. With other properties that we intuitively regard as concrete, *Opaque Constitution* is more doubtful. It seems that we cannot say that we completely understand the nature of physical substances, such as water, as long as we do not completely understand the nature of the subatomic particles that constitute them. However, since on the Russellian view all physical properties, including the property of being water, are functional or structural, this example would perhaps not show anything either. Goff gives

an example in support of *Opaque Constitution* from the realm of qualitative properties, namely how redness is constituted by scarlet (a shade of red). But this is not clearly a good analogy either. One difference is that this is a one-one relation, while mental combination is a many-one relation. Another difference is that redness could be regarded as an abstract or disjunctive property relative to scarlet – red is either the disjunction of all particular shades of red, or an abstraction of what they all have in common. A field of macroconsciousness, on the other hand, is neither a disjunction nor an abstraction. These worries about *Opaque Constitution* are perhaps not serious enough to rule it out for mental combination, but at least it adds to the amount of mystery that must be accepted along with constitutive panpsychism, and makes it even less clearly preferable to mysterian Russellian monism.

Arguing that epistemic gaps that are not closable in principle do not entail ontological gaps in mental cases is a final possibility, which nobody seems to have taken up. Chalmers mentions the option of type-B constitutive panpsychism, but sees no reason why it would not inherit the problems of type-B physicalism. Altogether, then, I think there is good, but not conclusive reason, to think that constitutive panpsychism faces a strong, but partial, analogue of physicalism's intelligibility problems. The analogy is only partial because physicalism is committed to saying that we have a positive grasp of physical properties via physical theory, but constitutive panpsychists are not equally clearly committed to saying that we have a positive grasp of microphenomenal properties or the nature of the relations they can stand in, and therefore one cannot rule out that the epistemic gap between the micro- and macromental is closable in principle. However, the element of mysterianism that comes with an appeal to the unknown nature of the microphenomenal gives rise to an analogue of the problems of non-panpsychist versions of Russellian monism instead.

Chalmers claims that the combination problem can be broken into three main problems concerning the combination of three different aspects of experience: subjectivity, qualitativeness and structure (Chalmers forthcoming-a: 4). When it comes to the combination of subjectivity, some philosophers have argued that this would involve not only an epistemic gap but also a contradiction.

Coleman (2012) argues that combination of many microsubjects into one macrosubject requires that the microsubjects survive the combination, because: "If one subject is left where formerly we had two, this means at least one subject has gone out of

existence, which is not combination but a fight to the death" (Coleman 2013b: 14). But if we allow the original subjects to continue existing, that rules out combination:

> What we wanted was to assemble our two subjects so as to constitute a unified higher-level subject, with its own point of view. If we are left with all and only the original pair of points of view then we still have a multitude, and are nowhere nearer to the genuine combination of subjectivities into one subject. (Coleman 2013b: 14)

Coleman thinks the problem requires that we let go of microsubjects and posit only *unexperienced* qualities all the way down, but as has been discussed, this amounts to abandoning panpsychism.

Pierfransesco Basile also argues that the combination problem points not only to a mystery but to a contradiction, given the following three assumptions about the nature of our experience:

> PHENOMENAL ESSENTIALISM – this is the view that, for an experience, to be is to feel a certain way. In the case of a pain, it seems pointless to draw a distinction between what the pain is in itself and the way it feels; the former – what the pain is 'in itself' – is wholly exhausted by the latter – its qualitative, felt dimension. […]
>
> PHENOMENAL HOLISM – this is the view that, within a person's total psychical whole, the nature of a single identifiable experience […] is essentially determined by the other experiences occurring along-side it – synchronically – within the whole. […]
>
> THE SHARING PRINCIPLE – this is the view that an experience can simultaneously occur within two distinct psychical wholes – i.e. the very same experience can be felt by two different feelers, in this case, by the 'lesser mind' of the neuron and the 'larger mind' of the human being. (Basile 2010: 107–108)

Basile argues that:

> Once these notions are accepted, it becomes clear that the notion of mental combination involves a contradiction. If mental combination is to be possible, an experience must be felt by two different subjects while remaining numerically self-identical. But if the being of an experience is wholly exhausted by the way it feels, then an experience cannot be numerically the same while being felt by two different feelers. So, it would seem, mental composition is an impossible conception. (Basile 2009: 109)

The reason an experience cannot be numerically the same while being felt by two different feelers is the phenomenal holism principle, according to which the experience would feel different (and thus be different, according to phenomenal essentialism) for each of its feelers, since they each experience it within different total phenomenal wholes.

Basile concludes that we should reject the sharing principle, and thereby accept that mental composition cannot happen constitutively. It seems that without sharing, microsubjects and macrosubjects cannot overlap in the way constitution requires. One might speculate about whether one of the other principles can be abandoned, in order to preserve the possibility of constitution. However, a view very close to phenomenal essentialism, that the nature or essence of an experience consists, at least partly, in how it feels, is necessary for the rejection of type-B physicalism and the phenomenal concept strategy. But phenomenal holism is perhaps less clearly indispensable. Barry Dainton (2010) notes that it was relatively commonplace among nineteenth century philosophers, such as James and F. H. Bradley. Timothy Sprigge defends it at length, for both metaphysical reasons and phenomenological reasons, i.e., it being apparent to introspection. Dainton points to many actual cases of phenomenal interdependence, from the way in which details in a painting can mutually influence each other's character, to the Müller-Lyer illusion and inter-modal dependencies such as the McGurk effect.[33] He shows that a growing body of evidence shows that this sort of interdependence is much more common than usually thought (Dainton 2010: 121–123). He argues, however, that one cannot infer from there being many cases of phenomenal interdependence to its being complete and necessary. It is not clear whether Basile's argument needs it to be complete and necessary, as opposed to contingent but relatively prevalent. In any case, constitutive panpsychists would need an argument that phenomenal interdependence is not complete and necessary. Alternatively, it must be argued that the contradiction is somehow only apparent. I will discuss the phenomenal holism principle and its justification further in chapter 5, sections 4.1.2 and 4.3.

James also thought that combination seemed not only mysterious but contradictory. As well as being the father of the combination problem, James is also the father of one of the most radical proposals for its solution. He was convinced that combination was actual, and hence *somehow* possible. In the absence of alternatives, he found himself compelled to "*give up the logic*, fairly, squarely and irrevocably", that he concluded was to blame for the problem (James 1909/1977: 96). So it appears that constitutive panpsychism has not only an intelligibility problem but also a worse problem of appearing impossible.

[33] In the Müller-Lyer illusion, two horizontal lines of equal length appear to be of different length depending on the direction of "tail-fins" at the ends. In the McGurk effect, the visual perception of lip movements makes us hear speech differently – seeing the lip movements of "gaaa" makes us hear a simultaneous "baaa" sound as "daaa" (Dainton 2010: 121).

A noted above, physicalism also arguably has an empirical problem, the exclusion problem, in view of the principle of microphysical causal closure. Seager argues that the kind of panpsychism according to which our kind of consciousness arises by what he calls conservative emergence – weak emergence that is really a kind of constitution – faces a problem that seems strongly analogous to the exclusion problem:

> […] conservative emergence is all that is required to rob complex consciousness of its efficacy and to generate the paradox of consciousness. The panpsychist hypothesizes that there are elemental mental properties which belong to the fundamental physical entities of the world. If these elemental features have their own causal powers (that is, are not themselves epiphenomenal) then by the logic of conservative emergence they will usurp efficacy from the complex conscious states which they subvene. (Seager 2012: 203)

If constitutive panpsychism is to be accepted on the grounds that it avoids *all* the problems of physicalism, then it must either be explained why the exclusion problem is not really a problem for physicalism, as some philosophers hold, or it should be explained why macroexperience is nevertheless not causally excluded by microexperience.

3.1.2 Combination Problems for Emergent Panpsychism

Strawson claims that while it may be that mental combination requires emergence in some sense, it does not require *brute* emergence. The kind of emergence required to fill in the gap between micro- and macroexperience is more acceptable: "unintelligible experiential-from-experiential emergence is not nearly as bad as unintelligible experiential-from-non-experiential emergence" (Strawson 2006a: 250).

How can it be justified that strong emergence within the mental is less bad? Strong emergence is by definition characterized by an epistemic gap that is not closable in principle – except, perhaps, insofar as the epistemic gap between causes and effects in general is closable in principle. If the presence of such an epistemic gap constitutes complete lack of intelligibility, i.e., bruteness, then there can be no such thing as non-brute strong emergence, and presumably no way in which emergence within the mental could be less bad. In order for intra-mental emergence to possibly be less bad, one would have to regard epistemic gaps as something that is incompatible with explanation by constitution but not all explanation; that their presence indicates not that there is no explanatory connection between two phenomena, but rather only that there is no explanatory connection of the constitutive kind. That there can be explanatory relations between phenomena separated by an ontological gap is affirmed by ordinary causal

explanation. There is an ontological gap between causes and effect, but we still explain – which is to render intelligible, in some sense – effects in terms of their causes.

The argument from non-emergence as Strawson presents it is compatible with thinking of epistemic gaps as a problem for constitution only. He can be read as first arguing that there is nothing about the physic(S)al in virtue of which the mental can be constituted by it, which can be seen from the epistemic gap, and then making the further claim that there is also nothing about the physic(S)al in virtue of which the mental could be necessitated by or emerge from it, in a way similar to the way in which physical causes produce physical effects (I presupposed this reading in the way I formulated the principles behind the total argument for panpsychism from philosophy of mind, on p. 26 above). Is there similar reason to think that there is nothing about the micromental in virtue of which the macromental could emerge, be produced or necessitated? Many find it intuitive that this is not the case; these two somehow have a closer connection. However, Strawson is not very explicit about what this connection would be, apart from perhaps just claiming that the macromental can emerge from the micromental in virtue of the latter being *mental*, which is not the most satisfying answer (as Goff (2006: 59) emphasizes). A more satisfying answer to the combination problem in terms of non-brute strong emergence should specify more precisely what it is about the micromental, as opposed to the merely physical, in virtue of which the macromental can emerge from it; or alternatively, why it is that the macromental can emerge from the micromental in virtue of the latter being mental.

Seager (2010) has argued that combination could be a matter of *combinatorial infusion*, a process where constituents lose their individuality as they fuse, or are absorbed, into a so-called "large simple". Large simples are "partless, yet extended", just like macrosubjects. He argues that the models we have of classical black holes show this structure. This means that:

> […] we are already in possession of a rigorous model of a kind of combination that has the properties demanded by combinatorial infusion. It is therefore possible to deploy this sort of model when thinking about how complex states of consciousness could emerge from simpler ones. (Seager 2010: 181)

The black hole would be categorized as an emergent phenomenon, in one sense of the term, but Seager claims that "there is no *radical* emergence here, but there is nonetheless the creation of a new entity" (Seager 2010: 181, my emphasis). Why is this kind of

emergence not radical, and presumably then, not brute? Seager does not answer this explicitly. *Prima facie*, the process of black hole formation is intelligible only on the basis of *a posteriori* laws of nature, which do seem completely epistemically brute: we do not know why the laws of nature are as they are, or why there are laws at all. Hence, the combinatorial infusion theory cannot be a complete solution to the intelligibility problem of emergent panpsychism. If the black hole model applies to combination, this could show how combination can be modelled without *formal* contradiction, but it does not show that no brute emergence is involved (which is perhaps to be regarded as a kind of metaphysical contradiction).

Emergent panpsychism faces not only an intelligibility problem but also an analogue of dualism's empirical problem. If macroexperience is distinct from microexperience, microexperience belongs to microphysical entities and macroexperience to macrophysical entities, and the microphysical is causally closed, then macroexperience is either epiphenomenal or an overdeterminer.

By the fusion account, this problem is *prima facie* avoided, because the fusion replaces its microscopic base, and so it would not be around to compete for causal relevance. However, this still leaves another kind of empirical problem, a problem of structural mismatch when linking combination to physical emergent processes. Seager says:

> DMP [deferential monadic panpsychism, i.e., Russellian monist panpsychism] requires that the mental realm shadows the physical, so we expect to find physical correlates of mental processes, including any case of combinatorial infusion. (Seager 2010: 181)

This is in line with the criterion of structural isomorphism or compatibility for Russellian monism explained above (p. 10). Now, as Seager points out: "there is little evidence that the brain supports any processes that could count as combinatorial infusion at the physical level" (Seager 2010: 181–182). Would this not entail that the combinatorial infusion theory of combination leads precisely to the kind of structural mismatch which is incompatible with Russellian monism, and its solution to the problem of mental causation? Seager thinks not, claiming that "it is not a requirement of DMP [i.e., Russellian monist panpsychism] that a mental instance of combinatorial infusion be accompanied by a physical instance of combinatorial infusion" (Seager 2010: 181). However, he gives no clear reasons for why this is not a requirement. In what other sense can the mental shadow the physical in the way Russellian monism requires? Until this is

explained, the theory does not constitute a complete response to the empirical aspect of emergent panpsychism's combination problem.

In summary, the combination problem threatens to undermine the line of argument for panpsychism from philosophy of mind by showing how panpsychism faces strong analogues of the problems of dualism, physicalism and, to some extent, non-panpsychist Russellian monism. These are problems that, according to the argument from philosophy of mind, the view was supposed to avoid. Constitutive panpsychism inherits an intelligibility problem and an empirical problem from physicalism: the epistemic gap and the exclusion problem. Emergent panpsychism inherits an intelligibility problem and an empirical problem from emergentism and dualism: the problems of brute emergence and the problem of microphysical (and hence micromental) causal closure. Some proposed solutions to the combination problem come at the expense of cancelling panpsychism's advantage over mysterian non-panpsychist Russellian monism, by appealing to unknown properties. In addition, constitutive panpsychism must respond to the charge that mental combination involves a contradiction relative to two important principles about the nature of experience, phenomenal holism and phenomenal essentialism/The Partial Revelation Thesis, the former of which can be supported by phenomenological reflection and empirical evidence; the latter of which is a presupposition of some of the anti-physicalist arguments that are part of the case for panpsychism in the first place.

3.2 Irreducible Dispositionality

Above, I noted a potential problem for the argument from categorical properties, namely that dispositional properties, such as causal powers, might be irreducible and not in need of categorical grounds. Pure, i.e., irreducible and ungrounded, dispositions or powers form the basis for views such as dispositionalism and dispositional essentialism. According to the former, dispositional properties ground or instantiate all physical structure; according to the latter, they ground or instantiate most of it. Dispositionalism is defended by e.g., Alexander Bird (2007), Stephen Mumford (2004) and Shoemaker (1980); dispositional essentialism is defended by Brian Ellis (2002) and George Molnar (2003). *Prima facie,* these views are in conflict with the reducibility principle which the argument from categorical properties relies on, which I put as its premise 2:

> 2) *Denial of ontological structuralism*: All physical structure must be instantiated by (individuals with) categorical properties.

I will now examine this challenge in more detail. I will show that the way in which premise 2 is usually defended is not sufficient to meet it, and while there are a few apparent quick fixes, they ultimately fail. Therefore, the notion of irreducible dispositionality constitutes a serious challenge to the argument from categorical properties.

As discussed above, arguments against ontic structural realism, including the Pythagorean *reductio* and arguments derived from Newman's problem, are mainly appealed to in defense of premise 2, as well as alternative ways of formulating the reducibility principle. However, it seems that these arguments really only support a weaker version of the premise, which denies only that physical structure exists uninstantiated, or can be self-instantiating:

> 2*) *Denial of ontological <u>pure</u> structuralism*: All physical structure must be instantiated by (individuals with) *non-structural* properties.

It is only uninstantiated or self-instantiating structure which is indistinguishable from mathematical structure (or empty space-time). The original premise 2 does not only say that physical structure must be instantiated by non-structural properties; it makes the stronger claim that it must be instantiated by categorical properties in particular. If all non-structural properties were categorical, the stronger claim would follow from the weak claim. But this is far from obvious, because dispositional properties are arguably both non-structural and non-categorical.

Dispositional properties include causal powers, potencies, or capacities. In the empiricist tradition, dispositions have been thought reducible to abstract relations between categorical properties. The power of charge might be thought reducible to a set of relations such as "will attract particles with opposite charge if they are nearby" and so on. If attraction is a kind of movement, movement is just a set of locations and location is a categorical property, charge is thereby reduced to a conditional relation between categorical location properties. One could also think of how higher level dispositions such as fragility, being disposed to break, are reducible to or explicable in terms of its underlying (relatively speaking) categorical molecular structure. Furthermore, since Hume, causal power in general has been attempted to be reduced to abstract relations between categorical properties. However, on various non-reductionist views, dispositions,

or at least fundamental dispositions such as charge, should be understood as having their own distinct nature.

Irreducibly dispositional properties have a nature distinct from categorical properties in virtue of having something relational about them. Molnar claims that "having a direction to a particular manifestation is constitutive of the power property" (Molnar 2003: 60). The nature of the power of charge consists in being directed against, or having a tendency or striving toward, attracting particles with opposite charge (and so on). Thus, powers are essentially related to things or states outside of themselves or to not yet actual possibilities. But in spite of this relational essence, dispositional properties are also by nature distinct from *abstract* relational properties that make up pure, logico-mathematical (or spatiotemporal) structure. Causal powers may be thought of as characterized by a kind of concrete force or energy, or "oomph", as some philosophers have called it. It is often claimed that this quality is something that we have a primitive or intuitive grasp of. This allows us to distinguish the physical from the mathematical; it is something that may "breathe fire into the equations" of physics, as in Stephen Hawking's phrase.

As mentioned above, even if they are not reducible to relations between the categorical, dispositions or causal powers are often supposed to exist alongside categorical properties, and necessarily so. For example, it may be held that most things have the (allegedly) categorical property of extension in addition to causal powers such as charge. However, according to the dispositionalism of Shoemaker, Bird, and Mumford, *all* properties are really dispositional, including extension (cf. Blackburn, p. 29). This view is also known as power structuralism (Marmodoro undated) or causal structuralism (Hawthorne 2001). According to dispositionalism, the physical world consists of nothing but powers acting on other powers.

Ellis' and Molnar's dispositional essentialism is a slightly weaker view, according to which the power structure requires a background of spatiotemporal location properties, which are classified as categorical, but all other properties are dispositional. But it seems the admission of spatiotemporal categorical properties by dispositional essentialists is not very helpful to panpsychists, given that panpsychism mainly requires that the things *in* space and time are mental, not necessarily space and time itself or the properties in virtue of which things are located in space and time (if one thinks such properties are indeed

required).[34] As mentioned, spatiotemporal properties are good candidates for fundamentally non-mental properties of an impure panpsychism, which some find more plausible than pure panpsychism.

In this way, both dispositionalism and dispositional essentialism constitute a challenge to the argument from categorical properties by contradicting premise 2. I will now consider two candidates for relatively quick fixes. Firstly, one might think this problem somehow results from formulating the argument in terms of categorical properties as opposed to intrinsic properties. While defending that dispositions need categorical grounds is very difficult, perhaps it is easier to defend the following:

> 2**) *Denial of ontological structuralism*: All physical structure must be instantiated by (individuals with) *intrinsic* properties.

The idea here is that categorical and intrinsic are not after all interchangeable terms – there are intrinsic properties which are not categorical, and while dispositional properties might not need *categorical* grounds, they will still need *intrinsic* grounds or properties to accompany them. However, insofar as premise 2** is now taken to rule out dispositionalism and dispositional essentialism, it would no longer be defensible by appeal to the arguments against ontic structural realism alone, given that dispositionalist structure is not uninstantiated or purely abstract or spatiotemporal. Therefore, additional substantive argument would still be needed, and it would not be a quick fix after all.

Furthermore, even if additional arguments for premise 2** could somehow be easily found, another problem is that it would arguably not rule out dispositionalism and dispositional essentialism anyway. As Molnar maintains (2003: ch. 6), irreducible dispositions can plausibly be regarded as intrinsic properties. This is because of the way in which powers can exist unmanifested. An electron might have charge even if it happened to never encounter another particle with charge, and its charge would therefore never manifest as attraction or repulsion. The directedness or tendency toward repulsion/attraction can exist independently of anything outside of it *actually* existing, and existing independently of other things is the criterion of being intrinsic. In other

[34] The admission of categorical properties in virtue of which things are spatiotemporally located would only be of help to panpsychists if coupled with a relationalist view of space or space-time, where all the non-spatiotemporal properties of things ground its location (as was Leibniz's view). This is not how dispositional essentialists think of location properties, however.

words, while the manifesting of a disposition is perhaps not intrinsic, because it is dependent on circumstances outside of it, the disposition itself is intrinsic.

The second potential quick fix is to claim that not only are dispositions intrinsic, they are in fact also categorical. John Heil and C. B. Martin (1999) and Strawson (2008a) defend the view that the dispositional and the categorical are really identical, or that all properties are fundamentally and necessarily both. The distinction between the categorical and the dispositional is to be regarded merely as a distinction between two aspects or ways of looking at one and the same property. If so, dispositionalism will entail premise 2 and not refute it.

However, if the identity view is correct, while it may save premise 2, it clearly does not save the argument. On the basis of the identity view, the challenge from pure powers can be reframed as an equally strong challenge to premise 3 instead. On the condition that we have a primitive grasp of dispositional properties – if we can understand them completely in terms of, say, a non-mental energetic quality – then there would be another, non-mental, categorical property that we know or can positively conceive of.

I conclude that panpsychists, in order to defend the argument from categorical properties, must find a substantive and convincing argument against dispositionalism, dispositional essentialism, and the identity view, or, more precisely, against those versions of the identity view that would allow us to conceive of the categorical via the primitively dispositional and not via the mental. Otherwise, there would be a set of views by appeal to which one could easily refute its premises.

4 OUTLINE OF THE THESIS: CAUSAL SOLUTIONS TO THE PROBLEMS

In the following chapters, I will give arguments that show how both the combination problem and the challenge from irreducible dispositionality can be avoided in ways that start from considering the nature of causation. I will begin with the argument that is relevant to the latter challenge.

In chapter 2, I will present a third argument for panpsychism, an argument from our acquaintance with the nature of causation in agency. I will go through the history of this kind of argument, a history which includes Hume, Leibniz, Schopenhauer, James, Russell and many others. Some of these philosophers have put the argument forth in defense of panpsychism, others as an alleged *reductio* of the premises.

In chapter 3, I will consider the argument from causation systematically. I will show that there is an argument of this type that can be put in a valid form, the same general form as the argument from categorical properties. The two arguments also share many premises. The most significant difference is that whereas the first argument asserts that the only categorical properties we know are mental properties, the second argument asserts that the only dispositional properties we know are mental properties. If this were true, it would pre-empt the challenge to the argument from categorical properties from irreducible dispositionality. *Prima facie*, though, the premise sounds very implausible. Based on considerations highlighted by historical proponents of the argument from causation, as well as some recent work on phenomenology of agency, I will argue that, contrary to initial appearances, there is actually good reason to accept it.

In chapter 4, I will present a number of objections to the argument from causation – some empirical and some more purely philosophical – along with replies to them. One objection will have to do with how the argument makes presuppositions that seem to intensify certain aspects of the combination problem. Another objection will be that the metaphysics of causation that follows from the argument leaves open some difficult questions, questions which are also relevant to the combination problem. They must therefore be answered after I have presented my account of mental combination in chapters 5 and 6.

In chapter 5, I will argue that understanding combination as a causal process enables a solution to the intelligibility aspect of the combination problem. It allows us to escape the dilemma between fully intelligible but seemingly unattainable constitutive grounding and wholly unintelligible brute emergence. I will begin by suggesting that although causation is a relation that is less intelligible than constitution, it is also more intelligible than brute emergence. Then I will propose an account of how causation is partially intelligible in spite of there being in principle not closable epistemic gaps between causes and effects. I will show how a notion of partially intelligible causation along these lines seems presupposed by the argument from non-emergence given by Strawson, as well as by many arguments across the philosophy of mind. It also fits well with science. Then I show how combination can be construed as causal process of the partially intelligible kind.

In chapter 6, I will argue that understanding combination as a causal process enables a solution to the empirical aspect of the combination problem. It allows us to eliminate structural mismatch and to integrate combined macromentality in the causal structure of

the world as science reveals it. I will show that construing combination as a causal process does not entail epiphenomenalism, because the kind of causal process combination will have to be is the kind of process that yields a fusion where the whole supplants its base. I will explain how this does not create a fatal structural mismatch between the mental and the physical even if we assume that the microphysical is causally closed; that is, even if there are no signs of fusions from the point of view of the physical sciences. I will also argue that there is more evidence for physical causal closure than there is for microphysical causal closure, and show how mental combinations construed as fusions would be the perfect structural match for the kinds of emergent behavior that could be most likely found in biological organisms, assuming microphysical causal closure is false.

In chapter 7, I will discuss how the two arguments and their conclusions relate to each other. The two arguments do not depend on each other – one could accept either one without the other – but I will show that the notions of intelligibility in causation they appeal to are fully compatible and to a large extent complementary.

2

The History of the Argument from Causation

As we have seen, panpsychism is today mainly defended by appeal to the Hegelian (against physicalism and dualism) and anti-mysterian (against non-panpsychist Russellian monism) arguments from philosophy of mind and the argument from categorical properties from philosophy of science and metaphysics. These arguments, the Hegelian argument in particular, can subsume many sub-arguments for panpsychism that have sometimes been presented as distinct arguments in the literature, such as arguments from the uniformity of nature and from the continuity of evolution. Other arguments have appeared that are not so subsumable; however, these may often seem incompatible with the naturalistic commitments of two main arguments, e.g., by starting from theological considerations. However, there is (at least) one additional argument for panpsychism that has appeared in the history of philosophy and that is neither clearly incompatible with, nor subsumable under or readily assimilated to any of these two arguments. This is an argument from our acquaintance with the nature of causation in our own agency, which I will refer to as the argument from causation for short.

The argument starts from the notion of causation which was the target of Hume's well-known criticism, the notion that involves natural necessity or power. Some, like Hume, regard these terms as nearly synonymous;[1] others find that it is important to separate them. Unless otherwise specified or implied, I will use the term causation so as to include both natural necessity and powers. Natural necessity or powers are part, or form the ground, of the kind of connection between cause and effect which according to Hume cannot be conceived of with the help of reason alone, and for which no adequate corresponding impression can be found. On the basis of this failure, Hume argues that we should replace the ordinary notion of causation with a notion of constant conjunction, an analysis in terms of regularities and similarity relations.

[1] "[…] the terms of EFFICACY, AGENCY, POWER, FORCE, ENERGY, NECESSITY, CONNEXION, and PRODUCTIVE QUALITY, are all nearly synonymous; and therefore it is an absurdity to employ any of them in defining the rest" (Hume 1739–40/1995: para. 4/36 p. 157).

According to the argument from causation, Hume is mistaken in this conclusion. The most significant premise of the argument is that causation can in fact be directly experienced in situations involving our own agency – and furthermore, that these experiences reveal causation as mental in character. From this, it is inferred that all causation must be of the same mental kind, the result of which is panpsychism.

Today, such an argument would probably strike most philosophers as highly implausible. But this has not always been so. In this chapter, I will give a historical overview – restricted to the early modern times and thereafter – of philosophers who have considered the argument from causation and regarded it as valid. This history will include not only panpsychists who have also regarded the argument as sound, and therefore put it forth in order to promote the conclusion. As I will show, it has frequently been presented also in its *modus tollens* version, as a purported *reductio* of some of the premises. This will make clear that the argument has been of some importance not only for panpsychism but also for reductionism in the metaphysics of causation. It seems that from the very beginning, reductionism about causation was motivated not only by empiricist principles, but also by worries that the ordinary concept of causation was intrinsically "anthropomorphic". I will also consider the relevance of the argument for contemporary debates – about causation, action and, of course, panpsychism.

I will not directly discuss the plausibility of the argument in this chapter. In the next chapter, however, I will give a version of the argument based on the more plausible features of different versions of the argument that will be seen in this chapter and argue that it is actually sound. I will also compare it to the argument from categorical properties and see how it affects the challenge from irreducible dispositionality.

The philosophers who have presented a version of the argument and inferred the truth of panpsychism include G. W. Leibniz, Arthur Schopenhauer, William James, James Ward, Ferdinand Schiller, George Frederick Stout and Charles Hartshorne. Those who regarded some version of the inference valid, but as a *reductio*, include Bertrand Russell, R. G. Collingwood, Ernst Mach, Karl Pearson, Leonhard Euler, Thomas Reid and David Hume himself. Additionally, the *reductio* has reappeared in recent literature on causal powers and dispositions.

1 LEIBNIZ

Gottfried Wilhelm Leibniz (1646–1716) may be the first philosopher in the early modern times to have presented the argument from causation.[2] There is some debate about whether Leibniz was in fact a panpsychist; if this is not his view, then his argument from causation would not be the type I am looking for. However, Leibniz clearly and repeatedly states that all substances are analogous to minds, and that the difference between minds and lower monads is only a difference in degree. He has also given versions of both the main arguments for panpsychism discussed so far, and it is reasonable to think that he did not regard these arguments as entailing a wholly different view. I will present his versions of these arguments, before presenting his version of the argument from causation.

Firstly, Leibniz gives the same kind of arguments against physicalism as discussed above. The famous thought experiment of the mill demonstrates an explanatory gap between the mental and the physical:

> Moreover, it must be confessed that perception and that which depends upon it are inexplicable on mechanical grounds, that is to say, by means of figures and motions. And supposing there were a machine, so constructed as to think, feel, and have perception, it might be conceived as increased in size, while keeping the same proportions, so that one might go into it as into a mill. That being so, we should, on examining its interior, find only parts which work one upon another, and never anything by which to explain a perception. (Leibniz 1714/1925: §17)

Furthermore, in the following passage he can be taken to argue that mentality must be fundamental because its emergence from the non-mental is impossible:

> […] every multitude presupposes *true unities* […] there must be force and perception in these very unities, since without that there would be no force or perception in all that which is made of them, which can only contain repetitions and relations of that which is already in these unities. And thus in bodies which have sensation there must be *unique substances*, or unities which have perception. (Leibniz, Letter to Princess Sophia, June 12th 1700, translated in Garber 2009: 342)

Additionally, Leibniz argues against Cartesian dualism on the basis of a conservation principle, which he seemingly takes to entail physical causal closure:

[2] Francis Bacon was perhaps ahead of him (as remarked by Hartshorne in a passage cited in section 7 below), but he states it without much discussion and very ambiguously.

> Descartes recognized that souls cannot impart any force to bodies, because there is always the same quantity of force in matter. Nevertheless he was of opinion that the soul could change the direction of bodies. But that is because in his time it was not known that there is a law of nature which affirms also the conservation of the same total direction in matter. Had Descartes noticed this he would have come upon my system of pre-established harmony. (Leibniz 1714/1925: §80)

These are the main elements of the Hegelian argument for panpsychism. His pre-established harmony, or parallelism of the mental and the physical, has many similarities with Russellian pure panpsychism.

When it comes to the argument from categorical properties, Leibniz was a pioneer. He was one of the first proponents of a relational view of the physical. Blackburn (see chapter 1, p. 29) mentions Leibniz's reduction of what for Descartes was the fundamental categorical property of the physical, extension, to the dispositional property of resistance. Seager mentions (see chapter 1, p. 27) how Leibniz furthermore regarded all extrinsic properties as reducible to intrinsic properties of substances. Leibniz thinks minds are the kind of things to which all relations can reduce. Via his principle of the uniformity of nature, "things are everywhere and always just as they are in us now" (1704/1997: 205–206), we can infer that the nature of substances in general is mindlike – we can extrapolate from the substance that is our own self to all substances.

A third argument for panpsychism, the argument from causation, also appears in Leibniz's writings:

> The clearest idea of active power comes to us from the mind. So active power occurs only in things which are analogous to minds, that is, in entelechies; for strictly matter exhibits only passive power. (Leibniz 1704/1981: 171)

> I found then that [the nature of substances] consists in force, and that from this there follows something analogous to sensation and appetite, so that we must conceive them on the model of the notion we have of souls. (Leibniz 1695/1989a: 139)

Leibniz's view seems to be that we are immediately and infallibly acquainted with the nature of causation, which he regards as involving active power[3] and force, in willing, acting and thinking. This is made clear in this passage where he argues against occasionalism that *our* causal powers are undeniable:

[3] Active power is opposed to passive power. The former is the power to affect and the latter is the power to be affected.

> Certainly if this doctrine [occasionalism] is pushed to the point of suppressing even the immanent actions of substances […], then nothing in the world appears to be more contrary to reason. In truth, who will question that the mind thinks and wills, and that many thoughts and volitions in us are elicited from ourselves, and that we are endowed with spontaneity? This would be not only to deny human liberty and to make God the cause of evil, but also to contradict the testimony of our inmost experience and of our conscience; through which we feel that those things are ours, which, without any kind of reason, our adversaries would transfer to God. (Leibniz 1698/1908: 126)

Causation, the principle of action, can be shown to be not only real but also intelligible on the basis of inner experience: "[…] this principle of action is most intelligible, since in it there is something analogous to what there is in us, namely perception and appetite" (1699–1706/1989: 180).

For Leibniz, activity is not something we *perceive* with an introspective sense which is analogous to the outer senses. Rather, we grasp it with what he calls the understanding:[4]

> But this indwelling force may indeed be conceived distinctly but not explained by images; nor, certainly, ought it to be so explained any more than the nature of the soul, for force is one of those things which are not to be grasped by the imagination but by the understanding. (Leibniz 1698/1908: 123)

He writes to his correspondent De Volder, who complains that he cannot perceive primary forces (another Leibnizian causal notion): "do you wish to imagine things that can only be understood, to see sounds and hear colors?" (1699–1706/1989: 180)

Leibniz can here be interpreted as claiming that we are directly acquainted with our own activity, as opposed to indirectly, as we are with perceptual objects that are revealed via "images". He links awareness of causation with the intimate awareness he thinks we have of the soul or the self – they seem to come down to the same thing: "And so, to your first question, what is the active principle, I answer in the same way as I would to the question as to what the soul is" (1699–1706/1989: 176, footnote 232).

Why does Leibniz think that everything has causal powers – a premise which is also needed to make the inference? Leibniz is firmly convinced that activity, in the irreducible sense, is the essence of being or substance. He claims that "to act is the mark of

[4] I am not attributing to Leibniz a fundamental distinction (as opposed to a gradual distinction) between perception and understanding. Kant famously claims Leibniz's lack of such a distinction between sensibility and understanding was one of Leibniz's great philosophical mistakes.

substances" (1695/1989b: 118) and that "what does not act does not exist" (1691/1965: 470).

Why does he infer panpsychism instead of that substances in general possess some non-mental kind of causal powers? Leibniz was committed to principles of the uniformity as well as the continuity of nature: "Nothing takes place suddenly, and it is one of my great and best confirmed maxims that nature never makes leaps" (1704/1981: 56). Since he was convinced that the active powers of the mind were real, and seemed to think that non-mental powers would be discontinuous with and distinct from the mental, these principles stop him from positing non-mental powers.

He also appealed to considerations of parsimony – why appeal to new properties when there are old ones that can do the job?

> We impose other things [than perceivers and perceptions, as well as the existence of those things which must be admitted in them] on nature and we then struggle with the chimeras of our mind, as with ghosts. (Leibniz 1699–1706/1989: 184, footnote 239)

> Indeed, everywhere and throughout everything, I place nothing but what we all acknowledge in our souls on many occasions, namely, internal and spontaneous changes. And so, with one stroke of mind, I draw out the entirety of things. (Leibniz 1699–1706/1989: 181)

Leibniz is well known for his view that monads have no windows, which is to say that they cannot be influenced by other monads. He holds that all interaction between monads is really a matter of divine pre-established harmony and not real causation. However, it is only *transeunt* or inter-substantial causation, causation between distinct monads, which Leibniz denies. *Immanent* or intra-substantial causation, the way in which one state of same the monad causes the next, is essential to his system, and this is the kind of activity that is referenced in the quotes given above.

2 SCHOPENHAUER

Arthur Schopenhauer (1788–1860) held a dual-aspect panpsychist view, according to which from one side, the world appears as representation (*Vorstellung*) and from the other as will (*Wille*). It is via representations that we know external objects, but representations only reveal their relational properties. The intrinsic nature of all objects is will. We are only acquainted with this aspect of one object, namely our own bodies, but we can infer

from this one object to the rest of the world and conclude that will is its inner nature throughout.

Schopenhauer links his distinction between representation and will to the Kantian distinction between appearances and things-in-themselves. As mentioned in chapter 1 (p. 30), Langton maps the same distinction onto the distinction between relational and intrinsic properties. Schopenhauer thinks Kant was right in separating appearances from things-in-themselves, but wrong in saying that the latter are unknowable. According to Schopenhauer, we can know the thing-in-itself precisely because we are *it:*

> […] we are not merely the *knowing subject*, but […] *we ourselves* are also among those realities or entities we require to know. […] *we ourselves are the thing-in-itself.* Consequently, a way *from within* stands open to us to that real inner nature of things to which we cannot penetrate *from without.* It is, so to speak, a subterranean passage, a secret alliance, which as if by treachery, places us all at once in the fortress that could not be taken by attack from without. (Schopenhauer 1859/1966b: 195)

Schopenhauer agrees with Kant that we can, in a sense, not *know* the will, the thing-in-itself, insofar as knowing something requires representing it. But while the will cannot be represented, it is still known in a deeper sense, directly and immediately.[5]

Schopenhauer repeatedly states that the will, the thing-in-itself, is not necessarily conscious. This might make it seem as though he is at best a panprotopsychist and not a panpsychist. However, he uses the term consciousness (*Bewusstseyn*) not as we would use phenomenal consciousness, but rather to denote a mode of knowledge. This is made clear at the same time as he denies that Will must be conscious: "This will is not essentially united with consciousness, but is related to *consciousness, in other words to knowledge*, as substance to accident" (Schopenhauer 1859/1966b: 199, my emphasis).

Schopenhauer's argument for his view has much in common with the argument from categorical properties. However, he thinks an intrinsic yet *dispositional* (i.e., causal,

[5] In Volume 2 of *The World as Will and Representation,* Schopenhauer concedes to Kant that while the will brings us closer to the thing-in-itself, it does not bring us into direct contact: "[…] in this inner knowledge the thing-in-itself has indeed to a great extent cast off its veils, but still does not appear quite naked. In consequence of the form of time which still adheres to it, everyone knows his will only in its successive individual acts, not as a whole, in and by itself. Hence no one knows his character *a priori*, but he becomes acquainted with it only by way of experience and always imperfectly. Yet the apprehension in which we know the stirrings and acts of our own will is far more immediate than is any other. […] the act of will is indeed only the nearest and clearest phenomenon of the thing-in-itself […]" (Schopenhauer 1859/1966b: 197).

dynamic) property is what must ground physical structure. Let us see how the argument proceeds.

Schopenhauer would agree with Russell that physics leaves a gap: science, or "pure mathematics, pure natural science and logic", "show us nothing more than mere connexions, relations, of one representation to another, form without any content" (Schopenhauer 1859/1966a: 121).

> Therefore, although all mathematics gives us exhaustive knowledge of that which in phenomena is quantity, position, number, in short, spatial and temporal relation; although etiology tells us completely about the regular conditions under which phenomena, with all their determinations, appear in time and space, yet, in spite of all this, teaches us nothing more than why in each case every definite phenomenon must appear just at this time here and just at this place now, we can never with their assistance penetrate into the inner nature of things. (Schopenhauer 1859/1966a: 121)

He furthermore asserts that inner natures, things-in-themselves, are necessary:

> Now if the objects appearing in these forms are not to be empty phantoms, but are to have a meaning, they must point to something, must be the expression of something, which is not, like themselves, object, representation, something existing merely relatively, namely for a subject. On the contrary, they must point to something that exists without such dependence on something that stands over against it as its essential condition, and on its forms, in other words, must point to something that is *not a representation*, but a *thing-in-itself*. (Schopenhauer 1859/1966a: 119).

Crude materialism is Schopenhauer's term for a view which denies any reality beyond what science reveals, and is in many respects analogous to ontic structural realism. Crude materialism, he claims, entails that:

> […] all content of the phenomenon would have vanished, and mere form would remain. […] finally mere phantom, representation and form of the representation through and through; one could not ask for a thing-in-itself. […] But this will not do; phantasies, sophistications, castles in the air, have been brought into being in this way, but not science. (Schopenhauer 1859/1966a: 123)

In fact, Schopenhauer goes on to claim, science presupposes the notion of *force* as always underlying phenomena and grounding the laws that relate them:

> […] we do not ask why 2 + 2 = 4, or why the equality of the angles in a triangle determines the equality of the sides, or why any given cause is followed by its effect, or

> why the truth of a conclusion is evident from the truth of the premisses. Every explanation not leading back to such a relation of which no *Why* can further be demanded, stops at an accepted *qualitas occulta*; but this is also the character of every original force of nature. Every explanation of natural science must ultimately stop at such a *qualitas occulta*, and thus at something wholly obscure. (Schopenhauer 1859/1966a: 80)

The nature of forces, that which answers the question as to "why any given cause is followed by its effect" remains occult:

> Mechanics, physics, chemistry teach the rules and laws by which the forces of [nature …] operate, in other words, the law, the rule, observed by these forces […]. But whatever we may do, the forces themselves remain *qualitates occultae*. (Schopenhauer 1859/1966a: 122)

> Men tacitly resigned themselves to starting from mere *qualitates occultae*, whose elucidation was entirely given up, for the intention was to build upon them, not to undermine them. Such a thing, as we have said, cannot succeed; but apart from this, such a structure would always stand in the air. What is the use of explanations that ultimately lead back to something just as unknown as the first problem was? (Schopenhauer 1859/1966a: 125)

Schopenhauer argues that the nature of force is not unknown and occult – we know it as will:

> For the purely knowing subject as such, this body is a representation like any other, an object among objects. Its movements and actions are so far known to him in just the same way as the changes of all other objects of perception; and they would be equally strange and incomprehensible to him, if their meaning were not unravelled for him in an entirely different way. Otherwise, he would see his conduct follow on presented motives with the constancy of a law of nature, just as the changes of other objects follow upon causes, stimuli, and motives. But he would be no nearer to understanding the influence of the motives than he is to understanding the connexion with its cause of any other effect that appears before him. He would then also call the inner, to him incomprehensible, nature of those manifestations and actions of his body a force, a quality, or a character, just as he pleased, but he would have no further insight into it. All this, however, is not the case; on the contrary, the answer to the riddle is given to the subject of knowledge appearing as individual, and this answer is given in the word *Will*. This and this alone gives him the key to his own phenomenon, reveals to him the significance and shows him the inner mechanism of his being, his actions, his movements. (Schopenhauer 1859/1966a: 99–100)

Representation only shows us *that* effects follow causes, what Hume would call constant conjunction, but leaves us in the dark as to *why*. When it comes to the connection between our motives and our actions, we have insight into why, or into the nature of the forces that underlie the connection. The Will is the "inner mechanism" of our bodily causation. He infers that there is will wherever there is force:

> [The Will] appears in every blindly acting force of nature, and also in the deliberate conduct of man, and the great difference between the two concerns only the degree of the manifestation, not the inner nature of what is manifested. (Schopenhauer 1859/1966a: 110)

Force is to be understood in terms of will, and not vice versa:

> […] if we refer the concept of force to that of will, we have in fact referred something more unknown to something infinitely better known, indeed to the one thing really known to us immediately and completely; and we have very greatly extended our knowledge. If, on the other hand, we subsume the concept of will under that of force, as has been done hitherto, we renounce the only immediate knowledge of the inner nature of the world that we have, since we let it disappear in a concept abstracted from the phenomenon, with which therefore we can never pass beyond the phenomenon. (Schopenhauer 1859/1966a: 111–112)

All causation must accordingly be understood as being driven by inner motivation:

> Only from a comparison with what goes on within me when my body performs an action from a motive that moves me, with what is the inner nature of my own changes determined by external grounds or reasons, can I obtain an insight into the way in which those inanimate bodies change under the influence of causes, and thus understand what is their inner nature. (Schopenhauer 1859/1966a: 125)

> […] from the law of motivation I must learn to understand the law of causality in its inner significance.
>
> Spinoza (*Epist*. 62) says that if a stone projected through the air had consciousness, it would imagine it was flying of its own will. I add merely that the stone would be right. (Schopenhauer 1859/1966a: 126)

Importantly, Schopenhauer claims that it is in inner motivation we find causal connectedness revealed to us, as opposed to between intentions and physical actions. The will is not related causally to its physical manifestation: "there is actually no causal connexion between the act of will and the action of the body, for they are directly

identical" (Schopenhauer 1859/1966b: 248). A physical action is a representation and will and representation are two aspects of the same reality. The body is the way in which the will makes itself visible, according to Schopenhauer. This can be regarded as a form of Russellian dependency.

As we saw, Schopenhauer rejected reductionism about causation, the elimination of what he calls force, on the basis that science presupposes force, and that abandoning this presupposition would leave the world as an "empty phantom" or "mere form". The latter part of this corresponds well with the Pythagorean *reductio* of ontic structural realism. Positing unknown properties, so-called occult qualities, would still leave physical structure "hanging in the air". But are there other non-mental but still not wholly unknown alternatives? Schopenhauer states that alternatives are inconceivable, and in particular, alternatives that would give the world of representations the same degree of reality as the will would give them:

> For what other kind of existence or reality could we attribute to the rest of the material world? From what source could we take the elements out of which we construct such a world? Besides the will and the representation, there is absolutely nothing known or conceivable for us. If we wish to attribute the greatest known reality to the material world, which immediately exists only in our representation, then we give it that reality which our own body has for each of us, for to each of us this is the most real of things. (Schopenhauer 1859/1966a: 105)

3 JAMES

William James (1842–1910) was committed to panpsychism throughout large parts of his life (see chapter 1, p. 25, footnote 23). His panpsychism seems to some extent motivated from an argument from continuity, which prohibits the diachronic, evolutionary emergence of consciousness from fundamentally non-mental matter (which I discuss in chapter 5, section 3.5). He also considers something akin to the argument from categorical properties: "Our only intelligible notion of an object *in* itself is that it should be an object *for* itself, and this lands us in panpsychism" (James quoted in Perry 1935: 446). Furthermore, he thinks panpsychism follows from his methodological stance, which he calls radical empiricism:

> If empiricism is to be radical it must indeed admit the concrete data of experience in their full completeness. The only fully complete data are, however, the successive moments of our own several histories, taken with their "objective" deliverance or "content." After the

> analogy of these moments of experiences must all complete reality be conceived. (James quoted in Ford 1981: 163)

Finally, he gives a version of the argument from causation:

> […] The concrete perceptual flux, taken just as it comes, offers in our own activity-situations perfectly comprehensible instances of causal agency. […] If we took these experiences as the type of what actual causation is, we should have to ascribe to cases of causation outside of our life, to physical cases also, an inwardly experiential nature. In other words, we should have to espouse a so-called 'pan-psychic' philosophy. (James 1911: 218)

Here, in the introductory textbook *Some Problems of Philosophy*, he endorses the validity of the inference, but only hypothetically entertains the premise that our experience of causation is veridical and not illusory. However, in *The Varieties of Religious Experience*, he explicitly affirms the premise, and adds that this is our *only* experience of causation:

> […] the recesses of feeling, the darker, blinder strata of character, are the only places in the world in which we catch real fact in the making, and directly perceive how events happen, and how work is actually done. (James 1902/1987: 448–449)

He then continues:

> Hume's criticism has banished causation from the world of physical objects, and "Science" is absolutely satisfied to define cause in terms of concomitant change. […] The 'original' of the notion of causation is in our inner personal experience, and only there can causes in the old-fashioned sense be directly observed and described. (James 1902/1987: 449, footnote 1)

Unlike Hume, James does not take the "old-fashioned" notion of causation to be confused and inadequate.

James considers the premise again as part of a lecture on pragmatism and religion:

> Our acts, our turning-places, where we seem to ourselves to make ourselves and grow, are the parts of the world to which we are closest, the parts of what our knowledge is the most intimate and complete. Why should we not take them at their face-value: Why may they not be the actual turning-places and growing-places which they seem to be, of the world – why not the workshop of being, where we catch fact in the making, so that nowhere may the world grow in any other kind of way than this? (James 1907/1975: 138)

His most extensive treatment of the argument is found in "The Experience of Activity" in *Essays in Radical Empiricism.* He starts out with a statement of radical empiricism:

> Everything real must be experienceable somewhere, and every kind of thing experienced must somewhere be real. […] By the principle of pure experience, either the word 'activity' must have no meaning at all, or else the original type and model of what it means must lie in some concrete kind of experience that can be definitely pointed out. (James 1912b: 160)

He points to our own experiences of activity, and claims that:

> If we suppose activities to go on outside of our experience, it is in forms like these that we must suppose them, or else give them some other name; for the word "activity" has no imaginable content whatever save these experiences of process, obstruction, striving, strain, or release, ultimate qualia as they are of the life given us to be known. (James 1912b: 167)

> No matter what activities there may really be in this extraordinary universe of ours, it is impossible for us to conceive of any one of them being either lived through or authentically known otherwise than in this dramatic shape of something sustaining a felt purpose against felt obstacles and overcoming or being overcome. What 'sustaining' means here is clear to anyone who has lived through the experience, but to no one else; just as 'loud,' 'red,' 'sweet,' mean something only to beings with ears, eyes, and tongues. The *percipi* in these originals of experience is the *esse*; the curtain is the picture. If there is anything hiding in the background, it ought not to be called activity, but should get itself another name. (James 1912b: 167–168)

James' argument can be seen to be based on a pragmatically motivated form of meaning empiricism.

He goes on to present an objection on behalf of someone who is not satisfied with activity being identical with the appearance of it, but rather demands that there be an "activity *an sich*" behind the appearances. The objector wonders whether "our feeling [does] more than record the fact that the strain is sustained? The real activity, meanwhile, is the doing of the fact; and what is the doing made of before the record is made" (1912b: 171). James' answer is the following:

> Sustaining, persevering, striving, paying with effort as we go, hanging on, and finally achieving our intention – this is action, this is effectuation in the only shape in which, by a pure experience-philosophy, the whereabouts of it anywhere can be discussed. Here is creation in its first intention, here is causality at work.[footnote omitted] To treat this offhand as

> the bare illusory surface of a world whose real causality is an unimaginable ontological principle hidden in the cubic deeps, is, for the more empirical way of thinking, only animism in another shape. You explain your given fact by your 'principle,' but the principle itself, when you look clearly at it, turns out to be nothing but a previous little spiritual copy of the fact. Away from that one and only kind of fact your mind, considering causality, can never get.[footnote omitted] (James 1912b: 183–184)

Note that James, unlike other philosophers I will later discuss, uses the term animism as something that stands opposed to panpsychism. He might have put "spiritualism" instead, as he appears to intend to refer to something which has similarities with the mind, but is somehow more transcendental or outside the world of appearances. Now, James' answer to the imagined objection seems to be that any notion of activity or causation as it really and transcendentally is in itself, behind the appearances, can only be conceived of on the model of the appearance, and therefore does not actually bring us away from the appearances and into the transcendental after all. This is in line with the meaning empiricism expounded earlier. He concludes on a pragmatist note, by noting that what we should really be interested in are concrete and individual causes:

> I conclude, then, that real effectual causation as an ultimate nature, as a 'category,' if you like, of reality, is just what we feel it to be, just that kind of conjunction which our own activity-series reveal. We have the whole butt and being of it in our hands; and the healthy thing for philosophy is to leave off grubbing underground for what effects effectuation, or what makes action act, and to try to solve the concrete questions of where effectuation in this world is located, of which things are the true causal agents there, and of what the more remote effects consist. (James 1912b: 185–186)

In *Some Problems of Philosophy*, he also notes the distinction between these two kinds of knowledge of causation: "Even so our will-acts may reveal the nature of causation, but just where the facts of causation are located may be a further problem." (1911: 216) Although panpsychism seems clearly implied throughout "The Experience of Activity", it is not mentioned explicitly until the end that these questions lead into the "region of panpsychic and ontologic speculation" (1912b: 189).

4 WARD

James Ward (1843–1925) was a contemporary of James, and like him a psychologist and philosopher. He criticized both the dominant metaphysical theories of his day, materialism on the one hand and absolute idealism on the other, and argued that a form of

panpsychism was a better alternative. Among his arguments for panpsychism, we find a version of the argument from causation:

> If pleasures and pains can be sufficient reasons, they too must be reckoned among the causes that animate nature, or at least among the causes that determine events. […] motives remain as a class of causes not yet admitting of mathematical treatment, still less mechanical interpretation. *De gustibus non est disputandum* here passes from a mere maxim almost into a metaphysical principle. In other words, wherever there is feeling there is something unique. Now, either this uniqueness appears in the physical world or it does not. The admission that it does will make it very difficult to stop short of regarding all the beings that compose the world – so far as 'being' implies any sort of unity or individuality – as feeling-agents, monads or entelechies. (Ward 1915: 172–173)

It is worth noting how Ward focuses on the efficacy of pain and pleasure, as opposed to will and effort. The argument is stated somewhat vaguely and hesitantly, and Ward does not give much more of an elaboration of it. It could perhaps be put more clearly and directly as follows: Pain and pleasure appear to be sufficient reasons for our actions. If they are in fact sufficient reasons, they must be causes. If they are *physical* causes, then all physical causes should be regarded as being capable of feeling, or of experiencing sufficient reasons like pains and pleasure for themselves – presumably on the basis of a principle of uniformity within the physical.

5 STOUT

George Frederick Stout (1860–1944) was another contemporary of James, and a teacher of Russell. He too was a psychologist as well as a philosopher. He argued for a panpsychist position he called animism, motivated to a large extent by a version of the argument from causation.

Stout affirms the following theses about the origin of our concept of causation:

> (1) There is nothing in the immediate content of sense-experience from which the notion of activity and its contrasted correlate, passivity, could be derived. (2) Their source is to be found in what we experience in our own doing and enduring. (Stout 1931: 15–16)

Here he emphasizes, like Leibniz and Schopenhauer, how causation does not appear in sense-experience, but that experience of causation constitutes its own category. He proceeds to explain how the causation in our own agency relates to causation outside of it:

> It is only our own action which we actually experience; but, in being aware of our own agency we cannot help being at the same time aware, or seeming to be aware, of another agency correlated with and opposed to ours within the same total situation. If this were not so, the words 'effort', 'resistance', 'pulling', 'pushing', etc., would lose an essential part of their significance. (Stout 1931: 17)

> In a third class of instances, neither effort not counter-effort are initiated by ourselves. When we perceive a chip swept along by a current, or a bent spring, what is presented as a sensible appearance includes relative position and change of position but nothing which could give to such words as 'force', 'stress', 'strain', 'power', 'energy', or to such phrases as 'swept away by the current', the significance which they bear in ordinary language. It is we who supply this significance by reading into the merely sensible phenomena some analogue of what we ourselves experience in active movement and in making efforts against resistance. (Stout 1931: 17)

On this basis, he claims that an animistic view is embedded in the common sense of "the plain man" and that this is not an "anthropomorphic fallacy" (1931: 33). He claims that:

> We are not warranted in rejecting the animistic position either on the ground that it is obviously absurd, or that it is in principle irreconcilable with the general position of science, or that in the history of mankind it finds expression in special ways which turn out to be untenable. (Stout 1931: 22)

On what grounds are we warranted in affirming it? Stout points to how Hume by rejecting our ordinary notion of causation – which includes what Stout calls "active tendency" – and replacing it by an analysis in terms of constant conjunction, as well as an explanation in terms of mere habitual expectation, came upon the problem of induction. Stout argues that only via animism can we restore the rationality of causal inference:

> If we reject this conception of active and actual tendency, causal inference is not logically inference at all. We have to give up logic for a very dubious psychology. All that we are justified in asserting is that in past instances we have found customary conjunction to be followed by expectation. (Stout et al. 1935: 52)

> I hold then that active tendency is required to fill a logical gap which would otherwise destroy the validity of causal inference. If we have a notion of active tendency that can fill this gap, we are justified in using it. But have we? If so, what is its nature and whence do we derive it? I answer that its source and the key to its nature is to be found and can only be found in the experience we have when and so far as we feel ourselves active. We all distinguish between our doing and what happens to us. (Stout et al. 1935: 53)

> Without [the animistic position] or some substitute for it, there seems no ultimate warrant for inference in matter of fact. Hence a heavy *onus probandu* rests on those how reject it. (Stout 1931: 36)

Stout furthermore considers whether consciousness, by which (unlike Schopenhauer) he appears to mean phenomenal consciousness, always accompanies the kinds of connections we experience in agency. At first, he concludes that we have reason to posit throughout the physical "the operative presence of mind, or at least of an unconscious analogue of mind" (1935: 61). However, he expresses doubt about whether unconscious mental action is coherent:

> But it is worth considering whether this conception of an unconscious analogue of mind and of mental action is tenable at all. To me it seems that it is not. I should say that an unexperienced conation is as absurd as an unexperienced pleasure or pain […]. (Stout et al. 1935: 64)

6 SCHILLER

Ferdinand Schiller (1864–1937) defended panpsychism on pragmatist, or in his preferred terms, Humanist grounds:

> We need not shrink from words like 'hylozoism,' or (better) 'panpsychism,' [...]. For at bottom they are merely form of Humanism, – attempts, that is, to make the human and the cosmic more akin, and to bring them closer to us, that we may act upon them more successfully.

In the history of panpsychism, Schiller is infamous for a passage which was ridiculed by Paul Edwards (1967) in an encyclopaedia entry on panpsychism:

> […] there is a common phenomenon in chemistry called 'catalytic action.' It has seemed mysterious and hard to understand that although two bodies, A and B, may have a strong affinity for each other, they should yet refuse to combine until a mere trace of an 'impurity,' C, is introduced, and sets up an interaction between A and B, which yet leaves C unaltered. But is not this strangely suggestive of the idea that A and B did not know each other until they were introduced by C, and then liked each other so well that C was left out in the cold? (Schiller 1907: 443)

Leibniz was always careful to point out that metaphysical explanations in terms of mental active power were to be confined to its own metaphysical domain distinct from science. We can see that Schiller was not so careful. In other work, however, he argues explicitly

for the metaphysical explanatory power of the mental. If we allow ourselves to be "anthropomorphic", he claims, then we have an answer available to Hume's question about the observability of causation. According to Schiller's analysis, Hume's objections to this proposal (to be discussed below), that the impression to which our ordinary notion of causation can be traced back is our experience of agency, are ultimately question-begging. Hume starts out with epistemic presuppositions that disqualify this proposal before it can be even properly considered:

> We [Humanists] start *ab intra* from the sequences which we most directly experience, and, treating them as typical, logically arrive at the conceptions of causal efficacy and necessary connexion. We admit, of course, that our method is sheer 'anthropomorphism' But then we are Humanists, and know it. You [Humeans] on the other hand only cripple yourself by trying to ignore the human character of your intelligence, and refusing to acknowledge the validity of your immediate experience. You insist on starting *ab extra* from the sequences which you observe in the outer world. You assume, that is, that you can know no more about yourself than about anyone else. (Schiller 1906: 104)

One might note how the distinction between aspects from within and without maps well onto Schopenhauer's distinction between will and representation – according to Schiller, Hume arbitrarily privileges external representations over the inner experience of will and imposes the loose and separate structure of the former on the latter. Schiller goes into further detail about the question-begging nature of Hume's arguments:

> How can anyone, e.g., confute a polemic which begs the point at issue with the superb audacity of Hume's argument in the Appendix to the Treatise? First he professes a desire to find a "perception" on which the causal connexion could be based; then he assumes (1) that "if perceptions are distinct existences, they form a whole only by being connected together;" (2) that "no connexions among distinct existences are ever discoverable by human understanding." Whence it would clearly follow that, even if we had a "perception" of causal connexion, it could not, *ex hypothesi*, serve as a principle of connexion, by the very fact of its being a "perception," and so doomed to remain a distinct and disconnected existence!
>
> Thus the very attempt to prove the existence of activity to those who insist on taking up a point of view from which it cannot be seen, is a mistake. The true retort to their attitude is to show that it is arbitrary, and that better alternatives exist. (Schiller 1906: 108–109)

Schiller reasons that his panpsychist Humanism would be a much better alternative, given that unlike Humeanism (note the extra "e"), it does not result in the "the utter cancellation

of […] agency" (1906: 109), and other consequences that would be disastrous from a pragmatist point of view.

7 HARTSHORNE

Charles Hartshorne (1897–2000) was a process philosopher and a follower of Alfred North Whitehead. He developed an extensive and complex philosophy which was both panpsychist and panentheist. In this passage, he refers to both the Hegelian argument and an argument from the unknown nature of matter in one:

> Psychical monism avoids the most obvious demerits of its two rivals. It is a monism, yet it is not a materialism […] We cannot remain mere dualists, for that means giving up the hope of universal explanatory principles; and we cannot agree upon the materialistic form of monism, not only because it is an attempt to explain away mind, but also because it leaves 'matter' essentially mysterious. (Hartshorne 1977: ch. 3)

He also claimed that panpsychism has an advantage with respect to causation:

> Psychicalism [i.e., panpsychism] has the signal advantage, hinted at by Francis Bacon, that it can construe causal connectedness of events in terms of generalized concepts of memory and perception. Materialism and dualism lack these resources and are in Hume's predicament about causality. Memory and perception are effects whose causes are intrinsically given to them. These are our only clues to the intelligible connectedness of events. (Hartshorne 1977: ch. 3)

In the article "Causal Necessities: An Alternative to Hume", he gives further details on how we can know causation via memory:

> As we shall see, human beings cannot hope ever to find the causal connectedness of the world simply transparent at any point. But they can to some extent grasp the principle of it, if they are able to overcome the habit of regarding psychical conceptions, such as memory, as secondary, merely special cases of something nonpsychical which is more fundamental. Is not a real connection of the present with the past, and of effect with cause, in one type of instance actually constituted by our human memory? The "errors of memory," which all have encountered, need not be so interpreted that they prevent us from recognizing memory as a given connection (for immediate memory is itself immediately remembered) of later with earlier experience. This given-retrospective relation, which I contend is involved in the very meaning of "succession," can, with certain qualifications, be construed prospectively, so that in terms of it the present can be seen both as cause really connected with its effect and as effect really connected with its cause. The future is that which will, and even logically must, have retrospective relation

> to the present we are experiencing, and this prospectively valid retrospective relation, adequately generalized and diversified to cover nonhuman and even nonanimal cases, furnishes in principle what we need by way of real causal connections. This holds, however, only for those who are willing to renounce […] the supposition of merely physical causation. Here Hume is right, the notion of mere matter furnishes no clue to causal connectedness. (Indeed, as Peirce said, it illuminates no problem whatever.) But in discussing psychical causation, Hume neglects memory and even, perhaps, the derivation of ideas from "impressions" for which his own doctrine seems to call. (Hartshorne 1954: 482)

Why does the relation that is represented in memory have to involve mentality wherever it is instantiated? Because the remembering itself, and the subject which remembers, is part of instantiating the succession relation:

> […] to be a conceivable "successor of E" connotes being a conceivable "rememberer of E," at least indirectly via memory of intermediaries one of which itself remembers E. (Hartshorne 1954: 490–491)

Hartshorne mainly focuses, unlike the other philosophers discussed so far, on what it is like to be an effect, i.e., successor, rather than what it is like to be a cause. Agency, or causing, is not out of the picture, however. Memory of agency gives a clear impression of a modal connection:

> Thus, if I have immediate memory of having just made an eager decision to act at once in a certain way, I must now either carry out the decision or else counteract it in some way, whereas without the decision I could neither yield to it nor resist its influence. (Hartshorne 1954: 491)

It is worth noting how Hartshorne affirms that the connection between cause and effect cannot in any sense be discovered either *a priori* or externally, but only *a posteriori* and internally – as in memory.

It is not clear throughout Hartshorne's article, however, at the end he explicitly confirms where he sees it as taking us: "The panpsychic theory, to which our argument leads, almost inevitably meets with resistance […]" (1954: 496).

8 HUME

After having gone through the main proponents of the validity and soundness of the various versions of argument from causation, I will now present some philosophers who

have only affirmed its validity – and then proceeded to reject one or more of its premises at least partly because they assume the falsity of the panpsychist conclusion.

David Hume (1711–1776) is often taken not only as a reductionist about the concept of causation, but also as an eliminativist about natural necessity and power regarded as concrete phenomena. The existence of natural necessity or powers is presupposed by the ordinary, pre-theoretical concept of causation, but these presuppositions disappear from Hume's final analysis. However, Hume scholars such as Strawson (1989) argue that Hume is a skeptical realist, not an eliminativist, about that which grounds or explains regularities in nature – that is, he proposes a reduction of the concept not because he thinks that we can be certain that natural necessity and powers, or causation in the ordinary sense, do not exist, but rather because we would lack epistemic access to these aspects of causation. The term Humean in the context of causation is often taken to entail not only reductionism about the concept but also eliminativism about irreducible powers or natural necessity. But because of this scholarly controversy, I will instead use the terms reductionism and regularity theory for views that entail such eliminativism.

As has already been mentioned (in section 6 on Schiller above), Hume considered, very thoroughly, the suggestion that an impression of necessity was to be found within the mind. In *A Treatise of Human Nature,* he gives a number of objections, which he expands on in *An Enquiry Concerning Human Understanding*. I will return to the plausibility of the argument in view of the objections of Hume and others in the next chapter. Here I will merely note one particular objection which is subtly added in the *Enquiry.* After first expanding on the objections from the *Treatise,* Hume inserts the following footnote:

> It may be pretended, that the resistance which we meet with in bodies, obliging us frequently to exert our force, and call up all our power, this gives us the idea of force and power. It is this *nisus*, or strong endeavour, of which we are conscious, that is the original impression from which this idea is copied. But […] we attribute power to a vast number of objects, where we never can suppose this resistance or exertion of force to take place; [...] to inanimate matter, which is not capable of this sentiment. […] It must, however, be confessed, that the animal *nisus*, which we experience, though it can afford no accurate precise idea of power, enters very much into that vulgar, inaccurate idea, which is formed of it. (Hume 1748/1995: 67, footnote 12)

Surprisingly, Hume here admits that we do have an impression of power, an impression that matches the *vulgar* idea of it. He argues, however, that the impression does not match any *accurate* or *precise* idea. Among Hume's reasons for dismissing the vulgar idea of

power is that he takes it to be in conflict with the view that matter is not capable of sentiment; that is, that he takes the falsity of panpsychism for granted. Thus it seems that he affirms the validity of an inference from acquaintance with causation in agency to panpsychism. It is interesting to note how anti-panpsychist commitments are here revealed as being part of the motivation for reductionism about the concept of causation from its beginning.

9 REID

Thomas Reid (1710–1796) is well known for one of the first counterexamples to Hume's analysis of causation as constant conjunction. Night and day are constantly conjoined, he pointed out, but we do not for that reason think that the night is the cause of the day. However, another of his important objections to Hume's view of causation consists in pointing to what he takes to be a clear and valid impression of power:

> […] in certain motions of my body and directions of my thought, I know, not only that there must be some cause that has power to produce these effects, but that I am that cause; and I am conscious of what I do in order to the production of them. (Reid 1788: 36)

He also holds that this is our *only* basis for the notion: "[…] of the manner in which a cause may exert its active power, we can have no conception, but from consciousness of the manner in which our own active power is exerted" (1788: 37).

He further argues that only conscious beings can have powers:

> If any man, therefore, affirms that a being may be the efficient cause of an action, and have the power to produce it, which that being can neither conceive nor will, he speaks a language which I do not understand. If he has a meaning, his notion of power and efficiency must be essentially different from mine; and, until he conveys his notion of efficiency to my understanding, I can no more assent to his opinion, than if he should affirm, that a being without life may feel pain.
>
> It seems therefore, to me most probable, that such beings only as have some degree of understanding and will, can possess active power; and inanimate beings must be merely passive, and have no real activity. Nothing we perceive without us afford any good ground for ascribing active power to any inanimate being; and every thing we can discover in our own constitution leads us to think, that active power cannot be exerted without will and intelligence. (Reid 1788: 40–41)

Reid considers panpsychism (in the guise of animism) primitive and unscientific.[6] But he is also dissatisfied with reductionism, the view that the behaviour of inanimate objects does not have any causes in the sense that involves active power at all. His preferred alternative is the following:

> With regard to the operations of nature, it is sufficient for us to know, that, whatever the agents may be, whatever the manner of their operation, or the extent of their power, they depend upon the first cause [i.e., God], and are under his control; and this indeed is all that we know; beyond this we are left in darkness. (Reid 1788: 37)

Thus, Reid's rejection of both panpsychism and reductionism, together with his affirmation of our acquaintance with our own causation in agency, leads him to accept a view somewhere in between occasionalism and skeptical realism.

One may view him as implicitly endorsing a *reductio* of the premise that all things have powers *of their own*, on the basis of the unacceptability of panpsychism or animism. The premise is not, however, rejected in favor of the view that things are not moved by any powers at all, as per reductionism; rather, he affirms that non-conscious entities, which are not moved by their own powers, are moved by the powers of other agents, either God or unknown agents that depend on God.[7]

10 ELIMINATIVISTS ABOUT CAUSATION AND NEWTONIAN FORCES

The mathematician and physicist Leonhard Euler (1707–1783) is described by Schopenhauer as arguing for the elimination of force on the basis of its connection with will:

> Further, it is worth noting that Euler saw that the inner nature of gravitation must ultimately be reduced to an "inclination and desire" (hence will) peculiar to bodies (in the 68th letter to the Princess). In fact, it is just this that makes him averse to the conception

[6] "Rude nations do really believe sun, moon and stars, earth, sea and air, fountains and lakes, to have understanding and active power. To pay homage to them and implore their favour, is a kind of idolatry natural to savages. […] When a few of superior intellectual abilities find leisure for speculation, they begin to philosophize, and soon discover, that many of those objects which, at first, they believed to be intelligent and active, are really lifeless and passive. […] As philosophy advances, life and activity in natural objects retires, and leaves them dead and inactive" (Reid 1788: 282–283).

[7] How are we capable of conceiving of the active power that derives from God? Only on the basis of our own experience: "Our notion, even of Almighty power, is derived from the notion of human power, by removing from the former those imperfections and limitations to which the latter is subjected" (Reid 1788: 278).

> of gravitation as found in Newton, and he is inclined to try a modification of it in accordance with the earlier Cartesian theory, and thus to derive gravitation from the impact of an ether on bodies, as being "more rational and suitable for those who like clear and intelligible principles." He wants to see attraction banished from physics as a *qualitas occulta.* This is only in keeping with the dead view of nature which, as the correlative of the immaterial soul, prevailed in Euler's time. (Schopenhauer 1859/1966a: 127)

This account can be seen to accurately match Euler's own statements (see Euler 1833: letters 50, 54 and 68). As Schopenhauer here alludes to, there was much dissatisfaction among scientists in general about the Newtonian concept of force. Newton himself did not, unlike some physicists after him, regard his second law $F = ma$ as an exhaustive definition of force, but as an additional fact about something that was essentially the producing cause of acceleration and whose nature was not known. He says about gravitation that it "must be caused by an agent acting constantly according to certain laws; but whether this agent be material or immaterial, I have left to the consideration of my readers" (Newton quoted in Jammer 1957: 139). And many of his readers did engage in extensive metaphysical and theological speculation about the true nature of force (see Jammer 1957: ch. 8). Others were skeptical of the concept for the same reason. Some skeptics came to see force as essentially anthropomorphic, such as the mathematician Karl Pearson (1857–1936):

> Primitive people attribute all motion to some will behind the moving body; for their first conception of the cause of motion lies in their own will. Thus they consider the sun as carried round by a sun-god, the moon by a moon-god, while rivers flow, trees grow, and winds blow owing to the will of the various spirits which dwell in them. Slowly, scientific description replaces spiritualistic explanation. The idea, however, of enforcement, of some necessity in the order of a sequence, remains deeply rooted in men's mind, as a fossil from the spiritualistic explanation which sees in the will the cause of motion. The notion of force as that which necessitates certain changes of sequences of motion, is a ghost of the old spiritualism. (Pearson quoted in Jammer 1957: 235)

Many scientists who recognized such anthropomorphic elements took this as a reason for adopting a purely relational definition of the concept of force (see Jammer 1957: ch. 11). As a term of physical theory, there are many methodological reasons internal to science for thinking that force should be defined purely relationally – and it is in full accordance with the Russellian view of the physical generally accepted by panpsychists. But whether one should also eliminate metaphysical concepts of causation or force, conceived as what

grounds or enforces physical relations, is an independent question. An anthropomorphic metaphysical concept of force does not necessarily prohibit a relational reduction of the term as it appears in physical theory. One cannot therefore eliminative all concepts of force and causation on purely methodological grounds internal to science – philosophical considerations must be brought in.

Ernst Mach (1838–1916) played an important role in the development of a relational reduction of force. He also held that the notion of causation should be eliminated partly on the basis of its anthropomorphic connotations:

> I hope that the science of the future will discard the idea of cause and effect, as being formally obscure; and in my feeling that these ideas contain a strong tincture of fetishism [i.e., animism], I am certainly not alone. The more proper course is, to regard the abstract determinative elements of a fact as interdependent, in a purely logical way, as the mathematician or geometer does. True, by comparison with the will, forces are brought nearer to our feeling; but it may be that ultimately the will itself will be made clearer by comparison with the accelerations of masses. (Mach 1897: 253–254)

This is somewhat curious, given that Mach was a proponent of a type of neutral monism that seems to borderline on panpsychism – like James in his neutral monist days, he took the neutral substrate to be pure experience. It therefore seems it was only animistic panpsychism, panpsychism that takes causation to be related to agency, that he dismissed.

R. G. Collingwood (1889–1943) was another eliminativist about causation. In his article "On the So-Called Idea of Causation", he claims that:

> In the first sense of the word cause, that which is caused is the free and deliberate act of a conscious and responsible agent, and "causing" him to do it means affording him a motive for doing it. (Collingwood 1937: 86)

This is "historically the original sense [of the word *cause*] […] and remains strictly speaking the one and only 'proper' sense" (Collingwood 1937: 85). From this he concludes that Newton, with his distinction between caused (forced) and uncaused (inertial) motions, effectively advocated a "reduction of physics to social psychology" (1937: 105).

Finally, Bertrand Russell (1872–1970), in his "On the Notion of Cause", argues that:

> […] the word "cause" is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable […]

> The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm. (Russell 1912: 1)

After pointing out a number of problems with applying the notion of cause in science, and on this basis arguing for its elimination, Russell explains that:

> The importance of these considerations lies partly in the fact that they lead to a more correct account of scientific procedure, partly in the fact that they remove the analogy with human volition which makes the conception of cause such a fruitful source of fallacies. (Russell 1912: 9)

Some of the central fallacies that, in Russell's view, result from the analogy between causation and volition are the following:

> (2) "Cause is analogous to volition, since there must be an intelligible *nexus* between cause and effect." This maxim is, I think, often unconsciously in the imaginations of philosophers who would reject it when explicitly stated. […] I do not profess to know what is meant by "intelligible"; it seems to mean "familiar to imagination." Nothing is less "intelligible," in any other sense, than the connection between an act of will and its fulfilment. […]
>
> (3) "The cause *compels* the effect in some sense in which the effect does not compel the cause." […] compulsion is a very complex notion, involving thwarted desire. So long as a person does what he wishes to do, there is no compulsion, however much his wishes may be calculable by the help of earlier events. And where desire does not come in, there can be no question of compulsion. Hence it is, in general, misleading to regard the cause as compelling the effect. […]
>
> (4) "A cause cannot operate when it has ceased to exist, because what has ceased to exist is nothing." […] The mistake in this maxim consists in the supposition that causes "operate" at all. A volition "operates" when what it wills takes place; but nothing can operate except a volition. The belief that causes "operate" results from assimilating them, consciously or unconsciously, to volitions. (Russell 1912: 9–11)

Partly, Russell seems to hold that understanding the notion of cause on the basis of volition is a fallacy because of panpsychist implications – in his point (3), in particular, it seems implicit that if causes compel effects, then desire would have to be involved in all causation. At the same time, he also argues that the analogy with volition leads to what he considers fallacies in virtue of not matching scientific procedure and the content of physics, as he understands it. Later in his career, Russell endorsed neutral monism of a

kind close to panpsychism, but around the time of the publication of "On the Notion of Cause" he was one of the doctrine's staunchest critics.[8]

11 RECENT *REDUCTIOS*

Reductionism about causation, which does not aim to eliminate the concepts of cause and effect, but does eliminate commitment to the existence of powers, natural necessity or other enforcing or producing elements of causation, gradually became orthodoxy in philosophy. But in recent times, non-reductionist theories have seen a revival. In chapter 1 (section 3.2), I discussed the dispositionalist, dispositional essentialist and categorical/dispositional identity views of properties, all of which presuppose irreducible powers and thereby non-reductionism.

Interestingly, along with the return of non-reductionism also came one of the clearest expressions of the *modus tollens* version of the argument from causation. Edward Madden and Peter Hare were early proponents of the revival of causal powers metaphysics. They argue that we should reject the view that causation can be experienced in agency because it will enable an inference from causal powers to panpsychism:

> It is most crucial to avoid what we like to call the "inferential predicament," because getting involved in it forces one inevitably into pan-psychism and animism, an unmitigated disaster in the eyes of a great majority of contemporary philosophers. […] The inferential predicament arises by taking volitional contexts as the only ones in which causal power is directly perceived, and then projecting such experienced power onto objects and events in order to make sense of causal necessities in the physical world. (Madden and Hare 1971: 23)

> The best, and perhaps only, way to avoid the inferential predicament and its pan-psychical consequence is to reject the premise that one is directly aware of causal power only in volitional situations […]. (Madden and Hare 1971: 25)

Madden and Hare proceed to argue for the view that we are rather directly aware of causal power in the external world.

David Armstrong gives an argument based on the same kind of inference, but takes it to indicate the falsity of dispositionalism about causation rather than the view that causation can be experienced in agency:

[8] Russell's first qualified endorsement of neutral monism was in a lecture given in 1918. In 1913, right after the publication of "On the Notion of Cause", he began writing his *Theory of Knowledge*, which contained a number of arguments against neutral monism (Tully 2003).

> […] a disposition as conceived by a Dispositionalist is like a congealed hypothetical facts or state of affairs: 'If this object is suitably stuck, then it is caused (or there is a certain objective probability of its being caused) to shatter.' It is, as it were, and inference ticket (as Ryle said), but one that exists in nature (as Ryle would hardly have allowed). That is all there is to a particular disposition. Consider, then, the critical case where the disposition is not manifested. *The object still has within itself, essentially, a reference to the manifestation that did not occur*. It points to a thing that does not exist. This must remind us of the *intentionality* of mental states and processes, the characteristic that Brentano held was the distinguishing mark of the mental, that is, their being directed upon objects or states of affairs that need not exist. This intentionality of the mental undoubtedly exists. But for physicalists such as myself it presents a *prima facie* problem. If the mental has intentionality, and if, as Brentano thought, it is also *ontologically irreducible*, then there is something here that would appear to falsify Physicalism. Physicalists about the mind are therefore found trying to give some ontologically reductive account of the intentionality of the mental. But if irreducible dispositions and powers are admitted for *physical* things, then intentionality, irreducible intentionality, has turned up in everything there is.
>
> Is this not objectionable? Does it not assimilate the physical to the mental, rather than the other way around? (Armstrong 1997: 79)

Intentionality is the manner in which thoughts, intentions, actions and other mental phenomena can be *about* or directed toward other things. Armstrong points to how dispositionality and intentionality share the feature of directedness. He does not, however, point to any experiences of either intentionality or dispositionality; rather, he compares their concepts. In his view, if causation is grounded in irreducible governing laws instead of dispositions the problem is avoided, so he does not take it as an argument for reductionism.

C. B. Martin and Karl Pfeifer also consider a similar *reductio*:

> We will show that the most typical characterizations of intentionality, including those discussed by W. Lycan in his article "On 'Intentionality' and the Psychological," and also Lycan's own suggested characterization and John R. Searle's more extended treatments of the concept all fail to distinguish intentional mental states from non-intentional dispositional physical states. Accepting any of these current accounts will be to take a quick road to panpsychism! (Martin and Pfeifer 1986: 531)

> Somewhat ironically, if we were to leave our discussion at this point, someone might interpret it as an argument for panpsychism, in that the characterizations of intentionality that we have discussed apply to anything (mental or physical) that has causal dispositions. For some, this may be a happy result – for us it is a *reductio ad absurdum* and an

> invitation to look elsewhere for an account of the intentional. (Martin and Pfeifer 1986: 551)

Martin and Pfeifer only take this as a *reductio* of a certain account of intentionality. Like Armstrong, they do not take it as something that casts into doubt any acquaintance we may have with the phenomena of either intentionality or agency via experience. Unlike Armstrong, they think dispositionalism can be made compatible with physicalism by revising the account of intentionality.

Armstrong's and Martin and Pfeifer's arguments in particular make evident the relevance of the argument from causation for the challenge to the argument from categorical properties discussed in chapter 1 (section 3.2), the challenge from irreducible dispositionality. If it could be established that Armstrong is right in endorsing the inference, and Martin and Pfeifer are wrong in rejecting it, then dispositionalism would entail panpsychism, and this challenge to the argument for panpsychism from categorical properties would be entirely pre-empted. In recent times, however, the argument from causation does not seem to have been put forth in support of panpsychism.

This concludes my historical review of the argument for causation. Before considering the plausibility of the argument systematically, in order to build some more motivation for undertaking this, I will briefly look at the relevance of the argument to present debates about causation and agency and its compatibility with contemporary panpsychism.

12 THE PRESENT RELEVANCE OF THE ARGUMENT

As noted, non-reductionism about causation seems to be on the rise.[9] A reasonable guess would be that this has something to do with the fact that empiricism – of the kind that motivated philosophers to look for an impression of causation lest it be eliminated – for various reasons no longer has the influence that it once had. However, as I have shown, not only empiricism, but also a wish to avoid anthropomorphism and panpsychism has played a role in the rise of reductionism. Therefore, non-reductionists who are not panpsychists should be interested in confronting the problem constituted by the argument from causation.

[9] At least it has been on the rise in the last decades. A survey conducted by David Chalmers and David Bourget (2013) shows that philosophers who identify as non-Humeans about the laws of nature outnumber those who identify as Humeans by 2:1, but the gap is smaller among younger people.

Versions of one central premise of the argument, that causation is experienced in agency, find support among a number of non-reductionists of causation. Huw Price claims that "we know causes by knowing what it is like to be an agent" (1991: 173). David Armstrong finds that "it is plausible to suppose that we have […] non-inferential awareness of the successful operation of our will" (1997: 215), and "in Humean terms, there is an impression from which we derive the idea of causality" (1997: 217). Stephen Mumford and Rani Lill Anjum hold that "causation can be perceived within one's own body" (2011b: 209), and that "we all have experience of dispositionality at work, through the exercise or our own powers and the action of other powers upon us" (2011a: 380).

Additionally, in philosophy of action, much attention has recently been devoted (following earlier relative neglect) to our phenomenology of agency, i.e., what it is like to be an agent. Sometimes it is claimed that the phenomenology of agency is accurately described as experience of causation. Carl Ginet argues that actions can be analyzed as mental events with an "actish phenomenal quality" that *seems* to represent productive causation by the subject:

> A simple mental event is an action if and only if it has a certain intrinsic phenomenal quality, which I've dubbed the "actish" quality and tried to describe by using agent-causation talk radically qualified by "as if": the simple mental event of my volition to exert force with a part of my body phenomenally seems to me to be intrinsically an event that does not just happen to me, that does not occur unbidden, but it is, rather, as if I make it occur, as if I determine that it will happen just when and as it does. (Ginet 1997: 87)

Terry Horgan similarly describes the phenomenology of agency as follows:

> […] the actional phenomenal dimension […] is the what-it's-like of self as source of the motion. […] You experience the bodily motion as generated by yourself.
>
> The language of causation seems apt here […]: you experience your behavior as caused by you yourself […] Metaphysical libertarians about human freedom sometimes speak of "agent causation" […] and such terminology seems phenomenologically apt regardless of what one thinks about the intelligibility and credibility of metaphysical libertarianism. (Horgan 2011: 79)

Tim Bayne and Neil Levy, although expressing skepticism about whether we have experiences with causal content, nevertheless claim that "in exerting effort it feels as though one is exerting a power of one's own against a force" (2006: 60), and that "the experience of effort involves an experience of the self as a source of force" (2006: 63).

They present empirical evidence that the feeling of effort reliably tracks physical energy expenditure.

It should be noted that the thesis that causation is experienced in agency is different from the stronger thesis that it is also experienced as *mental*, which is required as a further premise or as an additional aspect that must be added to it if the argument is to have a chance of being valid (as will be discussed in the next chapter). Still, it clearly seems that these two important (though insufficient) premises or prerequisites of the argument from causation, non-reductionism about causation and the thesis that we at least seem to experience causation (or power) in agency, have stronger positions in philosophy at present than they have had in a long time. Furthermore, according to the arguments presented in chapter 1, the conclusion of the argument, panpsychism, can no longer be simply dismissed as absurd. Therefore, taking the argument as a *reductio* of the premises is less plausible.

Although I gave the example of one philosopher, Mach,[10] who had panpsychist sympathies but nevertheless seemed to immediately reject an animistic version of it, contemporary panpsychism seems to be more than compatible with the argument. Russellian monism asserts that mental properties ground physical dispositions, and this is just what the panpsychists discussed in this chapter argue. It fits Strawson's broad statement that "energy is experientiality; that is its intrinsic nature" (Strawson 2006a: 234) and does not directly entail much more. Primitive animism, which asserts that things like the wind, the forest and the sun are moved by their own spirits, would not follow. The main problem with primitive animism is its proto-scientific nature: it professes to explain and predict nature with reference to indwelling spirits and is thereby incompatible with physical causal closure. The kind of animistic panpsychism the arguments reviewed here point to would – or at least could – be Russellian, and thereby respect physical causal closure and the autonomy of science. Furthermore, contemporary panpsychism does not generally attribute mentality directly to things like the wind and the forest, like primitive animism tends to. But nothing is directly implied by the argument from causation about what kinds of things, other than ourselves, should be regarded as having mentality directly, or being agents.

[10] Unlike Russell, Mach was committed to pure experientialism and neutral monism at the same time as he endorsed the *reductio.*

As I have shown, the validity of the inference has been recognized by philosophers from many different philosophical traditions: rationalism, German idealism, pragmatism, empiricism (radical and classic) and process philosophy. Therefore, it cannot be dismissed as an artifact of some particular dated methodological stance.

In conclusion, therefore, there is much that speaks in favor of reconsidering the argument for panpsychism from our acquaintance with the nature of causation in agency. Its premises and its conclusion are interesting and relevant for present debates, and they cannot be immediately judged as devoid of plausibility.

3

A New Defense of the Argument from Causation

In this chapter, I will try to formulate the most plausible version of the argument from causation – an argument for panpsychism that, as shown in the previous chapter, has been put forth by Leibniz, Schopenhauer, James and many others. I will also try to formulate it in a way such that it can be seen to pre-empt the challenge to the argument from categorical properties, which I discussed in chapter 1 (section 3.2), stemming from the possibility of irreducibly dispositional properties. I will consider some alternative formulations, and arrive at one where all the premises – except for one – are either easily seen to be plausible or will be defensible on the basis of what has already been discussed in chapter 1. Then I will give a new defense of the final premise, the thesis that we experience causation in agency in a way that reveals it as mental in nature – and that there is no other kind of causation whose nature we know or of which we have a positive conception.

1 FIRST FORMULATION

All the arguments presented in the previous chapter can be seen to fit the following general form:

> I) *Non-reductionism*: All physical things have irreducibly causal properties.
>
> II) *Mental causation*: The only irreducibly causal properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> III) *Anti-mysterianism:* The (nature of the) irreducibly causal properties of physical things are knowable or positively conceivable.
>
> Therefore,
>
> IV) *Panpsychism*: All physical things have mental properties.

The argument is put in terms of *irreducibly* causal properties because these are the kinds of properties that Hume denied we can experience, and it is experiences of these kinds of properties to which the arguments discussed in the previous chapter all refer. The experiences were not claimed to reveal anything about the nature of causal properties as reductionists would construe them, and cannot therefore directly support that reducible causal properties are mental.

Dispositional properties are a type of causal properties. If the only causal properties we know are also dispositional properties, it would clearly pre-empt the challenge to the argument from categorical properties from irreducible dispositionality. Dispositionalism, dispositional essentialism and the identity view would entail panpsychism. In order to make this evident, the argument can instead be formulated in terms of dispositional properties:

> I*) *Non-reductionism*: All physical things have irreducibly dispositional properties.
>
> II*) *Mental dispositionality*: The only irreducibly dispositional properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> III) *Anti-mysterianism:* The (nature of the) irreducibly dispositional properties of physical things are knowable or positively conceivable.
>
> Therefore,
>
> IV) *Panpsychism*: All physical things have mental properties.

Non-reductionism about causation comes in two main forms: those who posit irreducible governing laws (e.g., Armstrong) and those who posit causal powers, i.e., dispositions, belonging to individual things (e.g., Shoemaker, Ellis, Bird and Mumford). Formulating premise I* in in terms of dispositional properties rules out the governing laws view whereas the first formulation, in terms of causal properties, did not. Armstrong holds that singular events of causation, such as particular exertions of power or force, are *a posteriori* identical with instantiations of laws (1997: 218), and that laws are relations between categorical universal properties. Thus, he reduces dispositions to relations between categorical properties, although his conception of categorical properties and their

relations is not reductionist. An interpretation of the experiences of causation that the argument relies on as revealing Armstrongian causal properties will be problematic, however, because of how Armstrong's reduction of dispositions to instantiations of general laws proceeds via *a posteriori* identification. In Armstrong's view, the fact that the forces or dispositions he agrees that we appear to directly experience in agency are really instantiations of laws is something that can only be understood on the basis of empirical information about regularities and metaphysical reasoning. Reducing mental properties (which according to the argument includes experienced powers) via *a posteriori* identity is possibly in tension with arguments against type-B physicalism that are essential to the case for panpsychism from philosophy of mind (see chapter 1, p. 14–16). Furthermore, if the true nature of causation can only be known via *a posteriori* reduction of the powers we experience, then it seems we are not really directly acquainted with the nature of causation via experience after all. Since Armstrong's conception of causation therefore can be ruled out as an interpretation of the kind of experiences the argument will centrally rely on anyway, it is indirectly justified to rule it out already in the first premise, by putting the argument in terms of dispositional properties.

I will now consider the plausibility or defensibility of the individual premises of this argument. Then I will consider whether a better formulation of the argument is available with more easily defensible premises.

Premise II* – *Mental dispositionality* – in this formulation of the argument would be the least easily defensible. As I have shown, the view that we *seem to* experience dispositional causation in phenomenology of agency has the support of some present philosophers, however, even if they are right, this is merely a prerequisite for an argument in defense of premise II* – our experience also has to be veridical, reveal dispositionality as mental, and not be rivalled by other experiences or conceptions of dispositionality as non-mental. However, it is clearly essential to any argument that is going to reflect the views of the philosophers of the previous chapter, and a way of defending it must therefore be found. But before going into what can be said in defense of this premise, I will try to find the most plausible form of an argument for panpsychism that relies on it.

Premise III – *Anti-mysterianism* – is a counterpart to the anti-mysterian premise 4 of the argument from categorical properties, and a closely related principle ((viii) on p. 26) is part of the argument from philosophy of mind, as discussed in chapter 1. There I claimed that this principle must be justified on pragmatic, methodological or anti-skepticist grounds. Is this sufficient justification for premise III as well? The premise

rules out skeptical realism about causation, the view that irreducible causation exists, but that its nature is unknown. As mentioned in the previous chapter (section 8), this view has been attributed to Hume, and is also regarded by many (e.g., Strawson 1987) as highly defensible on its own terms. Therefore, it might seem unreasonable that premise III rules it out. However, it is consistent with the *motivation* for premise III to say that *if* it were true that we have no epistemic access to the nature of causation, skeptical realism would be preferable over eliminativism about irreducible causal powers or laws, i.e., ontological reductionism about causation, but in view of premise II, according to which we do have epistemic access to the nature of at least one kind of causation after all, skeptical realism is no longer justified. As with the argument from categorical properties, this reasoning could be made explicit, by replacing premise III with the methodological principle that, *if possible*, we should avoid positing properties that (or whose nature) are unknowable or not positively conceivable, as well as making some further modifications in order to restore the validity of the argument (cf. chapter 1, section 2.2, p. 37–38).

Premise I* – *Non-reductionism* – is more easily defensible than premise II*, though it is far from obviously true. As mentioned, non-reductionism about causation has many present supporters. However, trying to argue for non-reductionism over reductionism on a general basis would be a very extensive project, which I will not attempt here. Nevertheless, I will briefly review some different arguments for non-reductionism that have already been alluded to as part of the arguments presented in the previous chapter. This will be partly in order to show why it makes sense to try to find another basis for defending the argument, but also partly in order to clarify the theoretical role irreducible causation is often supposed to play, which will be helpful later on. I will then show that premise I* can instead be discharged by a reformulation of the argument, so that one would not need to go into the debate on a general basis (as opposed to specifically on the basis of the experience of agency) after all.

*1.1 Arguments for Premise I** – Non-Reductionism

The philosophers of the previous chapter motivate non-reductionism in various ways. Schopenhauer holds that forces are required in order to answer the question of why the laws of nature hold. Stout claims that causal powers (or in his terms, active tendencies) are required in order to restore the rationality of causal inference, i.e., induction. Leibniz, Schopenhauer, James and Schiller argue that we know (or have pragmatic justification) from our experience that we have causal powers, and via a principle of the uniformity of

nature or parsimony it follows that all causation involves causal powers as well. Finally, Schopenhauer also claimed the physical would be reduced to a mere relational abstraction, a castle in the air, if not underpinned by forces. Although these philosophers may intend these strategies to lead all the way to panpsychism, they can also be considered as strategies for establishing only some form non-reductionism. As strategies for establishing non-reductionism only, they all seem to have some plausibility. The first three of them, at least, have contemporary proponents.

The first strategy of Schopenhauer's has been taken up by Strawson (1987: 143–144). Strawson argues that we should posit what he calls objective forces, even if their nature is entirely unknown and unknowable, on the basis that there has to be something that explains the regular behavior of objects. Reductionists deny that the fundamental laws or regularities of nature, those who cannot be explained in terms of underlying regularities, have explanations. The basic regularities, according to reductionism, are brute facts that do not obtain for any deeper reason – it is just the way the world is. But, according to Strawson, that the world should be regular for no reason is deeply implausible:

> There are all those mind-independent physical objects knocking about out there completely independently of us, persisting and interacting, and nothing – *nothing* – governs or orders the ways in which they do this. And yet they persist and interact in a highly regular fashion. How can this be so? It is really rather extraordinary. For, *ex hypothesi*, nothing constrains them to behave and interact in a way that exhibits any order or regularity at all […] it seems that [reductionists] assert positively that all causation actually *is* an outrageous run of luck. (Strawson 1987: 265)

This objection does not only target reductive theories of causation formulated in terms of necessary and sufficient conditions, which can be regarded as refinements of Hume's original analysis in terms of constant conjunction. If the objection is effective against such theories, it would also be effective against Lewisian counterfactual theories. Lewis analyses causation in terms of counterfactual conditionals whose truth conditions do not include any irreducible necessities, only regularities and abstract similarity relations. The essential difference is that on the counterfactual theory, the relations extend across possible worlds as well as the actual. Although it will not be mysterious on the counterfactual theory that there are perfectly regular possible worlds, nothing clearly explains why the actual world happens to be regular.

Helen Beebee (2006) and Nicolas Everitt (1991) criticize this argument on the basis that positing forces to explain regularities will just be an arbitrary extra step in a

potentially infinite explanatory regress – at some point we have to accept brute facts. According to Everitt, the only way of ending the regress is by positing a necessary being, i.e., God, which most naturalists including panpsychists would like to avoid. Therefore, it makes sense to regard the ultimate empirical facts as brute facts, and not infer metaphysical qualities beyond them that would not be self-explanatory, and thereby finally end the regress, anyway. It might be retorted, for example, that this is overly verificationist, or that causal powers can be conceived of as having some special regress-ending properties; however, one could see how the debate would be hard to settle.

In accordance with Stout's strategy, both Armstrong (1983: 52–59) and Ellis (2010) have argued that positing irreducible causation (grounded in governing laws and dispositions respectively) solves the problem of induction. The reasoning is an extension of Strawson's: if irreducible causation is the reason for the world's regularities, and reductionists think that there is no such reason for the regularities observed so far, then neither would there be a reason for regularities to continue in the future. In other words, if nothing constrains the world to behave in an orderly fashion, it might as well stop behaving regularly at any moment.

The problem is the same on the counterfactual theory: if there is only one, or a few, possible worlds where our laws of nature hold without exception, but there are many possible worlds where things are regular up to this time, but turn chaotic in various ways afterwards, then it seems more likely that our world will turn out to be one of the suddenly chaotic ones, given that there are more ways of being chaotic (or first orderly and then chaotic) than there are ways of being orderly and each of these ways will correspond to a possible world.

One might object that the existence of necessary connections between causes and effects is surely not *sufficient* for induction to be rational. For example, the existence of necessary connections does not rule out, on most understandings of the nature of causal necessity, that the universe suddenly be annihilated for no reason or cause internal to it. However, it seems causal necessary connections are still a *necessary* condition for the rationality of induction – or at least a necessary part of a jointly sufficient but unnecessary set of conditions. But according to Beebee (2011), reductionists may have a distinct sufficient basis (or a necessary part of a distinct sufficient basis) for handling the problem of induction as well. Therefore, this consideration is not clearly decisive either.

The strategy of justifying non-reductionism on the basis on the experience of agency is pursued by Michael Esfeld. He claims that:

> [Humean] metaphysics is at odds with our experience as acting beings (agents) in the world and the conceptualization of this experience in folk psychology. That experience – and its folk psychological description – is veridical if and only if there is a causal relation between mental intentions and behaviour such that the intention brings about the behaviour in the sense that the intention makes it that the behaviour in question exists – given certain favourable background conditions in the body of the person and in the environment. (Esfeld 2007: 211)

> In short, if and only if there is a glue that ties the behaviour to the intention such that the intention necessitates the behaviour (given the mentioned background conditions), then we are agents – instead of simply undergoing contingent sequences of mental and behavioural property tokens that happen to satisfy certain regularities in virtue of what is going on elsewhere in space-time.
>
> Consequently, Humean metaphysics cannot admit our experience as acting beings (agents) in the world as being veridical. There is no room for agents in a Hume world. (Esfeld 2007: 212)

Since this argument for non-reductionism is premised on the thesis that we veridically experience causation in agency, it already presupposes an important aspect of premise II* – though not all of it, because it additionally says that we experience causation *as* mental. If premise I* can be defended on the basis of premise II*, this would obviously be an advantage, but it also suggests that the argument should be reformulated in order to accurately reflect the number of distinct theses the conclusion actually depends on.

As an alternative to pursuing any further a direct defense of premise I*, I will now suggest a different formulation of the argument, that on the basis of the discussion in chapter 1 would let premise I* be discharged. This alternative formulation is suggested by Esfeld's argument, but also, more directly, by the final strategy I credited to Schopenhauer and which I have not yet considered.

2 SECOND FORMULATION

The final strategy of Schopenhauer's (which can also be read as a strategy of Leibniz's) toward establishing the need to posit ubiquitous dispositional properties is his claim that the driving force of will must underlie physical structure in order for it to be real and not just an abstraction or an empty form. This brings to mind the formally similar, but in terms of content apparently completely opposite, premise 2 of the argument from categorical properties, which I suggested be formulated as follows:

1) *Structuralism about physical theory:* Science only tells us about the structure of the physical.

2) *Denial of ontological structuralism*: All physical structure must be instantiated by (individuals with) categorical properties.

3) *Mental categoricity:* The only categorical properties (whose nature) we can know, or positively conceive of, are mental properties.

4) *Anti-mysterianism:* The (nature of the) properties that instantiate physical structure are knowable or positively conceivable.

Therefore,

5) *Panpsychism:* All physical structure is instantiated by (individuals with) mental properties.

Schopenhauer's justification for why dispositional properties must be ubiquitous suggests a parallel argument from dispositional properties:

1) *Structuralism about physical theory:* Science only tells us about the structure of the physical.

2') *Denial of ontological structuralism*: All physical structure must be instantiated by (individuals with) irreducibly dispositional properties.

3') *Mental dispositionality:* The only irreducibly dispositional properties (whose nature) we can know, or positively conceive of, are mental properties.

4) *Anti-mysterianism:* The (nature of the) properties that instantiate physical structure are knowable or positively conceivable.

Therefore,

5) *Panpsychism:* All physical structure is instantiated by (individuals with) mental properties.

This formulation has many advantages. Although the soundness of the argument from categorical properties may be doubted, on the basis of the challenge from irreducible dispositionality in particular, there is no clear reason to think it is invalid. If it is indeed valid, then the argument from causation construed as a parallel argument from dispositional properties should be equally valid, having the same form.

This way of formulating the argument should also make clear how the challenge to the argument from categorical properties is pre-empted. If dispositional properties instead of categorical properties ground physical structure, but the dispositional properties we should posit are mental, the conclusion, Russellian panpsychism, still follows.

Furthermore, the connection between these two arguments can be exploited in order to render superfluous the type of arguments needed to establish non-reductionism, i.e., premise I* of the previous suggested formulation. If we assume that the premises of the argument from categorical properties, except premise 2, are adequately defensible, then we have a shortcut to the truth of the disjunction of premises 2 and 2'. If we reject pure structuralism, the view that physical structure can exist wholly uninstantiated or be self-instantiating, then it seems at least one of them (and possibly both), must be true – physical structure must be instantiated by categorical *or* dispositional properties.

The rejection of pure structuralism is, as discussed in chapter 1, defensible on the basis of the Pythagorean *reductio* and arguments derived from Newman's problem. It follows from these arguments that physical structure must be instantiated by non-structural properties. As I discussed, on the basis of dispositionalism and dispositional essentialism, there is reason to doubt that all non-structural properties are categorical properties. However, there is good reason to believe that all non-structural properties are either categorical or dispositional (or both).

Firstly, some simply define categorical properties as non-dispositional properties and dispositional properties as non-categorical properties. According to such a definition, it would be an analytic truth that all properties are dispositional or categorical. One might think it is implicit in the definition that both categorical and dispositional properties are also concrete properties, but all non-structural properties should be regarded as concrete properties (i.e., non-abstract, non-logico-mathematical) anyway, so this makes no difference. According to some views, however, such as the identity view defended by Heil, Martin and Strawson, categorical and dispositional properties are not mutually exclusive. Obviously, though, this view still entails that all non-structural properties are categorical or dispositional because they would be both, and the "or" should be taken as

inclusive. There are many other theories of properties that could also be considered, but all the theories which seem to be in discussion at the moment entail that all non-structural properties are categorical or dispositional.[1]

We can therefore combine both arguments into one – an argument from non-structural properties:

> 1) *Structuralism about physical theory:* Science only tells us about the structure of the physical.
>
> 2*) *Denial of ontological pure structuralism*: All physical structure must be instantiated by non-structural properties.
>
> 3a) *Mental categoricity:* The only categorical properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> 3b) *Mental dispositionality:* The only dispositional properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> 3c) *Exhaustiveness:* All non-structural properties are categorical or dispositional (or both).
>
> 3d) *Non-vacuity:* There are some non-structural properties that we know or have a positive conception of.
>
> 4) *Anti-mysterianism:* The (nature of the) properties that instantiate physical structure are knowable or positively conceivable.
>
> Therefore,
>
> 5) *Panpsychism:* All physical structure is instantiated by mental properties.

In this argument, premise 3' (from the argument from dispositional properties) is renamed 3b, and premise 3 (from the argument from categorical properties) is renamed 3a. I also leave some of the parenthetical qualifications in the previous arguments implicit. The

[1] All the views I have seen in discussion fall within one of these categories: dispositionalism, the identity view, the mixed view, according to which there are some purely categorical and some purely dispositional properties, and categoricalism, of which both Armstrong's view and the Humean/Lewisian view are varieties.

qualification of dispositional properties as irreducible has become superfluous, because if dispositions are reducible, they would be reducible to relations between categorical properties, and therefore still mental according to 3a. The justification for each of the premises can be summarized as follows:

Premises 1, 3a, and 4 are identical to premises from the argument from categorical properties, which I here assume are adequately defensible on the basis of the arguments discussed in chapter 1. Or, to be precise, I assume that premise 3a is defensible except insofar as it is threatened by a version the identity view according to which the categorical can be understood in terms of the dispositional, and the dispositional can be positively conceived as a non-mental energetic quality of some kind (see chapter 1, p. 55). If premise 3b is true, it pre-empts this threat.

Premise 2* is a weaker version of premise 2 in the argument from categorical properties. It rules out only the kind of structuralism that is arguably indistinguishable from Pythagoreanism, as well as kinds that would suffer similar problems as Pythagoreanism, such as spatiotemporal structuralism.

Premise 3c is either an analytic truth or a consequence of every theory of properties that seems to be in discussion at present.

Premise 3d is added to specify that at least one of the premises 3a and 3b must be non-vacuously true. In the argument from categorical properties, it is presupposed that premise 3 (now 3a) is non-vacuously true, that we know *some* categorical properties. The same is the case for premise 3' (now 3b) in the argument from dispositional properties: it presupposes that we know some dispositional properties. In the argument from non-structural properties, it is sufficient that only one of the premises is non-vacuously true, so the non-vacuous truth of both these premises *need* not be presupposed. If the argument is supposed to be stronger than either of the previous arguments, it must be allowed that one of them is only vacuously true, and so non-vacuous truth *should* not be presupposed.

The argument claims that a vast class of known properties are mental, both categorical and dispositional properties. Can this be seen as a kind of *reductio*? Does it say, in effect, that really all we know are our own minds, with no access to the external, physical world? This, one might think, must be based on a solipsistic error.[2] The response to this worry would be to reemphasize how it is only the non-structural known properties which are mental. Structural properties might include both mathematico-logical and

[2] Thanks to David Chalmers for pressing this objection.

spatiotemporal properties. According to a number of Russellian monists and ontic structural realists, many of whom have no sympathy with panpsychism, this adequately accounts for our knowledge of the external, physical world. Some structuralists about the physical are causal or dispositional structuralists, a position this argument would seem to rule out, but not everyone. Furthermore, as I will discuss in chapter 4 (section 2.6), there is reason to think that dispositions in a sense have a non-mental *aspect*, which I will argue might be compatible with the premises of this argument, because it would entail that dispositional properties are partially mental, even if not wholly. This would leave room for a sense in which there can be also dispositional structure after all.

If all this is granted, then this so far looks like a solid argument for panpsychism – except for the fact that the justification for premise 3b is still missing. Given the challenge from irreducible dispositionality to premise 3a, premise 3b must be non-vacuously true. I will now give a defense of this premise.

3 THE *MENTAL DISPOSITIONALITY* PREMISE

Premise 3b, *Mental dispositionality*, says that the only dispositional properties we know – or have a positive conception of – are mental properties. One possible argument for the premise would be to invert the *reductio* that philosophers most recently have worried about, the inference considered by Armstrong and Martin and Pfeifer (see chapter 2, pp. 85–85). That is, one could argue that our accounts of dispositionality and intentionality necessarily match and that intentionality is a mark of the mental.

As mentioned, Martin and Pfeifer argue that the account of intentionality can be revised, so that it no longer matches the account of dispositionality. Bird (2007: 114–129) argues that dispositionality does not share the characteristics of intentionality according to common accounts in the first place. U. T. Place (1996) and Molnar (2003), on the other hand, accept that the correct accounts of dispositionality and mentality necessarily match, but deny that intentionality is a mark of the mental. According to Place and Molnar, intentionality is rather the mark of the dispositional – and Molnar develops an analysis of dispositions in terms of physical intentionality (Molnar 2003: ch. 3). In view of this, I will not take for granted that intentionality is a mark of the mental. Nor will I start from *accounts* of intentionality, where the criteria for what makes them correct are to a certain extent left open so that they may be adjusted.

Most of the philosophers who regarded the argument from causation as valid (Martin and Pfeifer did not) took as their starting point that we seem to directly experience

causation, or more precisely, dispositional causation, as mental in character, but without mention of intentionality specifically. This will also be the basis for my argument, which will proceed as follows: first I will describe how we seem to experience causation in our phenomenology of agency – how as agents we experience something which fits Hume's central criterion for causation, but which Hume himself did not consider (sections 3.1–3.3). I will then argue that this experience is not illusory (section 3.4), that it presents causation as grounded in or instantiated by mental phenomena, including but not limited to intentionality (section 3.5), and that it is our only or most revelatory experience of causation and we have no other positive conceptions that contradict it (sections 3.6 and 3.7).

3.1 Marks of Dispositional Causation

What are the criteria for an experience of dispositional causation? Hume thinks that causation, in the robust realist sense which according to him corresponds to no impression, is essentially characterized by natural necessity. This means that causation involves a *necessary connection* between cause and effect. It is not merely so that effects *happen* to follow their causes. Rather, given the cause, and appropriate conditions, the effect *must* occur; it would be impossible for it not to. Necessitation is also *explanatory*: if the cause necessitates the effect, it explains why it occurs. And on a general level, if nature is pervaded by natural necessity, this explains why the basic laws of nature hold. By natural necessity, nature is constrained to behave regularly in the way that it does – the same causes will necessitate the same effects. Furthermore, necessary connections between causes and effects should obtain in virtue of the *natures* of things out there, not in virtue of something having to do with us – our concepts, projections or habitual judgments.

Proponents of the view that causation is grounded in irreducible dispositions or powers – which I will refer to as causal dispositionalists[3] – tend to agree with Hume about these criteria of causation; that it would involve natural, explanatory necessity. As discussed above (section 1.1), there are arguments for non-reductionism according to which irreducibly causal properties explain regularities, and provide a necessary condition

[3] Causal dispositionalism is compatible with dispositionalism, dispositional essentialism and the identity view of properties and seems entailed by all of them.

for the rationality of induction, and these arguments presuppose that causation has these features.

Causal dispositionalists also hold, against Hume, not only that there is good reason to think that causation exists, but also that we can say more about what it consists in, i.e., that we can have a positive conception of it. According to most causal dispositionalists, causes are things with powers, and in virtue of their powers they *produce* or *bring about* their effects, effects being the manifestations that the powers are directed toward.[4] The powers or dispositions of things are regarded as essential to them, i.e., as belonging to their nature, and powers are regarded as essentially directed toward their effects. This makes it the case that the laws of nature are necessary: we could not have the same things and different laws, because the behavior which can be described by law-statements follows from the very nature of these things. This is the general way in which causal dispositionalism explains laws or regularities.

Bringing about, or production, is a manner of defeasible necessitation: a cause will necessitate that which its powers are directed toward *if* no other powers interfere. In this way, powers form a ground for *ceteris paribus* necessary connections.[5]

Some causal dispositionalists, such as Mumford and Anjum (2011b) and Markus Schrenk (2010), hold that powers ground causation but not necessary connections. They argue that causation involves a *sui generis* modality which is weaker than necessity but stronger than contingency, and that the absence of necessity follows from the fact that there is always the possibility of interference from other powers. As far as I can see, though, this view still entails that causation would involve necessity of the defeasible kind, necessity *ceteris paribus.*

Some hold that dispositional causation is probabilistic – that powers only *tend* toward their effects and never guarantee them even in the absence of interference. I will argue

[4] Some think the fundamental dispositions are capacities or propensities, but the difference will not matter here – except insofar as propensities are probabilistic, and I discuss the relevance of this below.

[5] Some complain that *ceteris paribus* clauses trivialize necessary connections, reducing a claim such as "B will follow A" to the truism "B will follow A, except when it does not". This complaint may be valid if directed toward reductive analyses of causation which include *ceteris paribus* clauses, insofar as the nature of interference is not specified apart from being "that which occurs whenever B does not follow A". However, in the case of non-reductive causal dispositionalism, the qualified statement will rather read: "B will follow A, except when external powers directed toward non-B are present and are stronger that the power directed toward B". This does not reduce to a truism, because interference is understood in terms of the concrete powers whose nature we would, according to causal dispositionalism, have a positive, independent grasp of.

that we experience necessitation; however, someone who holds that causation involves only tendency may argue that the experience in question is more accurately described as an experience of a strong tendency, which would be very similar. Premise 3b might be as defensible on this basis, but I will assume that causation involves full-blown necessitation.

I will take it, then, that dispositional causation is characterized by defeasible, explanatory, necessary connections which hold in virtue of the natures of individual things, as opposed to external, governing laws, or merely our own habits, expectations or concepts. If our phenomenology of agency contains something with these features, and we have good reason not to dismiss the phenomenology as illusory, it should be taken as an experience of causation.

3.2 Efforts and Results – and Hume's Objections

Efforts seem to have many characteristics of dispositions, such as directedness toward an outcome, and being something in virtue of which such outcomes can be brought about. But do they seem to ground necessary connections? As already noted, Hume thoroughly considered the suggestion we find an impression of causation in the relation between our own acts of will and their successful results. He mainly argues that our efforts cannot be necessarily connected to their results because the results cannot be inferred and foreseen *a priori*, and illustrates this convincingly with the example of a man who is struck with palsy and does not understand that he cannot move before he actually attempts to and fails. Hume also argues that there is no *a priori* connection between efforts and results in the case of mental actions. And it does seem clear that one may suddenly be struck by, say, a stroke and, just like in the palsy case, not find out that one's mental powers have changed before one actually attempts some mental task and fails.

One might object that when our efforts fail, it is because of interference from other powers (such as the events in the brain that constitute the stroke), and that this is covered by the *ceteris paribus* clause that belongs to all causal connections according to causal dispositionalists. But even when we do succeed in our efforts, and thereby can infer that there has not been interference, we still do not experience that our success was necessary (we could at best infer it, given the prior assumption that efforts are powers that

necessarily bring about what they are directed toward in the absence of interference[6]). We can conceive of a scenario where we have the same experience of trying without encountering resistance, but the successful result did not follow our effort.

In my view, we should therefore grant Hume that we do not experience any necessary connections between our efforts of will and their results, either in the case of physical or mental actions. Still, there are other elements contained by our phenomenology of agency, which Hume did not consider, and between which I will now argue that necessary connections *can* be experienced in a way that satisfies even his strict requirements.

3.3 Necessary Connections Between Motives and Efforts

Efforts are not the only mental phenomena which seem dispositional. Motivational mental states also seem to dispose. In particular, pain disposes us toward trying to avoid it, and pleasure disposes us toward trying to pursue it.[7] Note that I only mention efforts or tryings,[8] which are mental events that need not be supposed to have any experienceable further necessary connection with subsequent mental or physical events that they are directed toward. Are there any necessary connections between motivational mental states, such as pain, and the efforts or tryings they motivate?

Clearly, pain does not always lead to avoidance attempts. Often we endure pain because we believe that it is beneficial. Sometimes we endure pain, or even pursue it, because we think we deserve it as punishment. Masochists seek pain in certain contexts because it will also give them pleasure. But this does not rule a necessary connection out, because specifying the absence of other motives or reasons to endure the pain is equal to inserting the *ceteris paribus* clause excluding interference.

Therefore, consider an instance of suddenly feeling really excruciating, torturous pain, where the subject of the pain experience has no other motives for enduring this pain, such as beliefs that it would in some way be instrumentally or morally good. We should

---

[6] In chapter 4 (pp. 135–137) I argue that the principle that efforts necessarily lead to success when there is no resistance can perhaps be inferred directly on the basis of the experience of effort. But it would not thereby enable the direct experience of the necessary connections themselves.

[7] At least two philosophers from the previous chapter took motivation as their starting point. Schopenhauer claims that "[…] from the law of motivation I must learn to understand the law of causality in its inner significance" (1859/1966a: 126). Ward claims that pain and pleasure, in particular, present themselves as sufficient reasons.

[8] Efforts can also be called acts of will or volitions. They are distinct from intentions, as ordinarily understood, because intentions can exist prior to efforts and sometimes end up not being followed by them. Searle, however, distinguishes between prior intention and intention in action (Searle 1983: 84–85), and intention in action seems to correspond to efforts or tryings.

also suppose that the subject has some belief about how to try to make the pain go away. Will the subject then necessarily try to make the pain go away? It is hard to conceive otherwise. That horrible pain, in and of itself, in the absence of any further reason to pursue it, should not make us to try to avoid it, in such conditions, is a scenario that seems impossible – in virtue of being inconceivable and unimaginable. And if it is *not* possible that pain does *not* lead to avoidance, then it is necessary that it does.

One might also consider whether pain and pleasure could be inverted in our motivational structure. Does it not seem impossible that the experience of pleasure should, in and of itself, motivate us to try to avoid it? With most regularities of nature we have no problem conceiving of them being otherwise. We can imagine, as cosmologists do routinely, that the values of the physical constants were different. We can imagine gravity working differently – that things went upwards instead of downwards when dropped, or, as in cartoons, that gravity did not take effect before someone notices that it is supposed to. With the regularities between pain and avoidance, and between pleasure and pursuit, it is different – it is not clearly possible to conceive of them coming apart under the conditions specified.

In this way, there is a distinct seeming of necessity. Unlike the connection between, e.g., efforts and a successful result, between pain and avoidance efforts we seem to experience a necessary connection. We can also experience any interference in the form of other reasons or motives. In the case of efforts and results such as bodily movements, it seems like we normally need to *infer* the presence of interferers when efforts fail and the absence of interferers when they succeed. When it comes to interfering motives, we can detect their presence or absence directly, because they would be our own mental states.

Why could this seeming of necessity not arise on the basis of inference and not direct experience? Causal necessity is inferred (at least partly) on the basis of the regularities it underlies. If the causal necessity between pain and avoidance attempts were merely inferred, we could not know it (as in subjectively justified knowledge as opposed to, e.g., reliable instinctive judgments) on the basis of a single instance. We would have to have observed many times that whenever we are in pain, we try to avoid it, unless we have another reason not to. But it seems we can know immediately, from experiencing pain just once, that this is something that we will systematically try to avoid. The number of repetitions does nothing to increase our confidence in this.

Not only do we know directly *that* pain and avoidance are connected, we also seem to have substantial insight into *why* they are connected. In order to understand this, we

only need to consider pain in terms of how it feels. There is something about how pain feels that seems sufficient to explain avoidance in a way that does not give rise to further questions. If someone genuinely wants to know *why* people tend to avoid painful experiences, the answer would be to let them know what a painful experience is intrinsically like. In this way, with the feeling of pain we have an experience of something with the power to explain regularities in the way powers or dispositional essences are supposed to. It cuts off the explanatory regress intelligibly and non-arbitrarily. With other laws or regularities of nature, we can never find such an answer to the question as to why they hold.

Why would the connection between pain and avoidance attempts be a matter of natural necessity, as opposed to merely analytic or conceptual necessity? Analytic functionalism notwithstanding, it does not seem to follow logically from any explication of the concept of pain that it would be impossible to experience it without trying to avoid it when there is no other reason not to. The thought of pain that does not make us try to avoid it when there is no reason not to is not like the thought of a square circle – a combination of concepts that generates a logical contradiction. The impossibility of pain that does not make us try to avoid it seems rather to have its source in its *nature* – given that the nature of pain is its phenomenal character, or *what it is like* to be in pain. Only when we consider pain in terms of how it feels, by either imagining it or having an experience of it, do we understand why it disposes us toward avoidance and why this could not have been otherwise.

Perhaps one could say that avoidance attempts in a sense follows with conceptual necessity, once we think of pain under a *pure phenomenal concept*. Chalmers defines pure phenomenal concepts as concepts that pick out phenomenal properties directly in terms of their intrinsic phenomenal nature (2010: 256), and my claim is that only when thinking of pain in such a way does the necessary connection appear. But necessity in virtue of phenomenal concepts is not like ordinary conceptual necessity, which can be argued to only reflect how we think about things, and not how things really are. Chalmers claims that for pure phenomenal contents, the quality of the experiences they refer to play a role in constituting the content of the concept, so that "one might say very loosely that in this case, the referent of the concept is somehow present inside the concept's sense […]" (2010: 265–266). A necessary connection in virtue of concepts seems to strongly indicate a necessary connection in virtue of the constituents of the concepts. If pure phenomenal concepts are in some sense constituted by the phenomenal nature of the property they

refer to, a necessary connection in virtue of such concepts strongly indicates a necessary connection also in virtue of phenomenal natures.[9]

3.4 Veridical Experience of Necessary Connections

So far, I have shown how the experience of our own motivated agency seems to contain distinct features of dispositional causation. But why should we think that this experience accurately represents reality? I will give two different reasons, one pragmatic and one theoretical, to think our phenomenology of agency actually reveals natural necessity and does not merely seem to, i.e., to think the experience is veridical or true and not just illusory. Firstly, it could be said that if our core phenomenology of agency is not veridical, it follows that we are not really agents at all. This entailment is presupposed, or at least regarded as highly plausible, in much philosophical literature on agency, but it is not often argued extensively for as its own thesis. Esfeld, as quoted above (p. 96), seems to take it for granted that it follows from (1) "Humean metaphysics cannot admit our experience as acting beings (agents) in the world as being veridical", that (2) "there is no room for agents in a Hume world" (2007: 212). Horgan (2007) makes the same assumption. He claims that:

> […] the right agenda to be pursuing, with respect to the generic philosophical project of defending our belief in our own agency, is the agenda of providing viable philosophical defenses of the various claims that jointly comprise [a position asserting (i) that humans are indeed agents of the sort they experience themselves to be …]. (Horgan 2007: 185)

Helen Steward (2012) argues extensively that our ordinary *concept* of agency needs to be satisfied in order for there to be agency at all. It is plausible that our ordinary concept of agency is in large part based on our phenomenology of agency. Steward holds, however, that our ordinary concept presupposes a certain kind of freedom, rather than necessity (a point I will return to in chapter 4, section 2.3, when considering objections). The point here is that she defends the entailment from an unsatisfied ordinary concept of agency to the denial of agency altogether, even though she may disagree about the essential content of the concept. Recall also Schiller, who defended that we have experience of causation

[9] Chalmers claims phenomenal concepts of qualities that we are not experiencing right now, but rather only have a memory of, are constituted by something like a faint Humean copy of the phenomenal quality (2010:272), but a connection in virtue of such a copy would still indicate a connection in virtue of the original.

on the basis that otherwise one result would be "the utter cancellation of agency" (chapter 2, p. 76).

That we are in fact agents is perhaps not theoretically self-evident. But having to deny it would clearly be a pragmatic disaster; a consequence that would be impossible to live with. As Jerry Fodor puts it:

> [...] 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. [10] (Fodor 1989)

Secondly, the hypothesis that there are in fact necessary connections between pain and avoidance efforts, and pleasure and efforts toward pursuing it, has the power to explain what would otherwise look like an incredibly fortunate coincidence of evolution. Consider this argument James gives against epiphenomenalism:

> *It is a well-known fact that pleasures are generally associated with beneficial, pains with detrimental, experiences.* All the fundamental vital processes illustrate this law. […] Mr. Spencer and others have suggested that these coincidences are due, not to any pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long-run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable. An animal that should take pleasure in a feeling of suffocation would, if that pleasure were efficacious enough to make him immerse his head in water, enjoy a longevity of four or five minutes. But if pleasures and pains have no efficacy, one does not see […] why the most noxious acts, such as burning, might not give thrills of delight, and the most necessary ones, such as breathing, cause agony. (James 1890/1981: 143–144)

This is an argument not only against supposing that pleasures and pain have no causal power at all, but also, more importantly, against supposing that pleasure and pain could cause *any* kind of behavior. In other words, it supports the view that not only is there actually a causal connection, in some sense, between pain and avoidance, and pleasure and pursuit of it; this connection must be necessary.

[10] As I will argue below, if more attention is given to phenomenology it is arguable that Fodor's description here actually rules out agency; that is, insofar as he is taken to claim that itching directly causes scratching, without an agent as an intermediary. Nevertheless, the quote serves to highlight the importance of real agency, or at least mental causation.

It seems like sound scientific reasoning to suppose that evolution shapes us to find certain experiences enjoyable and others noxious because these experiences will then be necessarily connected to pursuit and avoidance, respectively, and therefore selecting these motivations is a way of selecting the corresponding behavior. If this were not the case, if any motivational feeling could in principle give rise to any behavior, then evolution could just as well have linked pain and pleasure to different behavior, and as a result we could have creatures for which it felt terrible to breathe, as in James' example, but they still voluntarily did it, and it felt very pleasurable for them to suffocate, but they completely abstained from this after learning how pleasurable it felt.

Those who deny that there are necessary connections between pain and avoidance, and pleasure and pursuit, *ceteris paribus*, cannot explain why this is not the situation that obtains in nature. One would have to suppose that it is just an extremely fortunate coincidence that we find ourselves avoiding pain and pursuing pleasure, when these correlations could just as well be switched around if evolutionary history had been different.[11] Accepting that motivation and the behavior it motivates are necessarily connected explains this correlation and saves us from the implausible hypothesis that it is just a coincidence.[12]

An objection might go as follows: the disposition to avoid pain and seek pleasure is partly constitutive of rationality in general, and it is no mystery that there would be evolutionary pressure toward rationality. It can be granted that avoiding pain and seeking pleasure is in some sense rational, from the subjective and ethical point of view. But if so, it would follow that there would be very little evolutionary pressure toward rationality. In the scenario under consideration, pain is correlated with behavior or stimuli which is beneficial, and pleasure with behavior or stimuli which is detrimental. Therefore, avoiding pain and seeking pleasure would decrease chances of survival, and creatures who were rational enough to avoid pain would be selected against. Could avoiding pain and seeking pleasure be constitutive of rationality in any other respect, respects that

[11] Is this switching not precisely what obtains for patients with strong obsessive compulsive disorder? People with obsessive compulsive disorder often find themselves doing what they do not want to do. I think this is more of a problem of second-order motives losing to first-order motives; that is, of finding oneself with unwanted urges. It does not seem that compulsive behaviour is accurately described in terms of actions being caused solely by the desire not to perform the action.

[12] Reductive functionalists can explain it too, of course, although in a certain sense they rather explain the problem away by reducing the phenomenal quality of motivational states. So the argument must be supported by the kind of arguments against type-A physicalism discussed in chapter 1.

evolution could select for? If pain and pleasure is only contingently connected with avoidance and pursuing, respectively, then it should be equally contingently correlated with behavior and capacities which are rational in other respects, such as the capacity for logical reasoning and planning – so it is hard to see how this could be.

Another objection would be to say that the connection between events such as pain and avoidance could not really have been otherwise because it supervenes on the basic laws of nature, or alternatively, one might posit that the connection itself is a basic law of nature. Assuming the view that the basic laws of nature are metaphysically contingent, this would avoid the conclusion that the connection between pain and avoidance is metaphysically necessary, which is what the *Mental dispositionality* premise requires. But if so, the fact that such laws of nature actually obtain in our world could still be regarded as an extremely fortunate coincidence. The problem would have a lot in common with the problem of cosmic fine-tuning. Not only would we wonder: why do the physical constants happen to have values that make life possible? – but also: why do we happen to have laws that make the experience of being alive not completely absurd and hopeless? Or: why is it not the case that living creatures find themselves pursuing pain and avoiding pleasure for its own sake, or otherwise motivationally mismatched?

Now, the *ceteris paribus* necessary connection between pain and avoidance effort does not alone explain the correlation – there must also be a *ceteris paribus* necessary connection between avoidance efforts and actual avoidance. However, the connection between motives and efforts would still be a necessary condition for the connection between motives and successful action.

3.5 Experience of Mental Causation

Why does the causation that we experience in agency, according to the arguments so far, constitute mental causation? Or why are the causal and dispositional properties we experience to be regarded as mental properties?

One way in which the causation we experience in agency appears to be mental is in virtue of involving phenomenology – the experience shows that there is something that it is like to be a cause or part of a causal process. But it could then be objected that causation itself might be separable from the phenomenology. Since we have mentality, it could be held, we represent our own causal activity to ourselves through phenomenology. In inanimate nature, on the other hand, causation exists as something continuous with agency, but it is not represented. Therefore, causation involves phenomenology only in

our case, but it does not essentially involve phenomenology. This objection takes phenomenology of agency as a kind of perception of causation, insofar as perceived objects are generally regarded as existing independently of perceptual phenomenology. The view that causation is perceived in agency is expressed, though not explicitly defended, by Mumford and Anjum, who hold that "causation can be perceived within one's own body" (2011b: 209), and Timothy O'Connor, who speaks of "perception of the agent-causal relation" (O'Connor 1995: 196–7).

As seen in the previous chapter, Leibniz and Schopenhauer both emphasized that experience of causation is not a matter of perception, which is indirect and mediated, but rather a matter of direct acquaintance. The view that experience of causation is a kind of perception is problematic in and of itself. One central difficulty with it, it seems to me, is that it is incompatible with any causal theory of perception. If we assume that perception is a matter of being causally affected (in the right way) by the perceived object, as many naturalists would suppose, then perception of causation would require that causation itself causally affects us. This is metaphysically problematic in a number of ways.

However, I will not attempt to develop this objection any further, because the perception hypothesis, even if coherent, would not help avoid the conclusion that we experience causation as mental anyway. The reason is that, if the account I have offered so far is correct, the causation we experience in agency must be regarded as mental not mainly in virtue of being represented by phenomenology, but rather, mainly in virtue of what this phenomenology represents or contains.

Firstly, the phenomenology contains phenomenal qualities – not (primarily) as a representational vehicle, but as causal content. I have argued that the phenomenal nature of pain, what it is like to feel it, is what grounds the disposition to avoid it, and that in virtue of which we can understand why regularities such as "pain tends to lead to avoidance attempts" hold in a way that does not engender a regress of further why-questions. Therefore, even if it is granted that the causation that is revealed in agency could exist without being *represented* to the agent by phenomenology, this causation would still in involve phenomenal qualities, because it is grounded in them. Qualitativeness, or the what-it-is-like-ness, of experiences such as pain, is typically regarded as a paradigmatically mental phenomenon.

Secondly, implicit in the account I have defended is a reference to a subject of experience. Motives such as pains do not directly connect with efforts to avoid them. They are connected by a subject which constitutes a causal *nexus* between them. The

subject constitutes a nexus in virtue of experiencing the pain and having the power to make an effort toward avoidance. If the pain is not experienced by a subject, or the subject that experiences it does not have the power of agency, the power to try to avoid it, then there would be no experienceable necessary connection. Additionally, experienceable necessary connections obtain only when it is the same subject that experiences the pain that also exerts the effort. Thus the necessity that we experience is conditional on the existence of an agentive subject of experience, which is yet another paradigmatically mental phenomenon.

I am now making explicit that our phenomenology of agency is phenomenology of agent-causation, the kind of phenomenology I quoted Ginet and Horgan as describing in chapter 2 (p. 87). According to the agent-causal view, it is not the case that mental states such as pain cause actions directly. Rather, the agent causes the events that count as their actions (or together with their efforts constitute their actions, as some prefer to say) *in view of* motives. Some find agent-causation metaphysically problematic and I will consider some objections to it in chapter 4 (sections 2.3-2.5). For now, I will provisionally note that I take agent-causation to be fully compatible with both a deflationary (though not an eliminative) view of subjects and a lack of libertarian free will.

Now, if the phenomenal qualities of motives, subjects of experience, as well as the intentionality of efforts, are all parts of constituting that which explains regularities, or of what instantiates necessary connections, there is a clear sense in which these necessary connections are mental phenomena. If there are any mental phenomena at all, i.e., phenomena which either constitute or essentially carry a mark of the mental, it seems at least one of these elements must qualify. Intentionality might be excluded, if Molnar and Place are correct, but it is implausible to exclude both qualia and subjects of experience as well.

In summary, my argument that we know dispositional properties that are mental properties has been as follows: Dispositions are supposed to ground necessary connections in virtue of the natures of things, and we experience necessary connections between motives and efforts, such as pain and avoidance attempts, in virtue of the phenomenal nature of motives and the capabilities of subjects of experience. These are both paradigmatically mental phenomena, and when causal necessity obtains in virtue of them, this causal necessity should be regarded as mental. This experience should be regarded as veridical, because if our phenomenology of agency is illusory, then there is

no agency, and agency is something that we know, or at least pragmatically speaking need to believe is real. Furthermore, if there are no necessary connections between motives and efforts, such as pain and avoidance, the fact that evolution has still connected them would seem like an extraordinarily fortunate coincidence.

3.6 Other Experiences or Conceptions of Causation

Premise 3b, *Mental dispositionality*, does not just say that we know dispositional properties that are mental properties, it also says that the *only* dispositional properties we know, or have a positive and coherent conception of, are mental properties. Establishing this might start from establishing either one of these theses:

(1) There are no other experiences, and no other positive and coherent conceptions, of causation.

(2) There is only one kind of causation.

In order to establish thesis (1) one would have to seek out every other account of how we know, or may positively conceive of, causation and argue against these accounts. Firstly, there are other direct realists views, according to which causation *can* be experienced, but other places than in agency, such as the interactions between external things. This view is associated with G. E. M Anscombe (1971).[13] Secondly, there are various forms of conceptual primitivism, according to which the concept of causation is not grounded in experience, but is necessarily prior to experience. An example of conceptual primitivism is Kant's transcendental idealism.

Establishing thesis (2) – there is only one kind of causation – would be easier. It would be defensible largely on the same basis as one could argue for monist views in general, such as physicalism and Russellian monism, mainly the argument from the causal closure of the physical (which I discuss in more detail in chapter 5, section 3). If there are two or more kinds of causation, it would amount to interactionism or overdetermination. Establishing the final part of *Mental dispositionality* on the basis of thesis (2) might seem just as effective as going via thesis (1). If there is only one kind of causation, and we experience it – veridically – in agency as mental, then it seems any other experiences which do not represent causation as mental, and any other non-mental

[13] As I read Anscombe, she actually defends a version of conceptual primitivism, but direct realism is often attributed to her.

concepts of causation, must be either non-veridical experiences or unsatisfied concepts, or incomplete experiences or concepts that leave out something essential to causation. Hence, no positive conceptions actually refer to non-mental causation.

However, this simple defense involves a big as of yet unjustified presupposition, namely that the mental causation we experience is *irreducibly* mental. Only if we experience irreducibly mental causation can we really say that mental causation as experienced in agency is distinct from physical causation, so that they constitute separate kinds, in violation of thesis (2), or at least any reading of it that is supported by arguments from physical causal closure. Physicalists think that mental causation is reducible to a kind of physical causation, and therefore mental and non-mental causation existing side by side would not violate the causal closure of the physical. According to physicalism, mental and physical causation would not be different kinds in a sense ruled out by a plausible reading of thesis (2). In other words, for the defense to work, thesis (2) must be read as stating that there is only one *irreducible* kind of causation, and mental causation must be regarded as irreducible.

Have I not already argued that we experience irreducibly mental causation? Above, I did presuppose that subjects, qualia or intentionality (one or more of these) are marks of the mental, and I argued that the causation we experience carries all of these marks. But something can be, or carry, a mark of the mental and still be reducible to the physical. Physicalism only has a problem if fundamental things carry marks of the mental, and I have not argued that the mental causation we experience is fundamental, i.e., irreducible.

Can it be argued, then, that the kind of causation we experience in our phenomenology of agency is irreducibly mental? Agent-causation, the kind of causation our phenomenology represents, is associated with dualism, and many take agent-causation to contain significantly more obstacles for reduction than mental causation construed in some other way. However, with a mere appeal to the apparent irreducibility of mental elements involved in agent-causation, the problem that a veridical phenomenology of agency constitutes for physicalism will be more or less reduced to the problem of consciousness in philosophy of mind. Consequently, the argument from causation would, in having to presuppose that this problem for physicalism in philosophy of mind cannot be solved, end up not constituting much of a distinct argumentative strategy for panpsychism after all. Therefore, I will not presuppose the irreducibility of subjects, qualia, intentionality, or other elements of agent-causation, or defend it with appeal to considerations from philosophy of mind. I will rather try a different argument to

show that agent-causation is problematically different from physical causation as it can be positively conceived.

Even if it is granted, in order not to beg the question against physicalism, that mental causation of the kind we experience *could* be a kind of physical causation, it would still entail panpsychism if all physical causation is mental, i.e., if no physical causation is non-mental. Physicalists, in order to avoid this conclusion, would have to give a positive account of non-mental physical causation which is continuous with mental agent-causation. Non-mental physical causation should be as similar as possible to agent-causation, apart from not involving its mental elements; otherwise, one cannot expect agent-causation to reduce to it. The best prospect for coming up with an account of non-mental physical causation which is continuous with mental (and, by hypothesis, physical) causation, is to take the account that fits our experience of mental causation, and then remove, or abstract away, the mental elements from that account. But does abstracting away the mental elements of agent-causation yield a positive conception of non-mental causation to which agent-causation may reduce?

3.7 Abstracting Away the Mental

I have argued that causation as experienced is mental not in virtue of being experienced or being represented by phenomenology. I experience many physical things, and that does not prove that they are mental. Rather, what is experienced is itself mental. It seems like mental elements like pain qualia and subjectivity are part of *constituting*, *realizing* or *instantiating* necessary connections, rather than merely *representing* them. The proposal now under consideration, on behalf of physicalism, is whether the causal properties revealed in phenomenology can be abstracted away from their particular mental realizers.

In general, it seems that abstraction is an operation that yields abstract structure. The argument from non-structural properties is looking for non-structural properties in order to avoid Pythagoreanism (and spatiotemporal structuralism, which suffers similar problems), and by abstracting away the mental realizer we are perhaps left with nothing to realize physical structure after all. But it could be claimed that we can abstract away something which is not structural in the logico-mathematical (or purely spatiotemporal) sense. Perhaps we can abstract away the pure modal or metaphysical structure of agent-causation? This would presumably be the metaphysical structure that dispositional theories of causation already describe, given how it matches agent-causation.

Now, the experience of agency, if what I have said about it so far is correct, reveals that dispositional causal structure does not exist uninstantiated or on its own. It exists in virtue of some concrete mental realizers – qualia, subjectivity and effort. Presumably, non-mental physical causation would consist in the same metaphysical structure with non-mental physical realizers. The question for physicalists, then, is: what would the non-mental realizers of dispositional causal structure be?

Via the process of abstraction from causation as experienced in agency one could only arrive at an answer like: "the kind of properties which are non-mental and ground necessary connections in virtue of their natures". This does not count as a positive conception and is therefore incompatible with premise 4 – *Anti-mysterianism* – of the argument from non-structural properties. Therefore, one cannot get to a positive conception of non-mental causation via abstraction from the experience of agency.[14]

It is still possible that one might acquire a positive conception of the relevant type of non-mental properties in a different way than via abstracting from the experience of agency. A complete defense of premise 3b would therefore have to establish something like thesis (1), listed above, after all, the thesis that there are no other experiences, and no other positive and coherent conceptions, of causation. But at this point the number of alternatives that need to be refuted has been narrowed down. Only the theories of causation that offer a positive account of a kind of properties which are non-mental and ground necessary connections in virtue of their natures, resulting in a structure analogous to agent-causation so that agent-causation may reduce to it, must be taken into consideration, and no such theory obviously suggests itself. Varieties of causal dispositionalism would have been the obvious candidates, but as I have now argued, the fact that we experience the metaphysical structure this theory describes as instantiated by mental phenomena in agency reveals that this kind of theory in and of itself is too abstract to be a basis for such a positive account.

Therefore, if my arguments so far are correct, I take myself to have shown that premise 3b is quite plausible, even though establishing the now restricted thesis (1) would make it even more so. The premise should at least seem much more plausible in view of this than it does to most philosophers *prima facie*. Hopefully, though, I would have shown

[14] This reply takes inspiration from Schopenhauer, as quoted in chapter 2: "If, on the other hand, we subsume the concept of will under that of force, as has been done hitherto, we renounce the only immediate knowledge of the inner nature of the world that we have, since we let it disappear in a concept abstracted from the phenomenon […]" (Schopenhauer 1859/1966a: 111–112).

it to be plausible also to the extent that the onus would be on the physicalist (or other non-panpsychist) to show that the now restricted thesis (1) is false, i.e., that there *are* positive conceptions of non-mental causation of the relevant kind.

4 THE ARGUMENT FROM CAUSATION: SUMMARY

In chapter 1, I showed that the argument for panpsychism from categorical properties is threatened by dispositionalism, dispositional essentialism and certain conceptions of the identity view, according to which irreducibly dispositional properties can realize all or most physical structure. Dispositionalism and dispositional essentialism entail the falsity of the premise that physical structure must be instantiated by categorical properties, and the identity view could entail the falsity of the premise that all the categorical properties we know or have a positive conception of are mental. I have argued that instead of attempting to refute dispositionalism, dispositional essentialism and the relevant versions of the identity view, panpsychists can argue that they entail panpsychism. In chapter 2, I showed that an argument of this kind has been put forth by a number of philosophers since the early modern era, both in support of the conclusion, panpsychism, and as *reductio* of the premises.

In this chapter, I have shown that there is an argument of this kind, an argument from dispositional properties, which can be put in same form as the argument from categorical properties and would thereby be valid if the latter is. Both arguments can be combined into a stronger and more general argument for panpsychism, an argument from non-structural properties, where all premises – expect for one – are either easily seen to be defensible, or defensible on the basis of arguments discussed in chapter 1. The remaining premise is that the only dispositional properties we know, or have a positive conception of, are mental properties.

I have defended this premise by showing, first, how in our phenomenology of agency we seem to experience necessary connections, of the kind dispositional essences are supposed to ground, between motives and efforts, such as pain and avoidance attempts. This experience should be accepted as veridical, for theoretical reasons, such as James' evolutionary argument, and for pragmatic reasons, insofar as a non-veridical phenomenology of agency entails that agency itself is illusory. The dispositional properties known in this way would be mental, because the necessary connections experienced in phenomenology are grounded in or instantiated by phenomenal qualities, agentive subjects of experience and the intentionality of efforts – and at least one of these

must qualify as a mark of the mental. This knowledge of mental dispositional properties constitutes our only positive conception of dispositional properties, because, firstly, there is good reason to hold that there is only one fundamental kind of causation, and if our phenomenology of agency is veridical, then the mental dispositional causation we experience through it must be of this kind. Secondly, we seem to have no basis for a positive conception of non-mental causation of such a kind that the mental dispositional causation we experience in agency could reduce to a species of it.

4

Objections to the Argument from Causation

In this chapter, I will consider various objections to the argument from causation – first some that are based on psychology and pathological cases and then some based on the philosophical presuppositions or implications of the argument. Answering the final two philosophical objections will require considering the combination problem, and my full response can therefore only be provided in chapter 7, after having discussed the combination problem in chapters 5 and 6.

1 OBJECTIONS FROM PSYCHOLOGY AND PATHOLOGICAL CASES

1.1 Pain Asymbolia

Theories of pain often distinguish two aspects of pain. The sensory-discriminatory aspect is that in virtue of which we can discriminate pain from other sensations and phenomenal qualities in general – other bodily sensations, such as tickles, other sensations, such as visual qualities, and other uncomfortable experiences, such as negative emotions. The affective-motivational aspect is the negativity of pain, that aspect in virtue of which we are compelled to avoid it. We tend to believe that these two aspects are necessarily connected – that pain has negative affective value in virtue of its sensory quality, or in other words, that feeling like pain and feeling bad amounts to one and the same thing. But some medical cases seem to show that the sensory-discriminatory and the affective-motivational aspects of pain can come apart. Nikola Grahek (2007) has argued that there is one (and only one) case where we see the complete disassociation between the sensory-discriminatory and the affective-motivational aspects of pain, or in Grahek's terms, between pain and painfulness:

> Pain without painfulness is found in patients who suffer from a rare neurological syndrome known as pain asymbolia. Characteristically, these patients feel pain upon harmful stimulation, but their pain no longer represents danger or threat to them. These patients do not mind pain at all; indeed, they may even smile or laugh at it. (Grahek 2007: 3).

Does this show that the necessary connection between pain and avoidance efforts is illusory? It clearly threatens to show this – here we have the sensory quality of pain without any avoidance effort, where this cannot be explained by any interfering motives. However, if we go into the detail of this phenomenon, it can be seen that Grahek's account of pain arrived at on this basis does not in fact contradict my account.

The first lesson to learn from pain asymbolia, according to Grahek, is that the unity of the pain phenomenon, which is often taken for granted, is an illusion:

> […] although pain appears to be a simple, homogenous experience, it is actually a complex experience comprising sensory-discriminative, emotional-cognitive and behavioral components. These components are normally linked together, but they can become disconnected and therefore, much to our astonishment, they can exist separately. (Grahek 2007: 2)

If the emotional-cognitive component, corresponding to the affective-motivational painfulness or unpleasantness of pain, were to be revealed as purely functional, while the inert sensory aspect retains all the phenomenology, and function and phenomenology come completely apart, then it looks like the dispositionality of pain phenomenology must be an illusion. But if the emotional-cognitive component also involves phenomenology, the pure affective-motivational phenomenology of painfulness, negative tone, or unpleasantness, then there is no problem. The correct thing to say then would be that it is the painfulness part of normal pain experience which is necessarily connected to avoidance attempts. It is just because it is always, in normal cases, mixed with the sensory-discriminatory aspect that we fail to notice the distinction between them, and do not see that only one disposes toward avoidance.

Grahek endorses the latter view that enables the problem to be thus dissolved. "On the one hand," he claims, "there is pure pain sensation, and on the other hand, there is the pure *feeling* of unpleasantness, defying any further sensory specification". (Grahek 2007: 111, my emphasis). What kind of feeling is pure unpleasantness or painfulness? Given that there can be (sensory) pain without painfulness, it must also be possible to experience painfulness without pain – and Grahek shows that such a case has actually been documented. The case involves a patient who cannot feel sensory pain, and is subjected to cutaneous laser stimulation, which would normally be very painful:

> [...] the patient spontaneously described a "clearly unpleasant" intensity dependent feeling emerging from an ill-localized and extended area "somewhere between fingertips and

> shoulder," that he wanted to avoid. The fully cooperative and eloquent patient was completely unable to further describe the quality, localization, and intensity of the perceived stimulus. (Ploner, Freund, and Schnitzler quoted in Grahek 2007: 108–109)
>
> His dislike of the feeling of cutaneous laser stimulation and his accompanying wish to avoid it were not motivated by the quality and intensity of the sensation felt, by that distinctive and insistent quality which distinguishes pain from other sensations. Rather, the patient was motivated by the pure unpleasantness of the experience. Paradoxically enough, in this bizarre case of pain affect without pain sensation, his only answer to the query – "Why do you dislike that sensation?" – was "Because it is unpleasant." It is true that to give such an answer is really to give no answer, but merely to reaffirm one's dislike of the sensation. It tells us nothing new. But the point is that in this case, this was all the information available to the subject. (Grahek 2007: 111)

According to my account defended in the previous chapter, some substance can be read into the patient's answer after all – what he could mean to say is that he is disposed to try to avoid the experience (by expressing the wish to avoid the laser) because of its intrinsic unpleasant phenomenal character, and that if one knew what it was like, one would understand.

If pain is actually separable into two aspects in this way, is the sensory-discriminatory aspect revealed as inert and completely non-dispositional? Grahek assigns this aspect an important function:

> […] the sensation of pain (pain quality) plays an important role in the total human experience of pain. […] its central role is to distinguish pain sensations from nonphysical pain, and from other unpleasant sensations, as well as from other sensations that may be qualitatively similar to it but are of different modality. (Grahek 2007: 103)
>
> So, as far as the two basic components of human pain experience are concerned, it is obvious that both of them are necessary, but neither of them is a sufficient condition for pain. The two phenomena give us real pain only when they work together. This is how it should be, if Mother Nature devised the pain system to serve its primary biological function: to give an organism information concerning threatening or damaging stimuli and simultaneously to move it to self-preservation. (Grahek 2007: 111–112)

On this basis, it could be held that sensory-discriminatory pain quality is also dispositional, but not by disposing toward avoidance. It might rather, for example, dispose toward activities such as thinking about the source of the pain, or thinking about

the difference between this quality and other qualities, and – in combination with the affective-motivational aspect – toward a certain manner of avoidance.

In conclusion, then, pain asymbolia is compatible with my claim that ordinary pain disposes toward avoidance in virtue its intrinsic phenomenal character. It only turns out that normal pain phenomenology has two components, both of which are experiential, but only one of which disposes in this way. The other component of pain that does not dispose toward avoidance might still dispose toward other kinds of activities than avoidance in virtue of its intrinsic phenomenal character, so the phenomenon of pain asymbolia does not show that some phenomenal qualities are non-dispositional and inert. Below (section 2.1) I will discuss whether this would have been a problem.

1.2 Illusions of Agency I – Libet

Another empirical problem for my account would be cases of apparent illusions of agency, where subjects report that they acted, but there are empirical reasons to conclude that they did not act at all. This is in tension with the claim that phenomenology of agency provides immediate acquaintance with causation, as opposed to indirect awareness analogous to perception, because with acquaintance comes at least some degree of infallibility.

Benjamin Libet (1985) conducted some famous experiments that constitute such a challenge. These experiments purport to show – though the interpretation of the data is controversial – that certain brain events, readiness-potentials, systematically occur seconds before our conscious decisions. If these brain events are interpreted as unconscious decisions, then it seems natural to take the subsequent conscious decision as not another real decision predetermined to occur by the unconscious decision, but rather as a passive coming to be aware of the decision that has already unconsciously been made. Our phenomenology of making a decision, which can be regarded as a mental action, would according to this interpretation not really be an awareness of mental causation but at best an epiphenomenal (with respect to the outcome of the decision) representation of prior physical causation.

One response to this problem is to claim that there is no reason to interpret the preceding brain events, the readiness-potentials, as unconscious decisions – they could rather be interpreted as conscious urges to make a decision (Nahmias 2002: 532), or simply as unconscious antecedents of the decision with no appropriate description in psychological terms. If so, the experiment only confirms that our decisions have causal

antecedents, mental or physical, and this would not be in conflict with any claims about our phenomenology of agency that the argument from causation relies on.

Another response would be to grant that readiness-potentials are unconscious decisions, and that our conscious decisions are determined by these unconscious decisions, but then claim that the conscious decision could still be a real decision. What happens, it could be held, is basically that the unconscious mind[1] decides that the conscious mind should decide. If this sounds odd, I think it is mostly because it is odd that there could, strictly and literally, be such things as unconscious decisions in the first place. The oddness should therefore, if anything, be taken as a reason to reject the interpretation of readiness-potentials as unconsciousness decisions, in favor of interpretations of them as urges or non-psychological causal antecedents, rather than as a reason for doubting that the conscious decision is real.

It could be held, though, that there is nothing wrong with unconscious decisions, but that the idea that our conscious decisions are determined by unconscious decisions would still be objectionable, because the notion of *deciding to decide* is incoherent. In response, it could be granted that deciding to decide is not something we often, or perhaps even are able to, do consciously. However, when the first decision is unconscious, deciding to decide could make more sense – the unconscious could decide to produce a conscious reason, motive or urge that determines conscious decision, for example. This interpretation is also consistent with the phenomenology and would be available as long as it is not shown that the conscious decision (or brain event involved in it) is not even on the causal path toward the action – and Libet's experiments do not show this.

1.3 Illusions of Agency II – Wegner

Another challenge comes from Daniel Wegner, who has argued for a theory of *apparent mental causation*:

> According to this theory, when a thought appears in consciousness just prior to an action, is consistent with the action, and appears exclusive of salient alternative causes of the action, we experience conscious will and ascribe authorship to ourselves for the action. Experiences of conscious will thus arise from processes whereby the mind interprets itself – not from processes whereby mind creates action. Conscious will, in this view, is an

[1] Which according to panpsychism would in a sense be conscious, but it would probably be a system constituted by an aggregate of microsubjects and not a macrosubject of its own.

> indication that we think we have caused an action, not a revelation of the causal sequence by which the action was produced. (Wegner 2004: 649)

Wegner claims that it is common to see illusions of agency in cases of flipping coins and rolling dice: when someone is intensely hoping and intending for a certain outcome, and the outcome then occurs by coincidence, they may feel like they *made* it happen. Given that I do not claim that we are acquainted with causal relations between intentions or efforts and outcomes, there being a possibility of an illusion here is no problem. A problem would arise only if someone distinctly experiences the phenomenology of doing something, but from the outside it looks like they are doing nothing. Being deluded about *what* one is doing, remains fully possible on my account. In the dice and coin-flipping cases, the subjects were doing something, namely hoping and intending – and thereby at least firing some neurons. The subjects are just very much mistaken about the possible effects, or the appropriate description, of these mental actions, i.e., about *what* they were doing.[2] Since actions are always described in terms of their (typical) effects or outcomes, the fact that we do not have *a priori* knowledge of the effects or outcomes of actions entails that we cannot have *a priori* knowledge of the appropriate descriptions of our actions either.

But Wegner presents another case where subjects report that they acted, while it really seems they *were* doing nothing, not just something else. This is the so-called *I Spy* study, here summarized by Bayne and Levy:

> […] participants and an experimental confederate had joint control of a computer mouse that could be moved over any one of a number of images on a board (e.g., a swan). On certain trials the confederate forced the pointer to land on a target image while participants were primed with the name of the image via headphones at a certain temporal interval either after or before the pointer landed on the image. When the prime occurred immediately before the pointer landed on the target image participants showed an increased tendency to selfattribute the action, that is, to claim that they had intended to land on the image. Wegner and Wheatley argue that the prime creates an experience of agency in the absence of an exercise of agency. (Bayne and Levy 2006: 54)

This case is difficult to interpret. Firstly, one must be clear about what the participants are reporting. The study asked the participants to rank how much they had intended to make

[2] I assume that intensely hoping or intending are activities, not something one passively undergoes, so the subjects of the illusion are doing *something*.

the pointer stop on the image on a scale from 1–100 (Wegner 2003: 75), and the highest rating reported, when the prime occurred between 1 and 5 seconds before, was around 63 (Wegner 2003: 77). It is not clear whether reporting 63 % intentionality equals reporting actually having acted or tried. However, let us grant that it does; that the participants were self-attributing action. In what sense, then, did they feel that the action was theirs? Are they reporting to have experienced the distinct phenomenology of trying to move the pointer to the image? Or are they just reporting, for example, a feeling of being responsible for the movement of the pointer, the same kind of feeling one may attach to, say, the actions of one's children, or the actions of a group one is part of, even though strictly speaking they are not one's own actions (something one might easily even forget)? The participants might not have been clear about the many ways in which we can feel that an action is ours, ways which may only be distinguishable on reflection, and therefore not really be reporting what they seem to report.

But even if the participants are reporting having had the phenomenology of trying to move the pointer, this would not be a problem as long as this phenomenology was not present at the moment that the pointer was being moved by the confederate, but rather only present in a memory that they confabulated later. My account does not rule out *post hoc* confabulation or errors of memory, only that we can be mistaken about whether we are acting, i.e., exerting effort or power, as it is happening. Wegner himself points to the ubiquity of confabulation of agency after the fact:

> Studies of the confabulation of intention following action show that people often invent or distort thoughts of action in order to conform to their conception of ideal agency. People who are led to do odd actions through post-hypnotic suggestion, for example, often confabulate reasons for their action. Such invention of intentions is the basis for a variety of empirical demonstrations associated with theories of cognitive dissonance (Festinger 1957) and the left-brain interpretation of action (Gazzaniga 1983). (Wegner 2004: 659)

In summary, then, I would have to assume that the findings of the *I Spy* study can be explained as results of *post hoc* confabulation or misremembering, or as results of misinterpreting the reports of the participants as reports of actual phenomenology of trying instead of a more general feeling of being responsible, or of having in some sense intended without actually having acted.

Wegner presents a number of cases in support of his theory of apparent mental causation, and I cannot discuss them all here. In general though, the main thesis he is

interested in refuting seems to be the thesis that we have *a priori* knowledge of causal relations between will and its physical results, and it bears repeating that I fully agree that there is no reason to assume this. He also wants to refute the idea that we experience will in an especially metaphysically demanding sense, a view of will which:

> At the extreme […] makes the scientific study of [will] entirely out of the question, and suggests instead that it ought to be worshiped. Pointing to will as a force in a person that causes the person's action is the same kind of explanation as saying that God has caused an event. This is a stopper that trumps any other explanation, but that still seems not to explain anything at all in a predictive sense. Just as we can't tell what God is going to do, we can't predict what the will is likely to do. (Wegner 2003: 12–13)

While I am defending a view according to which the nature of will is to a certain extent outside the scope of scientific explanation (given that it belongs to the Russellian intrinsic natures of things), I am not claiming that will is a stopper for scientific explanation of any empirical phenomena. Since I am not defending any of the main theses Wegner is out to refute, many of the cases he presents are not relevant. Those that are relevant seem like they must all be approached along the lines of the responses just suggested.

1.4 The Transparency of Motivation

The two previous sections hinted at ways in which our own motivations are not always transparent to us. We cannot tell whether pain motivates as a whole, or just in virtue of one of its aspects. We may also confabulate and misinterpret our own motives and intentions. Does the *Mental dispositionality* premise presuppose that our own motives are always transparent to us? If so, it would be inconsistent with a host of evidence from psychology.

Schopenhauer famously claims that the will is groundless or arational – it want what it wants, and that is all there is to it. How come we can then explain it in terms of motives? When we explain our actions by their motives, we are, according to Schopenhauer, imposing the forms of representation on the will. We are no longer looking at the will directly, but rather comparing and relating one representation of the will to other representations of it. Therefore, we can make mistakes about our motives, while being infallibly in contact with the will.

I think of Schopenhauer's point as being that all we can infallibly *know*, in the case of every particular action, is that I tried to do this action now because I felt like it, because I felt the urge or compulsion to, or because I just wanted to. This is not an empty truism,

because the particular feeling, urge or want is concretely experienced and, by Schopenhauer's hypothesis (and mine), it actually reveals the nature and ground of causation. However, nothing scientifically interesting follows from it. When we *conceptualize* our feelings, urges and wants as being of a certain type (as opposed to just "*this* feeling") we can get something scientifically interesting, but we can also make mistakes. Conceptualizing a feeling involves comparing it to other particular feelings, which are not immediately present for comparison, and thereby the possibility of error enters.

Take the example of smoking. It is not always obvious to smokers why they are smoking – whether it is for the taste, the feeling, the stress relief, the addiction or something else. Consider trying to find out why you are really smoking a cigarette. Is it for the sake of enjoyment, in and of itself, of for example the taste of the tobacco or other feelings or sensations that come with it? Or is it in order to distract from negative emotions, such as stress, boredom or perhaps withdrawal symptoms? In other words, what is your motive? Finding the answer involves paying attention to the feeling of wanting to smoke, to other simultaneous feelings, reflecting on them and comparing them to representations of other feelings and urges you have experienced or can imagine which fall under some of the concepts in question, such as addition, relief, enjoying something for its own sake, and so on. This is clearly a difficult task, in which there is a lot of room for error. It does not, however, contradict the view that you are (or were, as it occurred) immediately acquainted with the feeling you are trying to find the proper concept for, and that you have infallible knowledge that *this* feeling (regardless of how it is properly described) ultimately explains *this* particular action.

This model explains how we can be acquainted with causation without being able to infallibly conceptualize our motives, and thereby makes plenty of room for confabulation and error of the kind psychologists have amply documented.

2 PHILOSOPHICAL PRESUPPOSITIONS AND IMPLICATIONS

Now I will consider some philosophical presuppositions and implications of the argument (apart from panpsychism!) that some might find objectionable.

2.1 Are All Experiences Motivational?

Does the argument presuppose that necessary connections are present in all motivation or just with pain and pleasure? If pain and pleasure were the only experiences that ground necessary connections, a hedonistic theory of motivation would follow – a theory many

find reason to resist. It would also be problematic if many phenomenal qualities were inert and non-dispositional on their own, and only became efficacious by being contingently connected with dispositional qualities, pain and pleasure. How could physical structure be systematically and intelligibly grounded in mental properties, and how could all mental properties have a causal role, if only some are intrinsically dispositional and some are purely categorical? If both categorical and dispositional aspects play a role in grounding the physical, it is hard to see how it could be that not all of the mental grounding properties have both aspects. If only one aspect plays a role in grounding the physical, it is hard to see what could be the causal role of the mental qualities that are without this aspect.

These problems, hedonism and a worrisome complication of the Russellian grounding relation, would not follow if all phenomenology, not only pain and pleasure, is intrinsically motivating and thereby dispositional. Hartshorne (1934) argues that all possible experiential qualities form an *affective continuum*. On his view, all qualities have an intrinsic affective aspect: "in feeling, 'striving' is implicitly incarnate […] (Hartshorne 1934: 37). For some qualities, like colors and sounds, this aspect is just not so noticeable. Hartshorne claims that "the 'gaiety' of yellow […] is the yellowness of the yellow" (1934: 7) – i.e., emotional tone is intrinsic to yellowness. Other possible examples are how blue seems calming and red seems energizing. These emotional aspects might be so weak that they are always drowned out by stronger motivations in our actual experience. But we can imagine that if they were properly isolated we could clearly experience their emotional tone and dispositionality.

I have already suggested that the sensory-discriminatory aspect of pain, the component that does not dispose toward avoidance, is not inert and non-dispositional on its own, but rather disposes one more weakly toward, e.g., thinking about the source of the pain quality, or thinking about the difference between this quality and other qualities. The dispositionality of these qualities would be weak in the same way the dispositionality of colors is weak. The role of such weak dispositions would perhaps usually be to combine with stronger disposition, such as an avoidance disposition, as modifiers or them. Sensory pain together with a pure unpleasantness would result in a different manner of avoidance than, say, the sensory quality of anxiety together with a pure unpleasantness (assuming that the unpleasantness in anxiety and bodily pain is similar).

If a theory like Hartshorne's, according to which all experiences are at least weakly dispositional, is plausible, then the argument would not entail hedonism and problems

with Russellian grounding. All qualities could be intrinsically motivating and dispositional without being coupled with pain and pleasure. The dispositional aspect of qualities could be regarded as essential to Russellian grounding, and any categorical aspects as necessarily connected to the dispositional aspect and thereby essential as well (I will discuss whether or not it follows that mental properties are also categorical in the next section). I find Hartshorne's theory very plausible, and will indirectly defend it by pointing to some further advantages of the continuum of qualities in chapter 5 (section 4.5). If such a theory cannot be made to work, however, perhaps a sophisticated version of hedonism can be found which escapes the most serious objections. Perhaps also a way of integrating both dispositional and purely categorical mental properties as grounds, systematically and intelligibly, can be found, even though how it would work is not immediately obvious.

2.2 Implications for Theories of Properties

I have argued that mental properties are dispositional. In chapter 1, I discussed some arguments that mental properties are categorical. The arguments can all be sound if the identity view of properties is true: if the categorical and the dispositional are two aspects of every single property. Considering how mental properties are arguably both categorical and dispositional could help to render the identity view more intelligible.

In principle, if the identity view is rejected, then some of the arguments must be mistaken. As I have shown, panpsychism would still follow if only one type of argument is correct. According to the argument from non-structural properties, only one of these premises needs to be (non-vacuously) true:

> 3a) *Mental categoricity:* The only categorical properties (whose nature) we can know, or positively conceive of, are mental properties.
>
> 3b) *Mental dispositionality:* The only dispositional properties (whose nature) we can know, or positively conceive of, are mental properties.

Because the conclusion follows anyway, the argument from causation, or dispositional properties, can in principle leave open the question of whether mental properties are purely dispositional or categorical as well (but their being purely categorical is ruled out). Given that there is room for error when applying concepts to our experiences (see section 1.4 above), this is not in tension with the presupposition that we are acquainted with the

nature of our experience (in some respects), as per The Partial Revelation Thesis and phenomenal essentialism (see chapter 1, p. 16 and 46).

2.3 Agent-Causation and Freedom

I have argued that our phenomenology of agency reveals necessary connections between motives and efforts in virtue of a subject of experience that experiences the motives and exerts the effort. This is the phenomenology of agent-causation.

Although, as noted, agent-causation is often associated with incompatibilist libertarianism about free will, agent-causation of the kind I have claimed we experience is perfectly compatible with determinism – in particular, determinism by reasons.[3] It would be bad news if phenomenology of agency seemed to reveal only libertarian agent-causation. I argued that we experience the kind of necessitation that could underlie the regularities of nature. Libertarian agent-causation involves no necessitation, because libertarian agents are uncaused causes, or partially uncaused causes, and this kind of spontaneous causation is more suited to underlie irregular behavior.

But with this rejection of libertarian freedom an objection may come up. The notion of libertarian freedom, regardless of its philosophical plausibility, is often held to have its source in our phenomenology of agency. A hard determinist might think that even though there is no such thing as libertarian freedom, we do still (non-veridically) experience ourselves as free in this way, and that this is perhaps the reason for the appeal of libertarianism in spite of its arguable philosophical problems. How can agency be where we experience necessary connections, and at the same time be the source of the notion of libertarian freedom, the opposite of necessity?

Whether we actually directly experience ourselves as free in the libertarian sense is debatable. It might be an inferred hypothesis or interpretation. Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer and Jason Turner have conducted a survey on the limited

[3] There can be no threat to free agency from physical determinism when it turns out that the laws of nature are grounded in mental properties, i.e., the intrinsic nature of phenomenal reasons. Determinism by reasons (i.e., motives of all kinds) has the advantage of being compatible with more views of freedom than physical determinism is, for example, the Kant-inspired view that freedom consists in being determined by rational motives, like moral imperatives, as opposed to arational or irrational motives, like desires and inclinations. Many find that rational determination is threatened by physical determination being fundamental, but on my view, physical determination is not fundamental so there is no such threat. A rational determination view of freedom is not only compatible with the view that there are necessary connections between motives and efforts but also requires it: an agent is free in this sense when their efforts are necessitated by rational motives. Rational motives could be experiential and motivate in virtue of their phenomenal natures assuming that there is such a thing as cognitive phenomenology.

data that exists documenting whether people in general take themselves to directly experience free will, with the following result:

> The data seem to support compatibilist descriptions of the phenomenology more than libertarian descriptions. We conclude that the burden is on libertarians to find empirical support for their more demanding metaphysical theories with their more controversial phenomenological claims. (Nahmias et al. 2004: 162)

This indicates (though very inconclusively) that phenomenology of agency alone is at least not the whole source of the libertarian notion of freedom.

However, if the indications would turn out to be misleading, if it is shown that we do actually experience ourselves as free in the libertarian sense, this would not necessarily be a serious problem for my argument as long as we do not experience ourselves as free all of the time. In particular, this phenomenology would have to be absent from some cases where pain motivates avoidance. Perhaps I could modify the *ceteris paribus* clause, and reformulate my claim as: there are necessary connections between motives and efforts in the absence of (1) interfering motives *and* (2) any capacity for spontaneous free action.[4]

2.4 Agent-Causation and the Persisting Subject

Agent-causation is also associated with the view that the subject, or self, is a *persisting* substance – which, again, many find reason to resist. But, just like with freedom, agent-causation does not entail that subjects persist over time. There is no reason to think only persisting things can be causes. However, some think that subjects are by nature necessarily persisting – unless they are annihilated by external forces they will naturally go on existing – and that by positing agent-causation we are forced to accept them. But this is not a given. Strawson defends a deflationary view of subjects (Strawson 2009, 2008b), according to which subjects do not persist through time for very long at all, but only last as long as a moment of experience – the reason being that subjects are not really distinct from the totality of their experiences. If Strawson's view, or another theory that

[4] The absence of the free capacity, as opposed to just the absence of the exercise of a free capacity, is required because it seems that if we (feel that we) have a capacity for free choice that we do not exercise, we (feel that we) freely choose not to exercise it. Many libertarians argue that the capacity for free choice is not always present, i.e., that free choice is possible only in certain cases and not always: in cases where desires and motives are closely balanced, but not in cases where desires or motives strongly point in one direction. Therefore, the clause is relevant to some, but not all, actually held views about how libertarian freedom works.

explains how subjects can be transient, is plausible, then we are not forced into accepting persisting subjects by accepting agent-causation.

Rejecting persisting subjects also lets the argument from causation avoid another potential problem. Empiricists, e.g., Hume, often claim that we have no impression of a self as something that persists in the background while experiences change. If we have an impression of causation, and causation is (partly) instantiated by a subject, then the subject cannot be equal to the unobservable persisting self because then there would be no impression of causation after all. If we accept a view whereby subjects are as transient as experiences, then this problem is avoided and there will be no need to argue against Hume (again) that there actually is an impression of something persisting.

In chapter 5 (section 4.1), I explain Strawson's view in more detail and argue that it might also be required in order to bypass the combination problem. I will discuss a further problem related to it in chapter 7 (section 1), concerning whether there could be proper causal relations of the kind I have described within a transient and deflated subject.

2.5 Agent-Causation and the Unified Subject

Agent-causation does not require that agentive subjects persist, i.e., are unified over time. What it does seem to require, however, is that subjects have a momentary unity as they exist. Firstly, this is because subjects are fundamentally characterized by a unity of consciousness within which all its experiences are contained. Acceptance of the momentary unity of subjects may be what fundamentally distinguishes deflationary views of the subject from eliminative views such as the bundle theory, and the agent-causal view is clearly not compatible with an eliminative view. Secondly, while agentive subjects can be complex in the sense of experiencing many different motives at the same time, when causally acting it only exerts one single effort at a time. Therefore, agentive subjects seem unified also in virtue of the structure of the power exerted by them.

As discussed in chapter 1 (section 3.1), panpsychism faces a combination problem. It has to explain how complex macrosubjects can either be constituted by a number of microsubjects, or strongly emerge from them. Agents being strongly unified in the way just described suggests that macrosubjects, who are agents, must result from strong emergence and not constitution. In particular, the unity of the powers or efforts exerted by agents seems too strong to be merely a result of aggregation of individual powers.

In this way, accepting agent-causation intensifies the combination problem for constitutive panpsychism. If we accept panpsychism on the basis of the argument from

causation we also accept a further reason to think macrosubjects cannot be explained via constitution. It may also be seen to somewhat intensify the empirical aspect of the combination problem for emergent panpsychism. As discussed in chapter 1 (section 3.1.2), a principle of microphysical, as opposed to mere physical, causal closure might entail that emergent macrosubjects are epiphenomenal or overdeterminers if their parts are still taken to be microsubjects. If one accepts panpsychism on the basis of the argument from causation, it becomes even clearer that this is the case and that epiphenomenalism is not an option. Agentive macrosubjects must have emergent causal powers, powers that are not constituted by the powers of its parts. If this is correct, then the argument from causation is dependent on there being a coherent theory of combination in terms of strong (but not brute) emergence of macrosubjects that supplant their parts, i.e., a fusion account. But – fortunately enough! – in the next two chapters, I will propose and defend a theory of combination of this kind.

2.6 Transeunt Causation

I have argued that we only experience necessary connections between motives and efforts within a subject that acts as a causal nexus. But on what basis can we understand causal relations between different subjects? This is the problem of inter-substantial, inter-subjective or transeunt causation. As noted in chapter 2, Leibniz thought this was impossible, and replaced real transeunt causation with a pre-established harmony.

In order to avoid this outcome, there are three main options for my position: (1) to take a mysterian, or skeptical realist, attitude toward transeunt causation, (2) to argue that we can infer a positive conception of the nature of transeunt causation from our acquaintance with the nature of powers in effort, and (3) to argue that we can abstract away from immanent causation, i.e., causation within a subject, a positive account of transeunt causation.[5] I think all three options are acceptable. They could also all be true at the same time, if some aspects of transeunt causation can be rendered partially intelligible in virtue of thesis (2) and (3), but other aspects remain mysterious.

[5] A fourth option is to say that for all transeunt causation there is an overlapping subject containing the experiences of both interacting subjects. Relative to the overlapping subject the causation is then immanent and intelligible. I think this is ultimately untenable for reasons such as: (1) Overlapping agentive subjects would give a serious exclusion problem. (2) Sharing of experiences is incompatible with phenomenal essentialism and phenomenal holism, both of which I find plausible (see Basile's argument discussed in chapter 1, p. 46).

According to the anti-mysterian premises from the different versions of the argument of causation (premise 4 from the arguments from dispositional and non-structural properties and premise III from the first formulation), the nature of causation is not unknown. But this does not necessarily rule out option (1), the view that we cannot know the nature of transeunt causation. One might read the premises as requiring only that the nature of causation is not *wholly* unknown. If the transeunt aspect of causation is unknown, but the immanent aspect is known, this makes the nature of causation *partially* known. However, one might find it objectionable that transeunt causation should be regarded as an aspect of causation and not as its own property or phenomenon, or that there is a distinction between aspects and properties with this kind of significance in the first place. Another way of looking at it, then, which does not rely on this subtle distinction, is seeing how positing unknown transeunt causal properties is at least compatible with the motivation of the anti-mysterian premises. The premises are motivated on pragmatic, methodological or anti-skepticist grounds according to which we should avoid positing unknown or not positively conceivable properties when possible, i.e., when theories in terms of knowable or positively conceivable properties are available. If options (2) and (3) fail, then there might be no positive conceptions of transeunt causation available and mysterianism or skeptical realism with regards to it cannot be avoided (given that Leibnizian eliminativism about transeunt causation is regarded as unacceptable). As discussed in the context of the first formulation of the argument from causation (chapter 3, p. 92), this reasoning could be made explicit in modified versions of the arguments, which would then be formally compatible with mysterianism or skeptical realism about transeunt causation.

It might be objected that if transeunt causation has a wholly unknown nature, it is not mental in any non-trivial sense. The result would then be a dualism of mental immanent and non-mental transeunt causation. However, this is not the kind of dualism that gives overdetermination, epiphenomenalism or problematic interaction. Transeunt and immanent causation complement each other; they do not compete. This view could go well with impure panpsychism, the kind of panpsychism according to which some physical properties are not grounded in mental properties (see chapter 1, p 9), because physical properties can then be understood as the properties that enable transeunt causal relations.

It is possible that even though we cannot directly experience necessary connections between efforts and result, our acquaintance with the nature of powers in making efforts

lets us infer a positive conception of the necessary connection between them, as per option (2). According to causal dispositionalism, powers necessarily produce the effects they are directed at if there is no interference. This could be a substantial truth and not a mere conceptual truth. If it is a substantial truth, our acquaintance with the nature of powers could be the source of our knowledge of it.

Some support can be found for the claim that the connection between efforts and results in the absence of interference or resistance is real, *a priori* knowable and substantial. James Woodward (2005) has developed and defended an influential theory of the epistemology[6] of causation known as interventionism or manipulability theory. An intervention, in Woodward's terminology, is basically an ideal experiment, where we take control over a causal variable by isolating it from all other causal relationships except the one we are going to test. Interventions cannot be defined in non-causal terms, but are essential to uncovering causal relationships, according to Woodward; if we start from mere correlations that we cannot assume in advance have the underlying causal structure of an intervention, we can never infer causation.

Woodward claims that we all have a primitive, i.e., *a priori*, belief that our actions have the causal structure of an intervention, and are causes of the effects that immediately follow them:

> […] human beings (and perhaps some animals) have (i) a default tendency to behave or reason as though they take their own voluntary actions to have the characteristics of interventions and (ii) associated with this a strong tendency to take changes that temporally follow those interventions (presumably with a relatively short delay) as caused by them. (Woodward 2007: 29)

This *a priori* belief is also reliable:

> […] it seems plausible that many voluntary actions do, as a matter of empirical fact, satisfy the conditions for an intervention. (Woodward 2007: 29)

Its reliability as well as its practical usefulness in being a starting point for causal inference strongly indicate that the belief is not only *a priori* but also true and substantial.

Interestingly, Woodward claims that it is our phenomenology of agency which triggers the primitive belief that our actions are interventions:

[6] Woodward mainly considers how we arrive at knowledge about particular causal relations and leaves open the metaphysical question of the nature of the causation.

> […] subjects must have some way of determining (some signal that tells them) when they have performed a voluntary action and this signal must be somewhat reliable, at least in ordinary circumstances. […] human subjects do have a characteristic phenomenology which is associated with voluntary action – they typically have a sense of agency or ownership of their behavior that is not present when they act involuntarily. (Woodward 2007: 29)

There is a sense in which our phenomenology of agency tells us, then, that our actions are likely to be successful interventions, i.e., that our efforts do tend to produce the results they are immediately aimed at.

Although Woodward does not mention it, it seems clear that if we experience the feeling of resistance we will not take our voluntary actions to be interventions after all. If we experience resistance, we will infer that something is in the way of our taking sole control of the causal variable we were trying to isolate, and that we are less likely to be in the process of making a successful intervention. On this basis we could infer that interferers, even when not experienced as resistance, will block interventions.

If Woodward's theory is correct, then, it would be plausible to say we have *a priori* substantial and reliable beliefs about whether our actions are likely to succeed in what they aim at, and that these have their source in our phenomenology of agency, including the phenomenology of our efforts being resisted. One way of explaining this is by saying that evolution has set us up so as to automatically and instinctively judge that our actions are interventions whenever we experience the phenomenology of unresisted efforts. This would mean that the phenomenology does not subjectively rationalize and justify the belief – it is a mere causal trigger of the belief. Another way of explaining it is by saying that the phenomenology of agency and resistance lets us become acquainted with the nature of powers and understand that this is how they work. On this explanation, the trigger of the belief also rationalizes and justifies it.

If it is already accepted that we are acquainted with causation in agency, maybe the latter type of explanation would be the best one. If so, there would be an argument for thesis (2) available on the basis of Woodward's theory. I will not go further into reasons for thinking this explanation is really the better one, or reasons to accept Woodward's theory. This has been only to indicate how an argument for thesis (2) could proceed.

Thesis (3), that we can abstract away from immanent causation a positive conception of transeunt causation, is (as already mentioned) not excluded by thesis (2), and they could also helpfully complement each other. In the next chapter, I will suggest some

principles by which transeunt causation, considered as either mental or physical, would be partially intelligible, in order to enable an account of mental causation as a causal process which avoids the intelligibility aspect of the combination problem. In chapter 7, where I discuss how the argument from causation and my account of combination as causation relate in general, I will explain how these principles can perhaps be seen as abstractions from immanent agent-causation, and thereby arrive at yet another possible answer to the objection from the intelligibility of transeunt causation.

5

Combination as Causation: the Intelligibility Problem

The combination problem threatens to undermine the line of argument for panpsychism from philosophy of mind. In the following two chapters, I will argue that understanding mental combination as a causal process enables a solution to it. This chapter will be concerned with the intelligibility aspect of the combination problem, and the next chapter with the empirical aspect. I will begin by reviewing and extending my analysis of the combination problem from chapter 1.

1 FURTHER ANALYSIS AND OUTLINE OF MY SOLUTION

According to the Hegelian argument, panpsychism avoids the main problems of both physicalism and dualism. Physicalism mainly has an intelligibility problem: how can the mental be grounded in the physical, given the epistemic gap between them? Dualism mainly has an empirical problem: how can the mental be causally efficacious when evidence tells us that the physical is causally closed? Physicalism additionally has an empirical problem of exclusion: if not only the physical but the microphysical is causally closed, how can macrophysical mental properties be causally efficacious? Dualism and emergentism additionally have an intelligibility problem: how can the mental be produced by the physical, either as an emergent phenomenon, or as mental effects of physical causes? Panpsychism appears to avoid all these problems. But so do other versions of Russellian monism, such as mysterianism and neutral monism. Panpsychism can be regarded as preferable to these on the basis of either the demonstration of an epistemic gap between neutral or mysterian and mental properties relevantly similar to the gap between physical and mental properties, or the principle that positive theories which actually give explanations are to be preferred over theories that appeal to unknown or unknowable properties or only provide a schema for explanation. This latter principle can also be regarded as the main motivation behind the *Anti-mysterianism* premise from the arguments from causation and from categorical/dispositional/non-structural properties.

Therefore, I proposed to understand the argument from philosophy of mind as the claim that panpsychism has the advantage of being compatible with the following principles, which are jointly incompatible with any other (so far suggested) view:

(i) There is an epistemic gap between mental and physical properties that is not closable in principle.
(ii) Epistemic gaps that are not closable in principle entail ontological gaps.
(iii) There is nothing about the physical in virtue of which the mental can non-brutely emerge.
(iv) Brute emergence is impossible.
(v) The mental is not epiphenomenal.
(vi) There is no systematic overdetermination.
(vii) The physical is causally closed.
(viii) The mental is not grounded in or emergent from properties whose nature is unknown or not positively conceivable.

The respective problems of the other views can be put in terms of these principles as follows: Physicalism's intelligibility problem is its dilemma between inexplicable grounding and brute emergence. The first horn requires the denial of (i) or (ii), and the second horn requires the denial of (iii) or (iv), as well as an arguable collapse into dualism. Dualism's empirical problem requires the denial of (v) by affirming epiphenomenalism, the denial of (vi) by affirming overdetermination, or the denial of (vii), by affirming interactionism. Physicalism's empirical problem, the exclusion problem, also requires the denial of (v), (vi) or (vii); however, this presupposes a strong reading of (vii) as stating the principle of not only physical but microphysical causal closure.

All non-panpsychist forms of Russellian monism will be in conflict with (ii) or (iv) given that one endorses the strong principle, defended by Strawson, according to which the mental cannot be grounded in or necessitated by the non-mental. However, this principle is controversial among defenders of panpsychism. Somewhat less controversially, one could accuse non-panpsychist Russellian monism, except positive neutral monism, of having to deny (viii) by positing unknown non-mental Russellian properties. Positive neutral monist views can be eliminated via the independent

demonstration of an epistemic gap between them and the mental, in combination with principle (ii).

The combination problem is the problem of explaining how macromentality arises from micromentality without facing problems strongly analogous to those it is claimed to solve. In other words, it is the problem of accounting for mental combination without running into conflict with some of the principles (i)–(viii), and thereby cancelling panpsychism's advantages according to the overall argument from philosophy of mind. There is reason to believe that the main kinds of panpsychism, constitutive and emergent, both run into such conflict, and that each inherits both an intelligibility problem and an empirical problem.

Constitutive panpsychism's intelligibility problem is analogous to the inexplicable grounding horn of physicalism's dilemma, and this constitutes conflict with either (i) or (ii). There is an epistemic gap between the micromental and the macromental that can be demonstrated in the same way that the gap between the physical and the mental can be demonstrated. If it is closable in principle, the fact that it is not actually closable must mean that we are ignorant about either the nature of our own macromentality in some respects (as Strawson suggests) or about the nature of micromental properties or the relations between them (as Chalmers and Goff suggest). Specific, perhaps indecisive, worries can be raised about each proposal, but both face the general objection that appeal to ignorance is in tension with principle (viii).

Emergent panpsychism's intelligibility problem is an analogue of the brute emergence horn of physicalism's dilemma and this constitutes conflict with (iv). The emergence of the macromental from the micromental looks just as brute as the emergence of the mental from the physical – or at least, the intuition some have that it is not brute has not been properly spelled out and defended.

Constitutive panpsychism's empirical problem is an analogue of the exclusion problem and this constitutes conflict with (v), (vi), or the strong reading of (vii) which affirms microphysical causal closure. If the microphysical is causally closed, then the micromental should also be causally closed, and the micromental would thereby threaten to causally exclude the macromental.

Emergent panpsychism's empirical problem is an analogue of dualism's problem of physical causal closure, except that it only arises for emergent panpsychism if the microphysical is causally closed. It constitutes conflict with (v), (vi), and the strong reading of (vii). If the macromental is an emergent phenomenon distinct from its

micromental base, but there are no emergent macrophysical phenomena, then macromental emergent phenomena must either be epiphenomenal, overdeterminers, or, if they supplant their micromental bases, it is hard to see how they can ground physical structure which does not clearly exhibit such patterns of emergence.

In this chapter and the next, I will not argue that all the principles (i)–(viii) are certainly true. Neither will I argue that physicalism, dualism, and non-panpsychist Russellian monism are unavoidably in conflict with some of them and that the arguments discussed in chapter 1 in particular succeed in showing this. My aim is only to defend panpsychism against the charge that *if* physicalism, dualism, and non-panpsychist Russellian monism indeed face such problems, as the arguments discussed in chapter 1 would show if they are sound, then so does panpsychism. In other words, I will argue that panpsychists can in fact account for mental combination without facing any problems that are strongly analogous to the main problems of physicalism, dualism and non-panpsychist Russellian monism, and that panpsychism therefore cannot be accused of merely moving and repeating the problems that it is argued to solve.

The account I will propose is a version of emergent panpsychism. Most panpsychists seem to prefer constitutive panpsychism over emergent panpsychism, and I am ready to grant that *if* constitutive panpsychism can be made to work, it would clearly have some advantages over an emergent version.[1] My own suspicion, though (for what it is worth), is that constitutive panpsychism cannot avoid its combination problems because it preserves too many of the features of physicalism – not only the monism, but also the reductionism. Getting entirely away from the anti-reductionist arguments usually directed against physicalism (or strong analogues thereof) requires not just changing the reduction base from non-mental matter to mental matter, but freeing panpsychism of reductionism altogether. Not only will mentality have to be fundamental, but, in a certain sense, our variety of it will have to be as well – as will be the result of emergent panpsychism.

In this chapter, I will argue that the macromental can emerge from the micromental while avoiding the intelligibility problem. In the next chapter, I will argue that it can avoid the empirical problem. Emergent panpsychism's intelligibility problem is its

[1] These might include the advantage of making macroconsciousness fully intelligible relative to microconsciousness (on my emergent account, it is only partially intelligible, as will be explained shortly), and causally integrating it in a more systematic way given that the microphysical and not only the physical is causally closed (as will be discussed in chapter 6).

apparent conflict with (iv), the principle that rules out brute emergence. Solving it requires explaining how the emergence of the macromental can be non-brute.

Examples of non-brute emergence are often really matters of highly complex or otherwise epistemically impenetrable constitution. The emergence of liquidity of non-liquid elements, patterns generated by cellular automata, such as the Game of Life, and phenomena that result from deterministic chaos are all examples of in principle intelligible constitutive grounding, or *weak* emergence.[2] As mentioned, *strong*, non-constitutive, emergence is when the properties of a whole are not predictable in principle from complete knowledge of the properties and configuration of its parts as these properties manifest in isolation or in different wholes. There are various accounts of what kind of non-constitutive determination relation would have to underlie strong emergence. I will suggest that when it comes to the emergence of the macromental the nature of the determination relation must be causal, i.e., that micromentality *causes* macromentality. This is helpful with respect to the intelligibility problem, because causation is arguably a *partially*, but not fully, intelligible relation. On an account where macromentality is caused in a partially intelligible way, then the coming into being of *our* consciousness would be less intelligible than on a constitutive account, but more intelligible than on an account that offers only brute emergence. Bruteness equals complete absence of intelligibility, and demonstrating partial intelligibility is therefore sufficient to avoid brute emergence and thereby the conflict with principle (iv), which constitutes the intelligibility problem.

This chapter will proceed as follows: I first explain why, and in what sense, we should think that causation is partially intelligible in the first place. Then I will offer an account of the partial intelligibility of causation, as well as the unintelligibility of radical emergence – the kind of emergence that would have to underlie the psychophysical relation given physicalism, according to the arguments for panpsychism – in terms of a broadly Aristotelian theory of change. This suggestion might sound obscure, but I will defend it by pointing to how it seems already to some extent implicitly presupposed within philosophy of mind, and showing that it fits well with science. After having thus spelled out and defended the account of the partial intelligibility of causation on a general

[2] Some patterns and events that emerge from cellular automata and deterministic chaos are in principle not predictable except by simulation; however, simulation is a form of prediction, and therefore the emergence counts as weak/constitutive. There are principled reasons besides complexity for why they are not (always) predictable except by simulation (see, e.g., Seager 2012: 79–80).

basis, I will show how mental combination can be construed as a causal process that complies with it. It seems this is possible only if we adopt a deflationary view of subjects, aspects of which have been defended by Strawson and Parfit. I will show that when understood in such a way, it is possible for microsubjects to combine into macrosubjects through a causal process of fusion or blending – in a way more akin to how rivers combine into a sea than to how bricks combine into a wall.[3]

My argument will not be based on the account of the intelligibility of causation defended in chapters 3 and 4. There I argued that causation as experienced in agency is intelligible – we understand *why* motives such as pain tend to lead to efforts such as avoidance attempts. However, this was an account of immanent causation, causation within the same subject, only. If microsubjects cause macrosubjects to come into being, this would be a matter of transeunt causation, causation between different subjects. I will therefore propose a different account of the intelligibility of causation, considered as transeunt and with any mental aspects abstracted away. This account can be accepted independently of the arguments and theses defended in the previous chapters. As mentioned already, in chapter 7 I will explain how my accounts of immanent and transeunt causation nevertheless systematically relate. Until then, I will set aside the view that causation can be experienced in agency in a way that reveals it as intelligible and mental. In chapter 6, however, where I will discuss emergent panpsychism's empirical problem, the view that we experience our own causation in agency will play a role, but not the role of revealing it as intelligible or mental.

2 THE PARTIAL INTELLIGIBILITY OF CAUSATION – WHY AND IN WHAT SENSE?

Why should we think that causation (that is, transeunt causation which is not necessarily mental), is partially intelligible? In the context of the combination problem, mainly because it seems to be a presupposition of the argument of which the intelligibility problem from emergent panpsychism is an analogue, namely the argument from non-emergence, put forth by Strawson and Nagel. According to this argument, consciousness appears brutely emergent relative to the physical, and brute emergence is impossible (as per principle (iv)). However, according to reductionism about causation, causation itself is a metaphysically brute relation, i.e., in principle unintelligible: effects follow causes for

[3] Basile (2010: 189) offers this comparison as a suggestion for how to think of mental combination.

no underlying reason (cf. chapter 3, section 1.1, p. 93-95). If, as reductionists are committed to (given that causation is ubiquitous), the great majority of dependence relations between distinct properties are brute, it can hardly be objectionable that the psychophysical relation is one of them, as emergentists are accused of affirming. There is no reason to think brute emergence is impossible if brute causation is possible.[4] Therefore, the argument from non-emergence (as well as the justification for principle (iv)) must depend on causation not being metaphysically brute.

Strawson clearly hints to it being a presupposition of the argument that causation is not metaphysically brute, claiming that the process by which the notion of brute emergence has gained currency:

> […] is underwritten by the wild radical-empiricism-inspired metaphysical irresponsibilities of the twentieth century that still linger on (to put it mildly) today and have led many, via a gross misunderstanding of Hume, to think that there is nothing intrinsic to a cause in virtue of which it has the effect it does.[footnote omitted] (Strawson 2006b: 19)

At the same time, Strawson holds that causation, while not metaphysically brute, is entirely epistemically brute – referring to "Hume's wholly correct, strictly epistemological claim – that so far as we consider things *a priori* 'any thing may produce any thing'" (Strawson 2006b: 19, footnote 34). This is the skeptical realist view about causation, according to which causation has a nature beyond regularities (and similarity relations across possible worlds), but this nature cannot be known.

Now, full commitment to skeptical realism about causation does not seem entirely compatible with the argument from non-emergence. If we have absolutely no insight into the nature of causation, into what may produce what, how can we say with certainty that the physical (i.e., the narrowly/t-/physicSal) cannot causally produce the mental? The negative knowledge that certain phenomena could *not* be causally related must constitute or be based on some limited insight into the nature of causation. Those who reject brute emergence are thereby committed to causation being actually (not just in principle) partially intelligible. If causation were entirely epistemically brute, if we knew nothing at all about its nature or inner workings, then what basis could there be for the negative

[4] To be precise, there would be no reason to think brute *strong* (non-constitutive) emergence impossible. There would still be reason to think that brute *weak* (constitutive) emergence is impossible, namely that weak constitutive emergence is not brute by definition.

judgment that the argument from non-emergence depends on, i.e., the judgment that the non-experiential cannot produce, causally or otherwise, the mental?

If the argument from non-emergence presupposes that causation is (actually, epistemically) partially intelligible, then its analogue, the intelligibility problem for emergent panpsychism, presupposes it too. If causation is not partially intelligible, then both problems would dissolve. The outcome for emergent panpsychism would be more or less neutral, dialectically speaking: it would lose one important argument in its favor, but also one if its most serious problems. It would still have the support of the arguments from mental causation – insofar as it can get around its empirical problem – and categorical/dispositional/non-structural properties. Therefore, I will not attempt to defend the view that causation must be partially intelligible. I will instead just hedge the question: if causation is partially intelligible, in what does its intelligibility consist?

2.1 Intrinsic and Extrinsic Intelligibility

In a sense, everyone, even reductionists and skeptical realists, agree that causation is actually intelligible, or confers intelligibility to its relata: giving the causes of a phenomenon constitutes explaining it, and explaining something means rendering it intelligible. Also, independently of their views about causation, most philosophers find the idea that something physical can produce something mental more mysterious than the idea that something physical can produce something else that is physical. If ordinary, physical causation is comparatively non-mysterious in contrast with the psychophysical relation, it must also in a sense be partially intelligible.

But this is not necessarily the same sort of intelligibility that is presupposed by the argument from non-emergence. The intelligibility of causation, as well as the comparative unintelligibility of the psychophysical relation, can be acknowledged by reductionists and (fully committed) skeptical realists if it is accounted for in purely *extrinsic* terms. An account of the extrinsic intelligibility of causation, and the extrinsic unintelligibility of psychophysical emergence, would go roughly as follows: A causal relation obtains when two events or entities are correlated in a way that fits into a unified system of laws, and being thus integrated constitutes being rendered intelligible. Highly extrinsically intelligible causal relations are instances of laws or generalizations with universal validity, laws that are part of a simple set of laws which subsume many phenomena. The psychophysical relation, on the other hand, is not very well integrated with other causal relations. The mental looks like a nomological dangler, causal dead-end or otherwise

inelegant complication. It would be an instance of a law or generalization that seems to apply to only one domain, e.g., the biological or neurological, so it is not universal. Furthermore, it explains just one phenomenon, consciousness, and nothing else, so it does not increase systematicity and unity in exchange for expanding the set of fundamental laws. This constitutes its lack of intelligibility and explains why it is often categorized as not causal, in the ordinary sense, but emergent.

If this were really all there was to the intelligibility of causation and the unintelligibility of the psychophysical relation, then the argument from non-emergence, which is a complaint about the *intrinsic* bruteness of the psychophysical relation, would also dissolve, or collapse into arguments from mental causation. The argument from non-emergence implies that the mental emerging from the non-mental is unintelligible in and of itself, regardless of how often and systematically it tended to happen, and is thus a worry about intrinsic, not extrinsic, intelligibility. In turn, the intelligibility aspect of the combination problem for emergent panpsychism would also dissolve, or collapse into its empirical problem. Demonstrating the extrinsic intelligibility of the micro–macromental relation requires only integrating it into the causal structure of the world as science reveals it, and solving the empirical problem amounts to the same. Again, the outcome for emergent panpsychism would be neutral, dialectically speaking. Therefore, I will hedge the question again: if causation is partially *intrinsically* intelligible, in what does its intelligibility consist?

2.2 Ordinary and Emergent Causation; Radical and Brute Emergence

Before proceeding to suggest an answer to the question just posed, I will clarify my use of some terms, and introduce some distinctions. The distinction between extrinsic and intrinsic intelligibility enables a distinction between two types of emergence, *radical emergence* and *emergent causation*, as follows:

	Intrinsic intelligibility (to us)	**Extrinsic intelligibility**
Radical emergence	None	(not defined)
Ordinary causation	Some/partial	High
Emergent causation (strong, but not radical, emergence)	Some/partial	Low

Radical emergence is the relation that, according to the argument from non-emergence, would have to underlie the psychophysical relation given physicalism or emergentism. I will understand radical emergence as a relation which is wholly intrinsically unintelligible, in principle, *for creatures like us*[5] – in terms of facts we are equipped to grasp, there is nothing about the base in virtue of which the emerger results. The reason for understanding radical emergence in terms of unintelligibility *to us* is so that radical emergence will not *by definition* be brute, and hence always impossible, according to principle (iv) above. *Brute emergence* I will understand as wholly intrinsically unintelligible *full stop,* or even to God, as it were[6] – there is nothing whatsoever about the base in virtue of which the emerger results. I take it that in at least some cases, as in the case of the psychophysical relation, it is reasonable to infer that an instance of radical emergence would also be an instance of brute emergence.[7] For the purposes of this chapter, it could readily be granted that all radical emergence is brute, because I will argue that micro–macromental emergence is neither,[8] but the distinction will still be useful to have available.

[5] In what respects must creatures be like us? Intelligibility to us is supposed to contrast with (the metaphorical) intelligibility to God. In order to be minimally restrictive, but preserve the contrast, one could substitute "finite or natural creatures" for "creatures like us".

[6] Strawson explains brute emergence in these terms, but emphasizes that "'intelligible to God' isn't really an epistemological notion at all, it's just a way of expressing the idea that there must be something about the nature of the emerged-from (and nothing else) in virtue of which the emerger emerges as it does and is what it is" (2006b: 15).

[7] Roughly, it seems to me these would be the cases where there is reason to think we have an adequate grasp of the base, the emerger or both, so if their relation were intelligible at all, it should be intelligible to us. For the psychophysical relation, the inference from radical to brute emergence is licensed by our grasp of what physical (*qua* structural or fundamentally non-mental) properties are supposed to be, and our acquaintance with the nature of mental properties.

For reasons to be discussed below, there are other scenarios that could also qualify as radical emergence, such as divine creation or the Big Bang (when understood in a certain way). In these cases, it is arguable that we cannot infer that the radical emergence is also brute: since we have no adequate grasp of the base (either God or the nothingness prior to the Big Bang) we cannot conclude that there is nothing whatsoever about them in virtue of which the emerger, in this case the universe, can emerge. On the other hand, it is also arguable that these types of emergence would also be brute. Perhaps we have a good enough grasp of divine properties such as omnipotence to say that it would lead to contradictions (such as the problem of evil), or of the nothingness that some suppose precedes the Big Bang, to say that there is no way in which the universe could emerge from that.

[8] If one admits that micro–macromental emergence is radical, it seems unavoidable that it is also brute (for reasons of the kind discussed in the previous footnote 7). Also, even if it were arguably radical but not brute, this theory would be a version of McGinn-type mysterianism and conflict with principle (viii).

Causation, on the other hand, can (for the reasons discussed so far) be assumed to be partially intrinsically intelligible for creatures like us,[9] and thereby not brute – there is something intrinsic to the cause, of which we have at least a partial grasp, in virtue of which the effect results. Within partially intrinsically intelligible causal relations, however, one can distinguish highly extrinsically intelligible relations from less extrinsically intelligible ones – or relations that fit very systematically in with other causal relations from relations that would fit in less systematically. I will use the term *ordinary causation* for the former type and the term *emergent causation* for the latter type.

Emergent causation is an important notion, because emergent causation is a kind of strong, but not radical, emergence. If the only kind of strong emergence that emergent panpsychism is committed to is non-radical emergent causation, which is partially intelligible, then it avoids the intelligibility aspect of the combination problem. Emergent causation still represents an empirical problem – what would be its particular structure, and how can it be integrated into the causal structure of the world as science reveals it? – but this has been set aside until the next chapter. In this chapter, I will only consider whether mental combination can be intrinsically intelligible to us – whether it can occur without radical emergence. From now on, I will use the term intelligibility, unless otherwise noted, to mean intrinsic intelligibility to us.

3 THE INTELLIGIBILITY OF CAUSATION AND THE UNINTELLIGIBILITY OF RADICAL EMERGENCE

If causation is a least partially intelligible (again, considered apart from the view defended in chapters 3 and 4), then how or by what principles is this so? What makes it more intelligible than radical emergence – while at the same time not as intelligible as constitution?

There are many ways of differentiating various notions of emergence from causation. For example, causation is often supposed to be diachronic (causes preceding effects), and emergence synchronic (the emerger being simultaneous with its base). However, philosophers have defended both diachronic emergence (e.g., Humphreys 1997; Stephan 2002; O'Connor and Wong 2005) and synchronic causation (e.g., Kant – see chapter 7, section 1), so these do not seem to be essential characteristics and, in any case, not

[9] And presumably fully intelligible to God.

something that confers intrinsic intelligibility. In general, structural differences seem relevant to extrinsic intelligibility only.

Strawson makes a few further claims about the nature of radical emergence. It requires that real physicalists (who think that experience is real and physical) "throw away the conservation principles and say that brand new physical stuff (mass/energy) is produced or given rise to when experiences are emergent from the non-experiential" (Strawson 2006b: 23–24). Non-radical, acceptable emergence, in contrast, is "essentially conservative in the sense of the conservation principles" (Strawson 2006b: 24). Radical emergence is emergence from non-existence – and "*ex nihilo nihil fit*, whatever anyone says (Nobel Prize winners included)" (Strawson 2006b: 19, footnote 34).

This brings to mind Aristotle's theory of change, which he offers in response to Parmenides. Parmenides was the first to argue that nothing comes from nothing – *ex nihilo nihil fit.* He holds that creation out of nothing is not only unintelligible but also impossible. Furthermore, according to Parmenides, all change is necessarily creation out of nothing, and therefore change is impossible. The reasoning behind this seems to be that if a state constitutes a change at all, it would have to be a result of its own negation. Change is when A comes from non-A, which is nothing relative to A. If A comes from A, it does not come from nothing, but then there is no change after all.

Aristotle agrees with Parmenides – and Strawson – that creation out of nothing is impossible, but offers a tripartite analysis of change according to which change can occur without it. Change involves not two elements, a state and its negation, but three: two *forms* and one portion of *matter*. A change consists in the *same* matter passing from one form to another. If A and non-A are both forms of the same matter B, they would not be nothing relative to each other after all, because they have B in common, but neither would they be identical, so that change is negated.

According to Aristotle, all substances are composites of matter – *hylē* – and form – *morphē,* hence hylomorphism. In most changes, a substance remains persistent, and thereby both form and matter, as when a man turns from being unmusical to being musical. In some changes, the substance itself is transformed and only matter remains, as when a seed transforms into a tree. In this change, however, some primary body or element remains persistent. But Aristotle further holds that the primary bodies, earth, air,

water and fire, are also transformable into each other (Cohen 2012: 219). What remains persistent in this change is *prime matter*.[10]

Aristotle claims that prime matter is not perceptible, but is knowable only "by analogy" (Aristotle 1984: 191a9). Many interpreters have taken it as a kind of pure potentiality (Cohen 2012: 220), unable to exist alone, without a particular form that actualizes it. Prime matter, and only prime matter, is ultimately permanent; no forms or substances, composites of matter and form, are fundamental and incorruptible. He seems to derive a conservation law for prime matter (Aristotle 1984: 192a25–34, 318a24–25).

The dispositional theory of causation, discussed in the previous chapters, derives from Aristotle. It should therefore not be too surprising if also this aspect of Aristotle's view has its place in contemporary non-reductionist metaphysics of causation. I will now argue that an Aristotelian, hylomorphic theory of change could be the basis for the distinction between radical emergence and causation, and that on such a basis causation would be partially intelligible.

3.1 Causation as Conservation of Matter and Continuity of Form

I will propose the following distinction between causation and radical emergence: causation always relates, or is always ultimately grounded in relations between, different determinates of the same fundamental determinable, or different forms of the same kind of matter or stuff. Matter should be understood as a fundamental and concrete determinable, which does not necessarily have any properties of actual physical matter, such as extension, mass or inertness. A form should be understood as any determinate of such a determinable – such as a structure, arrangement, configuration, shape or quality. Causation is the kind of change through which matter in this sense remains persistent, while form varies. Radical emergence, on the other hand, is the kind of change through which nothing remains persistent. It is when a new fundamental determinable, or new matter or stuff, comes into existence. This gives rise to three important features of causation: qualitative conservation, quantitative conservation and continuity. Radical emergence will have none of these features.[11]

[10] Prime matter is, as was already remarked by Alter and Nagasawa as quoted in chapter 1, a controversial notion. Many interpreters claim that it is incoherent, but nevertheless, it is clear that Aristotle posited it in some sense (Cohen 2012: 220).

[11] Pat Lewtas (2012) has also argued (in a conference presentation) that the notion of conservation is needed in order to distinguish acceptable modes of combination from radical emergence. He offers various suggestions as to how combination can work within this and other constraints, however, his overall

3.1.1 Qualitative Conservation of Matter

Matter is a fundamental determinable, and a fundamental determinable is a determinable which cannot also be conceived as a determinate of another determinable. For example, color is a determinable but not a fundamental one. It is a determinable relative to redness, but can also be conceived as itself a determinate of, say, sensory or secondary qualities. A molecule is a determinate organization of atoms,[12] which are again determinate organizations of quarks, and quarks are determinate forms of whatever fundamental physical stuff there is, the stuff which is either identical to or transformable into energy. Energy, or the stuff which is in some way a duality of mass and energy, is where the regress of physical determinables end. It is that of which everything physical is a determinable or form, and which is itself a form or determinable of nothing else. Physical causation, causation which relates physical phenomena, would trivially, by definition, always relate determinates of the same determinable if physical energy-matter is a determinable out of which all physical phenomena are constituted. However, on the account I now suggest, causation could not occur at all between the physical and the non-physical if they are not revealed as actually having some deeper determinable in common.[13] This would rather constitute radical emergence.

Radical emergence is when a new kind of matter comes into existence, an entity which is not a determinate of physical energy-matter, and of which physical energy-matter could also not be regarded as a determinate. Radical emergence, whereby a non-physical thing stands in a dependency relation with a physical thing, is closer to creation than causation. An example of radical emergence of the physical from the non-physical is a non-physical God producing a physical world.[14] The reverse relation, the emergence of the mental from the physical, is similarly sometimes likened (e.g., by Strawson, as seen above) to creation *ex nihilo.* The first feature of the account, then, is that new kinds of matter, stuff or determinables are created or radically emerge, while forms or

---

approach is different from mine in many respects, most importantly in assuming constitutive rather than emergent panpsychism.

[12] It might seem incorrect to say that a molecule is a determinate organization of atoms, because atoms are already fully determinate objects. But *organization* of atoms is not fully determinate, because atoms can be organized in different ways. So I take the indeterminacy to be in the organization, not necessarily in the atoms organized.

[13] Neutral monism would be a view according to which the physical and the mental have a deeper determinable in common, the neutral determinable.

[14] This is not necessarily brute emergence, because it is arguably in principle intelligible (but not to us) how God can create the physical, namely in virtue of divine attributes such as omnipotence, infinity, etc.

determinates are caused. In other words, causation is qualitatively conservative when it comes to matter; radical emergence is not.

3.1.2 Quantitative Conservation of Matter

Causation furthermore requires that not only the same *qualitative* kind but also the same *quantity* of matter or a determinable remains constant throughout a change. If a change involved the production of more matter, it would also amount to radical emergence, even if the new matter were of the same kind that the thing it emerged from was made of. An example of radical quantitative emergence would be perpetual motion machines, physical machines that create more physical energy-matter. The second main feature of the account is thus: causation is quantitatively conservative; radical emergence is not.

3.1.3 Continuity of Form

There are indications that all the possible forms of a kind of matter form a continuum, and that any form of a kind of matter could in principle be transformed by gradual steps into any other form. For many determinables, such as color and geometric shape, their determinates form a continuum. In many respects, this seems to also be the case for physical matter. For example, not only can one gradually transform a lump of gold into, say, a gold coin; gold itself can be gradually transformed into other elements such as lead – in principle, any element can be transformed into any other by gradually "rearranging" the subatomic particles of atoms to form different ones. However, it seems one cannot demonstrate that all determinates of any determinable form a continuum, and furthermore, physics provides some possible examples of necessarily non-gradual transitions that are still conservative (to be discussed in section 3.4 below). So a principle of continuity of form must perhaps allow for exceptions or approximations (that will also be discussed below).

However, it will still provide for an important contrast with radical emergence, because for radical emergence there will be no exceptions: discontinuity is a necessary feature of it. Consider first radical qualitative emergence, an event which violates qualitative conservation. One could not gradually transform one fundamental kind of matter into a different kind, because the two portions of matter would have nothing in common that they would both be degrees *of*. Consider then radical quantitative emergence, which violates quantitative conservation. There is no way for additional quantities of matter of the same kind to come gradually into existence either. A quantity of matter can come into being portion by portion, but the individual portions, no matter

how miniscule, would have to appear suddenly, because there is no intermediate state between existence and non-existence. The existence/non-existence distinction is strictly binary – something either exists or it does not.[15] With different forms of the same matter, there are (almost) always intermediate states (as with colors), or, with absolute distinctions, vague, borderline states in between (as with baldness/non-baldness).

Radical emergence thus entails discontinuity, but the entailment might not go both ways – it could be that discontinuity sometimes has other explanations. If there is discontinuity, then the default hypothesis, one might say, should be that the explanation is radical emergence, a crossing of the absolute boundary between non-existence and existence. However, if another explanation can be provided of the discontinuity, or if other signs of the persistence of the same type and quantity of underlying matter or determinable are clear and numerous, then one need not conclude that there is radical emergence. Other signs of persistence of matter could be, for example, that the forms are continuous across many dimensions but not all, or that the forms can be ordered on a spectrum which is discrete but clearly much more fine-grained than the binary spectrum of existence/non-existence.

Therefore, the third main feature, is that causation *tends to* relate possibly continuous states, while radical emergence relates necessarily sharply discontinuous states. In other words, discontinuity is to be regarded as a defeasible indicator of radical emergence.

3.2 Hylomorphic Change as the Ground of Causation

Can hylomorphic change, where forms change but matter remains, constitute or ground all causation? Our ordinary concept of causation allows causation by absence, as when the death of a plant is caused by nobody watering it, which entails that there are causal relata which have no matter at all. It also does not require that causes are always transformed into their effects. For example, when we say that smoking causes cancer, there is no transformation of smoking to cancer. Nevertheless, it is not unreasonable to think that hylomorphic change, i.e., transformations of energy-matter, *underlies* all causal relations.

Energy, as already hinted at, seems to have a number of features in common with Aristotelian prime matter, the substratum of all change: it is imperceptible – not directly

[15] Some philosophers dispute this, and claim that there are degrees of existence. If so, my account could perhaps still work, if one could find some other criterion for recognizing radical emergence than discontinuity. But as it stands, my account is premised on there not being degrees of existence.

measureable, it always exists in a certain form and can never be found alone, it is always conserved – never created or destroyed, and it is a kind of pure potentiality. Theories of causation in terms of transference of conserved quantities, such as energy (see, e.g., Fair 1979; Dowe 1992), arguably have a number of shortcomings as reductive analyses of causation (see Dowe 2008: sect. 6), but it is more plausible that transference of conserved quantities non-reductively underlies all causal relations. Conserved quantities include energy, momentum and charge. It could be that all the conserved quantities together constitute prime matter or are different aspects of it, or that a conserved quantity theory of causation can be formulated only in terms of energy.

How is it plausible that transference of conserved quantities, energy in particular, underlies all causation just as prime matter would? It could underlie causation by absence, insofar as an absence can only cause an event if there is some energy transference in the effect that this absence allows to occur. An absence cannot cause something unless there are further positive causes present as well. To see how energy transformation can underlie causation more generally, consider how the total state of the universe at a moment can be regarded as the total cause of the total state of the universe at the next moment. For this causal relation, the most general one of all, it would be correct to say that the total cause is transformed into the total effect. It is the very same energy-matter that made up the universe a moment ago that makes it up now – in this way its transformation underlies the sum of all causation. More particularly, one might think that all direct, productive causation by concrete positive causes, as opposed to absences, involves the cause being *partially* transformed into its effect by transferring *some* of its energy to it. In this way it could underlie the causal relation between smoking and cancer, insofar as they are linked by a mechanism that would involve the transfer of some of the energy of smoke particles to lung cells, in virtue of which the lung cells are changed in a way that ultimately leads to cancer. If (productive/positive) causation is partial transformation, this means that an object will eventually be wholly transformed, and thereby disappear, if it keeps on causing changes (productively/positively) without being causally affected in return. This is what happens according to physics: if something loses all its energy, it must also have lost all its mass and disappeared.

3.3 Partial Intelligibility in Virtue of Hylomorphic Principles

How does the hylomorphic theory make causation intelligible? And how does it account for the unintelligibility of radical emergence?

If causation is fundamentally a matter of things being made into other things, or states being transformed into other states, we can to a certain extent understand *how* a cause produces its effect: by letting itself, or a part of itself, be transformed into it, or by providing the matter for it. We cannot on this basis understand how a cause is transformed in exactly the way that it is, but we can understand how there can be a transformation of *some* kind. With two things that are made out of different stuff, on the other hand, as in radical emergence, we correspondingly understand why they *cannot* be transformed into each other. A cause cannot provide the matter for an effect if it does not already possess the kind, or quantity, of matter required. Some think creation out of nothing is still possible, for example, because they think this is what God does. But nobody claims to understand *how* God could do this, and it cannot be by transformation.[16]

Hume posits as a criterion of causal intelligibility that it enables *a priori* derivation of effects from causes. A corresponding criterion of partial intelligibility would be that it enables *a priori* derivation of some, but not all, facts about effects from causes. The hylomorphic theory does seem to enable this. Firstly, we can derive from the theory that effects must be made of the same prime matter as their causes.[17] There are no clear empirical counterexamples to this. Furthermore, if we take energy to be prime matter or a fundamental aspect of it, as it seems we should, and prime matter can neither be created nor destroyed (as per the criteria of qualitative and quantitative conservation), only transformed, then the law of conservation of energy follows *a priori*. There are no clear empirical counterexamples to this either.

It might sound suspicious that a law of physics suddenly becomes partly *a priori* on the basis of a metaphysical theory. It could indicate the theory is not really *a priori* after all, but is covertly derived *a posteriori*. However, it is already widely held that the law of

[16] As already noted (footnote 14), divine creation is not necessarily brute emergence, because, in divine creation, there is clearly something about the "base" in virtue of which the emerger is produced, e.g., omnipotence, infinity, or other divine attributes. Insofar as omnipotence and infinity and so on cannot be grasped by finite creatures like us, divine creation would be in principle unintelligible to us and therefore still an example of radical emergence. If divine creation were demonstrably metaphysically brute (as some would argue), it too would be ruled out by principle (iv) above, but there is nothing in this discussion that clearly indicates or depends on this being so (that is to say, the arguments for the bruteness of the psychophysical relation would not be analogous to the arguments for the bruteness of divine creation, and so one can accept the former type of argument and reject the latter). No emergentist would claim that brains literally have a divine power to produce consciousness out of nothing, so not saying anything to rule out the possibility of divine creation should not be a problem for the argument for panpsychism from non-emergence.

[17] Insofar as the causes and effects are made of *something*, i.e., are not absences.

conservation of energy is not purely a matter of empirical discovery, but that it is better regarded as in part being a regulative principle, an *a priori* truth, or even a tautology. Barbara Montero discusses this view and recounts some events in the history of science that render it plausible:

> About a century ago, Poincaré claimed that we would never reject the conservation law for energy because any apparent violation would be rectified by positing a new form of energy. Poincaré took this to show that the law of the conservation of energy is 'outside of the reach of experiment and reduces to a sort of tautology'. And in 1930, to some extent bolstering Poincaré's view, Wolfgang Pauli rectified certain apparent violations of the laws of conservation of energy and momentum by positing a new, virtually unobservable form of energy, the neutrino. Eventually, empirical support did emerge when in 1956 huge detectors revealed that neutrinos exist. Nonetheless, the history of the neutrino indicates that the conservation of energy law, if not tautologically correct, will most likely not be given up easily. (Montero 2006: 391–392)

One might think that the idea that *some quantity* is conserved throughout all change is the regulative, *a priori*, or – as Poincaré thinks – tautological part of the law of conservation of energy, whereas how to define and empirically measure the conserved quantity is an *a posteriori*, empirical matter. At least this is what the history of science seems to show. The identification and measurement of the conserved quantity has been a long and difficult process. Scientists have considered candidates such as *vis viva* (mv^2), force (ma), and many others, before settling on energy (as well as charge and momentum). But many have been explicitly motivated by *a priori* considerations, metaphysical or theological, in their search for the conserved quantity, including Descartes, Joule, Faraday and Helmholtz (Coopersmith 2010). One could also mention Kant, who aimed to clearly demonstrate, in his *Metaphysical Foundations of Natural Science* (1786/2004), how conservation of *matter* was partially an *a priori* truth – in Kant's view, conservation of substance is *a priori*, but the concept of matter is empirical. At the same time, he also argued for a dynamical theory of matter, according to which all its fundamental properties arose from the interplay of forces, so altogether this clearly approximates the law of conservation of energy.

3.4 Hylomorphism and Science

I have already pointed to a number of ways in which the hylomorphic theory of change fits well with science – especially via the conservation criterion. I will now add some further considerations to substantiate this, and address some possible counterexamples.

Patrick Suppes has argued that Aristotle's basic doctrine of matter and change "is correct in a strong sense: it can be used as a basis for interpreting the results of modern science" (Suppes 1974: 27). He argues that other concepts of physical matter, Descartes' and Boscovich's, are inadequate, and that Aristotle's concept (and to an extent Kant's, which he claims is very close to Aristotle's) is supported by, among other things, the following developments in physics:

> As quantum mechanics developed […] it was also recognized that matter was not indestructible, contrary to ancient ideas of an atomic sort, but that it could be converted into energy. […] The pursuit of particles continued and as the energy levels became higher it became apparent that the world is full of particles that are continually undergoing processes of generation and corruption, as Aristotle would put it. Methods for observing this generation and corruption were brought to a fine point by bubble-chamber apparatus and other related methods. […] The empirical evidence from macroscopic bodies and also from high energy particles is that the forms of matter continually change. […] The collisions of electrons and other particles to produce new particles as observed, for example, in cloud-chamber and other experiments is simply good Aristotelian evidence of the change of form of matter. The cloud-chamber data especially support Aristotle's definition of matter. As we observe change there must be a substratum underlying that which is changing. What is the substratum underlying the conversion of particles into other particles, or the conversion of particles into energy? The answer seems to me clear. We can adopt an Aristotelian theory of matter as pure potentiality. The search for elementary particles that are simple and homogeneous and that are the building blocks in some spatial sense of the remaining elements of the universe is a mistake. There is a continual conversion of the forms of matter into each other; there is no reason to think that one form is more fundamental than another. The proper search at a theoretical level is for the laws that describe these changes of form, and not for the identification of elementary particles that are in some fundamental and ultimate sense simple and homogeneous. (Suppes 1974: 46–47)

Suppes here emphasizes the impermanence of individual substances or things. On the hylomorphic view, substances or things, like atoms and particles, are composites of matter and form. Since forms are impermanent, substances must be as well. Science has shown many times that particles first thought to be indestructible, first atoms, then subatomic particles, are impermanent forms of energy which can be transformed into other forms of energy.

Does physics support that all possible forms of energy-matter can be ordered on a continuum? I will argue that it at least supports the more moderate claim that *most*

possible forms of matter can be ordered on an approximate[18] continuum, i.e., a discrete but very fine-grained spectrum. This preserves the contrast with radical emergence – the relata of an instance of radical emergence can never be ordered on even an approximate continuum: the binary spectrum of existence/non-existence is maximally coarse-grained. The moderate claim can be upheld in view of some of the most significant indications of discontinuity in physics, which I take to be the following: (1) the possible discreteness of space or time, (2) quantum jumps and (3) phase transitions.

If space or time is discrete, as some theories in physics entail, then it would not be possible for spatiotemporal states to be ordered on a continuum. But it would be possible to order them on a discrete but very fine-grained spectrum, which is compatible with the approximate continuity principle.

At the quantum level, quantum jumps seem to represent discontinuous events. Bound particles, for example electrons in an atom, have discrete energy levels. The transition of an electron from one energy level to another can therefore not be gradual. However, the energy levels of electrons are not ordinary determinate (i.e., classical) states. They are related to wave functions that give the probability of finding the electron at different locations. In a "jump" between states A and B involving different energy levels, the location of the electron during the transition may be a place in space and time where the probability for finding the electron is non-zero, independently whether the electron is in state A or B. Thus, it is in principle possible that the movement of the electron can be completely continuous both in time and space while the transition from state A to B takes place, even if their energy levels change discontinuously.[19] In other words, a discontinuity between indeterminate, non-classical states does not entail, nor suggest,[20] discontinuity between any determinate states.

If nothing entails discontinuity, if it is in principle possible that the movement of the electron is completely continuous during the quantum jump, this still marks a great difference with radical emergence. With a "jump" between non-existence and existence,

[18] I intend this term in an informal, loose sense. Mathematically, a discrete spectrum of countably many states, even if infinite, cannot really approximate a spectrum of uncountably infinite states, i.e., a continuum.

[19] I have confirmed this with associate professor Arnt Inge Vistnes, Dept. of Physics, University of Oslo.

[20] This is fundamentally because quantum theory, unaccompanied by any metaphysically motivated interpretations, does not suggest *anything* about whether the movements of jumping electrons are discontinuous or continuous; it only gives probabilities of where they are to be found upon measurement.

as in radical emergence, there is no sense in which this *could* have happened continuously, and discontinuity is entailed.

In thermodynamics, phase transitions, such as the transition of water from liquid to gaseous form (vapor), are mathematically modelled as singularities, which are sharp discontinuities. However, as has been debated in philosophy of physics, this model cannot possibly be fully realistic, because it actually leads to a paradox. The paradox appears when the definition of phase transitions from thermodynamics as singularities is combined with statistical mechanics. As Craig Callender puts it: "The problem is that phase transitions as understood by statistical mechanics can only occur in infinite systems, yet the phenomena that we are trying to explain clearly occur in finite systems" (2001: 549). Callender recommends that we should therefore not treat phase transitions as singularities, i.e., discontinuities, after all: "we should say that real finite systems give rise to the sort of behaviour associated with phase transitions in thermodynamics even when the partition function is not singular. After all, the fact that thermodynamics treats phase transitions as singularities does not imply that statistical mechanics must too" (Callender 2001: 550). But what if it turns out the paradox must rather be dissolved in a different way, and the discontinuity of phase transitions is confirmed? Since most physical changes are not phase transitions, this would be compatible with the principle that most possible forms of the same matter can be ordered on a continuum, or that all possible forms can be ordered on a spectrum which is mostly continuous but has some gaps. It also seems different phases will be discontinuous only along some property dimensions, but continuous along many other property dimensions, so the gaps are in this sense not big gaps.

I have assumed that radical emergence does not occur within the domain of physics or physical sciences, and therefore it is important to show that the criteria I have proposed for radical emergence does not indicate it. However, there is one event posited in physics that would qualify as radical emergence, namely the Big Bang, insofar as the Big Bang is really preceded by nothingness (as opposed to, say, a Big Crunch following a previous Big Bang and so an *ad infinitum*). If one holds that brute emergence is impossible, but is not prepared (as seems reasonable) to reject the hypothesis that the Big Bang could be preceded by nothingness on that basis, one could argue that this is an instance of radical emergence which is not brute, i.e., that how the universe could emerge from nothing is perhaps in principle not intelligible to us, but it would be intelligible to God, metaphorically speaking. Admitting the possibility of radical emergence in the Big Bang

would presumably not threaten the argument from non-emergence for panpsychism, because it would not be a plausible defense of emergentism to say that consciousness emerges systematically out of nothing in the same way that the whole universe emerged from nothing. If physical–mental emergence is claimed to be analogous to the Big Bang, it would, if anything, make emergentism seem more implausible, not less.

3.5 Hylomorphism and Philosophy of Mind

I have now shown how hylomorphic criteria of causation map onto the kind of changes we find in physics, where radical emergence does not take place – except perhaps with the Big Bang. Now I will show that the opposite criteria of radical emergence map well onto physicalism's problems with consciousness. Many problems concerning consciousness that arise from assuming physicalism can be expressed by saying that within this metaphysics its appearance would have all the characteristics of radical emergence. Correspondingly, if consciousness could be construed as coming into being in the way that had the characteristics of causation as I have defined it, it seems physicalism would have no special problem with it, in spite of there being an epistemic gap between it and the non-mental.

One problem with consciousness is that it cannot be conceived as nothing but physical matter (as it is ordinarily understood) in a certain form. It rather appears like a new kind of matter – qualifying it as radically emergent. As William James puts it:

> The point which as evolutionists we are bound to hold fast to is that all the new forms of being that make their appearance are really nothing more than results of the redistribution of the original and unchanging materials. The self-same atoms which, chaotically dispersed, made the nebula, now, jammed and temporarily caught in peculiar positions, form our brains; and the 'evolution' of the brains, if understood, would be simply the account of how the atoms came to be so caught and jammed. In this story no new natures, no factors not present at the beginning, are introduced at any later stage.
>
> But with the dawn of consciousness an entirely new nature seems to slip in, something whereof the potency was not given in the mere outward atoms of the original chaos. (James 1890/1981: 149)

If physicalists could make clear that consciousness is just physical matter in a certain form, there would be no big mystery about it, even though it still would not be intelligible why there exists a law such that physical matter ever enters that configuration, and the appearance of this configuration of matter would be accompanied by an epistemic gap between it and its causes.

The discontinuity criterion of radical emergence is also relevant for many arguments. Consciousness is often held to be the kind of thing that cannot come gradually into being; its appearance constitutes a discontinuity in nature. Dim or weak consciousness is still consciousness, not an indeterminate state in between. James also characterizes the problem of consciousness in these terms:

> The demand for continuity has, over large tracts of science, proved itself to possess true prophetic power. We ought therefore ourselves sincerely to try every possible mode of conceiving the dawn of consciousness so that it may not appear equivalent to the irruption into the universe of a new nature, non-existent until then.
>
> Merely to call the consciousness 'nascent' will not serve our turn.[footnote omitted] […]
>
> It is true that the word signifies not yet quite born, and so seems to form a sort of bridge between existence and nonentity. But that is a verbal quibble. The fact is that discontinuity comes in if a new nature comes in at all. The quantity of the latter is quite immaterial. The girl in 'Midshipman Easy' could not excuse the illegitimacy of her child by saying, 'it was a very small one.' And Consciousness, however small, is an illegitimate birth in any philosophy that starts without it, and yet professes to explain all facts by continuous evolution. […] (James 1890/1981: 151–152)

Discontinuity figures in many anti-physicalist arguments. One widespread worry is that there must be a definite point in the evolutionary continuum where consciousness shows up, but any point we select will look arbitrary (William K. Clifford (1874/1886) is regarded as the originator of this problem). There are also worries that consciousness cannot be identified with a function or computation, because it can be vague or indeterminate whether something realizes a function or computation, but not whether something is conscious (Searle 1990). All such arguments have as a premise that consciousness is an either/or-phenomenon, while nothing physical stands out from the continuum in the same way.

Finally, as already mentioned, in much of the literature in philosophy of mind, it is presupposed or argued that the coming into being of consciousness constitutes a deeper mystery than physical causation. When philosophers pose the hard problem of consciousness, the question of *how* or *why* the mental can arise from the physical, they generally do not assume that the same *how/why*-question must be asked about ordinary causation, even though an epistemic gap is present there as well. Just as we cannot, according to the knowledge argument, deduce the mental from the physical, we cannot deduce effects from their causes, and just as we can, according to the conceivability

argument, conceive of any physical configuration without phenomenal consciousness, we can also conceive of causes occurring without their usual effects. But most philosophers find that these respective gaps are still not of the same type, and they also tend to regard the problem represented by the former gap, the hard problem of consciousness, as somehow more serious than the problem represented by the latter gap, the problem of the intelligibility of ordinary causation.

As discussed, some might hold that the difference is one of mere extrinsic intelligibility, of fitting the psychophysical relation systematically in with all other causal relations, which would mean that the mystery of consciousness reduces to the problem of mental causation alone. If the problem of mental causation is solved by identifying mental properties with physical properties, one can explain the preoccupation with the mental–physical gap by the fact that identity, as opposed to causation, is a relation in which epistemic gaps do not belong. But it seems many think the coming into being of consciousness is also intrinsically mysterious, much more so than ordinary causation, and furthermore, it is regarded as mysterious in this way also (and perhaps in particular) by those who do not accept the identity theory. This could be explained by causation already being partially intelligible in a way that the relation between the mental and the physical cannot also be intelligible, as would be accounted for by the hylomorphic theory. Otherwise, it seems hard to justify that the mental–physical gap is systematically treated as intrinsically more problematic than the gap that is present in all causation.

4 COMBINATION AS CAUSATION

According to the hylomorphic theory of change, the fundamental problem of physicalism when it comes to consciousness is the following: (1) physical (prime) matter is fundamentally non-mental,[21] (2) mentality is not merely a form of non-mental matter, therefore, (3) the mental will have to radically emerge from the physical – which is at best

[21] In chapter 1 (section, 2.2, p. 36), I quoted Alter and Nagasawa as mentioning Pereboom's suggestion that prime matter could be the non-mental Russellian inscrutable, and claiming that the notion was notoriously obscure. I have now claimed that the notion of prime matter should not be so dismissed, and it may then seem that the suggestion could be the basis for a non-panpsychist Russellian monism which is nevertheless compatible with the problem from non-structural properties after all. But this would not be the case. Physical prime matter has so far been specified either schematically as "that which underlies physical change (and which is non-mental)" or "the determinable which all physical states have in common", which does not positively characterize it, or as pure potentiality, which is not distinguishable from pure dispositionality and would therefore be vulnerable to the arguments from dispositional properties (and non-structural properties).

unintelligible and at worst impossible. Assuming panpsychism and hylomorphism, on the other hand, matter itself – prime matter – is fundamentally mental. This fits Strawson's statement that: "energy is experientiality; that is its intrinsic nature" (Strawson 2006a: 234), if energy is identical to or an aspect of the prime matter of the actual universe. If macroconsciousness can be construed as a form of mentality and mentality itself as prime matter, it is possible that it is caused by microconsciousness being transformed into it. Microconsciousness would have to be regarded as a form of mentality as well, and thereby as fit for going through "generation and corruption" just as the fundamental particles of physics. Their transformation into macroconsciousness would then not constitute an intelligibility problem. It might, as mentioned above, constitute an empirical problem, but, as noted, this is set aside until the next chapter.

However, microconsciousness belongs to microsubjects and macroconsciousness to macrosubjects. How can *subjects* be reduced to a form of mentality, or a determinate of an experiential determinable? Traditionally, subjects are regarded as the most fundamental mental element. Subjects are that in which mental properties inhere, the havers of experiences and that which remains constant while experiences change. They seem more like what all other mental properties or elements are forms or determinates *of*, and thereby as though they should be in the position of matter. If so, their coming into being would constitute radical emergence. The coming into being of subjects also looks like radical emergence according to the discontinuity criterion. Subjects are not the kinds of things that seem able to come gradually into being – they either exist or do not exist.

Is this just another illustration of how panpsychism only succeeds in moving a physicalist problem and of how the combination problem is after all inescapable? I will argue that it is not, because the problem can be avoided by rejecting the view of subjectivity that gives rise to it. The view of subjects that I will propose should replace the traditional conception is the view defended by Strawson that I already discussed briefly in chapter 4 (section 2.4) as a way of avoiding some metaphysical difficulties with agent-causation. It involved construing subjects as less substantial by taking away their permanence. Now I will explain this view more fully, and show how I think it enables subjects to be finally reduced to pure form, with mentality or experientiality as their matter.

4.1 The Identity View of Subjects and Experiences – or: "Das Ich ist Unrettbar"[22]

The common view of subjects as permanent things or containers with changing experiences as properties or contents puts the subject in the position of matter or substance – as something which is not a determinate or accident of something else. The alternative view I suggest is that experience, consciousness or *what it is like*-ness is matter or basic determinable stuff, and that particular experiences are forms that experiential matter can take. Experientiality, like matter, can be thought of as a determinable that cannot exist undetermined, without any particular form. It seems there cannot be experience without any particular quality,[23] just as there could not be Cartesian matter without a particular extension, or energy without a particular form of energy.

It is not unnatural to treat experientiality as such as a general determinable. We seem to be doing it all the time. When we think that there is something that it is like to be a bat, but cannot imagine what it is like in particular, we seem to be thinking that being a bat involves the determinable experientiality, without thinking about any determinate experience. If we could not treat experientiality as a determinable, this thought and similar ones would not make sense.

If experientiality is matter and particular experiences are its fundamental forms, where does this leave subjects? Strawson has argued for the view that subjects are identical to their experiences (Strawson 2009; 2008b, and other places). If experiences are forms of experiential matter, it follows, if Strawson's view is correct, that subjects would also be forms of experiential matter.

I will now explain what happens to particular aspects of subjectivity when this view is accepted. My aim with this is only to show that the view can account for the main aspects of subjectivity which are phenomenologically apparent – as opposed to present in, e.g., intuitive judgments or presuppositions of ordinary language – and that it does so *prima facie* without incoherence. I will not argue that it should be adopted on any other basis than that it helps solve the combination problem. Many will have objections to the view, but for the purposes of a solution to the combination problem, as I have defined it, it does not need to be defended from all kinds of objections. It must only be the case that

[22] "The ego is unsalvageable" (Mach 1886: 18, footnote 12).

[23] Even in meditation, where some people claim to experience pure consciousness, this is described as a positive emptiness or openness, which does not sound like the absence of any quality but rather a very general quality.

the view is not clearly false and that it does not conflict with principles (i)–(viii) listed above. If my account of combination ends up replacing the problems of physicalism, dualism and non-panpsychist Russellian monism with different, non-analogous problems concerning the identity view of subjects and experiences, this would constitute progress (as opposed to the regress that characterizes the combination problem).

4.1.1 The Subject as the Experiencer

Subjective experiences have two aspects: one the one hand, there is the content and the phenomenal qualities; on the other hand, there is the fact that there is some point of view on these contents or qualities, an experiencer. Some think the experiencing subject is a distinct entity, which must be posited in addition to experiential content in order to yield an actual phenomenal experience. However, according to Strawson, it is not so clear that the experiencer and the content of experiences are really distinct. He holds that it is a necessary truth that there can be neither such a thing as experiential content without an experiencing subject, nor a subject of experience without an actual experience with experiential content. What ultimately makes this mutual dependency true, Strawson argues, is that the terms subject and content do not really pick out distinct portions of reality. They pick out aspects or poles of something that is fundamentally unified, namely a single experience. On reflection, we can see that subjects can be regarded as nothing over and above their experiences and contents, meaning that the relation between a subject, the contents of its experience and the experience itself could be *identity*. The distinctions only exist in our concepts and result from different ways of regarding the experiential unity. The result is that experiences really have themselves. An experience is its own subject.

This view avoids certain difficulties of the cruder bundle theory of subjects, as the bundle theory in effect eliminates the unified subject pole of the experience and leaves only the multitude of the content pole. "What bundles the bundle?" then remains the question. Strawson's identity view affirms the reality of both poles and claims that on reflection their necessary co-existence can be grounded in them being aspects of the same phenomenon. If the identity claim is coherent and phenomenologically adequate (as Strawson argues at length in his (2009)), this would seem to be the simplest and most parsimonious explanation of what underlies their co-dependence.

4.1.2 The Subject as the Unifier of a Multitude of Experiential Content

It is difficult to see how the multitude of experiential content can be straightforwardly identical to a fundamentally single subject. It is often supposed that a distinct subject is required to explain the unity of consciousness, the fact that a multitude of contents is unified into a single experience where different contents are all co-conscious. If the subject is something that *produces* a unity out of a priorly disunified collection of contents, it must indeed be distinct. In order to avoid this conclusion, one must reject the idea that there is such a thing as a *prior* multitude of contents. Rather, the total experiential field must be prior to its parts, meaning that, fundamentally speaking, we are always having *one* experience, not many. Individual qualities within the experience, such as redness, blueness, noises and moods are not priorly carved out building blocks but rather posterior delimitations of a holistic experiential field. Concepts of individual experiences are ways of carving up an intertwined whole that does not have clear joints already built in.

The result is not that individual qualities have no unity and that the experiential field can be carved up into individual qualities in any way we like. There are objective relations of similarity and difference between qualitative fields that constrain how they can be carved into contiguous sections or groups. But the precise divisions between individual qualities can be vague and depend on conceptual schemes. In some cultures, for example, there is no distinction between blue and green, so where I would count two colors, blue and green, they would count only one. Relative to my conceptual scheme, they would be wrong and I would be right, but they would be right relative to theirs, and it is hard to say that my conceptual scheme is correct – or that there is one scheme that carves the color continuum at its joints.

The boundaries between simultaneous total experiential fields (the experiences of different organisms and entities), on the other hand, seem absolute. There is only one correct way of counting the number of total experiential fields that exist right now. In this way, it seems that the total experiential field has a much stronger unity than the individual qualities contained within it.

If we accept this view, whereby the unity of a total experiential field is prior to and stronger than the unity of its parts, then the unity of consciousness does not have to be explained by appealing to the unifying power of a subject acting on a multitude of individual qualities. Rather, the unity of consciousness would just *be* the unity of the

experiential field, which can be regarded as fundamental to it, and therefore there will be no need to posit a distinct subject to explain it.

The identity of the unity of subjects and the unity of total experiential fields is also part of Strawson's view:

> The unity or singleness of the (thin) subject of the total experiential field in the living moment of experience and the unity or singleness of the total experiential field are aspects of the same thing. (Strawson 2010: 81)

He also puts it more strongly as "two aspects of the same unity" (Strawson 2010: 91), and is open to their being fully identical.

4.1.3 The Subject as the Grounder of Personal and Diachronic Identity

The view has so far eliminated subjects as distinct experiencers and unifiers of synchronically co-conscious qualities. A third aspect of subjects is their role as *diachronic* unifiers, something that makes it the case that temporally separate experiences belong to the same temporally extended person or self. If a subject is identical with an experience, it would mean that every time we have a new experience, which we do at every moment, we also have a new subject, and so it cannot account for the persistence of a self. The only thing that remains constant when an experience changes is experientiality itself, but if this is supposed to account for the idea of the persisting subject or self, it could not be unique to any human being; in fact there could only be one self in the whole universe insofar as there is only one experiential determinable, and it would be the subject of every particular experience of every person (and every thing, given panpsychism).

Strawson argues that individual experiences are temporally extended – and that subjects, being identical with them, would last as long as experiences do. But experiences cannot last anywhere near long enough to preserve the kind of permanence entailed by the traditional conception of subjects.[24]

As already discussed in chapter 4 (section 2.4), there are many empirical problems with the notion of persisting subjects. Nothing permanent is apparent to introspection, according to empiricists such as Hume, only impermanent perceptions:

[24] Strawson makes the empirical bet that experiences only last for up to two seconds in the human case (Strawson 2008b: 161). See chapter 7, section 1, for further discussion of the duration of experiences.

> For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe any thing but the perception. (Hume 1739–40/1995: para. 3/23 p. 252)

Furthermore, no physical counterpart can be found in organic bodies which is strongly and fundamentally permanent in the way subjects are on the traditional conception. That there is nothing fundamentally permanent about persons, neither physically not mentally, in which personal identity can be grounded, and that personal identity must hence be regarded as a non-fundamental property constructed out of relations of similarity and causal continuity, is a view that has been defended by Derek Parfit and has won many adherents. Further support for abandoning the permanence of everything related to the subject, except experientiality itself, can be found in some Buddhists traditions, where a central doctrine is that the persisting self when considered in one way is an illusion, but when considered in another way is universal.[25] This view is held not as a religious dogma, but as something that can be seen through arguments or upon reflecting on it in meditation, which is partly a phenomenological investigation.

I think panpsychists will have to join this group and let go of the diachronic subject or self as something fundamentally permanent. The diachronic subject or self must reduce to a sequence or stream of momentary (i.e., at most very briefly temporally extended) total experiential fields, connected by causation and similarity, in line with Parfit's psychological criteria for personal identity. In the following, I will use the term subject (and micro/macrosubject) to mean subjects that can persist for longer than an individual experience, but it will be presupposed that they are not fundamentally persisting, but rather reducible to a sequence of short-lived fundamental subjects identical to their experiences. I will sometimes use Strawson's term *thin subject* to refer to subjects in the latter, fundamentally unified and short-lived sense.

4.2 Combination as Experiential Fusion

With this account of subjects, where all their properties can be identified with properties or aspects of experientiality itself, it is possible to construe mental combination as a process where there is no radical emergence. Combination could simply be thought of as a kind of Parfitian psychological fusion, a fusion of streams of total experiential fields.

[25] In the Mahayana tradition, it is held that the true nature of the self is Buddha-nature, which is the same in all sentient beings (see O'Sullivan 2010).

Parfit imagines someone who has control over the fibers that connect our two brain hemispheres, the corpus callosum (Parfit 1971: 18–19). The brain hemispheres are disconnected, and each starts performing different parts of a long calculation, one writing it down with the left hand and the other with the right, as though they were separate persons. Each brain hemisphere will have its own stream of psychologically connected experiential moments, but the two streams will not be psychologically connected to each other – each hemisphere having no idea of what the other hemisphere is thinking. Then the hemispheres are reunited, and this results in an experience in the whole brain of remembering both calculations. A single, twice as informed, stream, will then replace the two individual streams, but the new stream will also be psychologically connected with both of the previous streams by remembering them, so the temporarily distinct subjects of each of the calculations would both survive. This seems like a case of two subjects combining into one.[26]

Combination of microsubjects could be analogous to this scenario of unification of hemispheres, where two streams of experiences fuse. In a simple case of combination, we would have two microexperiences in separate streams constituting individual microsubjects. Combination happens when two experiences at some point jointly cause a single new experience that is equally similar and equally strongly causally connected to both of them, so that they both count as "surviving" as it.[27] The new experience will, in some sense, have to be twice as rich and complex as the two previous experiences, having similarities to both of them. Let us say that each of the streams consist of experiences that are simple in the sense of having one quality. One is the experience of, say, pure red and the other is the experience of pure blue. The experience they jointly cause would be an experience of red fading into blue with some purple in between.[28]

Such a process *prima faci*e (I will examine it in more detail in section 4.4 below) fulfills all the criteria for causation devised above. Microexperiences and macroexperiences are both forms of the same kind of matter, or determinates of the same fundamental determinable. The fact that two experiences jointly cause one experience

[26] Not necessarily an empirically possible case, of course, but that should not matter.

[27] The identity of a microsubject could probably not be defined in terms of continuity of personality and memories, but we may assume that there are other kinds of similarities that would be relevant.

[28] Insofar as we can speculate about the character of microexperience at all, I do not think this is realistic. In line with the argument from causation, something like pain and pleasure, or comfort and discomfort, are better candidates.

should not be any more mysterious than two physical events jointly causing a single subsequent event. There is no discontinuity, because there are no new subjects being created, nor old ones annihilated. The microexperiences that jointly cause a new macroexperience do go out of existence, but this is not radical annihilation (which is the reverse of radical creation/emergence and just as bad) because their matter remains and is transformed into other experiences. On the view of subjects presupposed, it is only metaphysically possible for a subject to survive by being followed by a stream of experiences that is continuous with its current experience, and in this sense the original subjects do persist; they evolve into the macrosubject instead of being annihilated and replaced by it.

Coleman argues, as already quoted in chapter 1, that combination requires that all subjects involved survive, because: "If one subject is left where formerly we had two, this means at least one subject has gone out of existence, which is not combination but a fight to the death" (Coleman 2013b: 14). If combination is like Parfitian fusion, Coleman's paradox is avoided. It is a process whereby two subjects can both survive in one and the same successor subject, and in this way there can be a reduction of the number of subjects without a "fight to the death". Instead, combination can now be understood according to a comparison offered (though not further developed) by Basile: "Very roughly, if composition of some sort has to be possible, then lesser experiences do not have to come together into a larger one like bricks in a wall, but like rivers in a sea" (2010: 189).

Combination construed in this way is very different from constitution. In order for the criterion of conservation of experientiality to be respected, the microexperiences must disappear[29] (or at least significantly diminish and descend to an even deeper level of microscopicness) as they produce the macrosubject by (fully or partly) transforming into it. If a macroexperience were produced by two microexperiences alone, and the microexperiential steams continued to exist (undiminished) as distinct from the new macroexperiential stream, it would look like a new portion of experiential matter had radically emerged. Microexperiences will therefore strictly speaking exist only in the causal history of the macroexperience. In constitution, on the other hand, it is absolutely required that the microexperiential constitutive parts continue to exist simultaneously

[29] When microexperiences disappear their identical thin subjects also disappear, but their diachronic subjects do not, because their identity is grounded in a stream of continuous experiences.

with the macroexperiential whole, because things cannot be constituted by no longer existing parts.

Still, combination as causation would retain some similarity with constitution. It involves a kind of material, as opposed to material *and* formal, constitution, because the new macroexperience is constituted by the same matter as the microexperiences were, but not the same forms of the matter. For illustration, consider again how the combination of subjects on this account is more like the combination of rivers in a sea, as opposed to bricks in a wall. When rivers combine, they contribute their matter, the water they were made of, to the sea, but they lose their form.[30] There will no longer be water in the form of rivers, and therefore no rivers, when the sea begins. When bricks combine into a wall, on the other hand, they contribute both their matter and their form; there will still be cement in the form of bricks, and therefore there will still be bricks, when the wall is up.

One difference between water and experiential matter is that the forms of the latter are much more diverse. The different possible forms of water are mainly structural or geometric. The forms of experiential matter are both qualitative and structural – sensory qualities, emotional qualities, cognitive phenomenology, agentive phenomenology, etc., distributed in different ways within total experiential fields, would all be forms of experientiality. My claim is that both structural and qualitative forms are intelligibly related in virtue of continuity.

Is continuity the only relation in virtue of which quality combination is intelligible? I will now consider some other principles – holism and blending – by which it is arguably intelligible, and see whether they are compatible with my account.

4.3 Phenomenal Holism and Blending of Qualities

In chapter 1 (p. 46), I discussed Basile's argument that assuming three principles, phenomenal essentialism, phenomenal holism and sharing, combination as constitution would involve a contradiction. Phenomenal essentialism is the thesis that the essential nature of an experience is to feel a certain way, phenomenal holism is the thesis that the nature of a single identifiable experience is essentially determined by the other experiences occurring along-side it within a total experiential field, and sharing is the thesis that many subjects can share a numerically identical single experience. My account

[30] In order for the example to bring out the important contrast, it must be imagined that water is an infinitely divisible homogenous substance. Actually, river combination is just like bricks-in-a-wall combination, because H_2O molecules retain their form, just like bricks.

of combination rejects sharing, by having macrosubjects transform into macrosubjects instead of having them overlap. Thereby I avoid the contradiction, but also abandon constitution, since, as Basile says, sharing is required to make sense of this (Basile 2010: 110). Phenomenal holism and phenomenal essentialism can both be retained.

Phenomenal holism concerns what happens to individual qualities as they come together in an experiential whole. Here is Basile's explanation of the principle and its motivation:

> PHENOMENAL HOLISM – this is the view that, within a person's total psychical whole, the nature of a single identifiable experience […] is essentially determined by the other experiences occurring along- side it – synchronically – within the whole.
>
> This principle involves a rejection of the traditional atomistic view of the mind, one typically exemplified by Hume's notion of a perception as a self-subsistent, substance-like entity that could, as a matter of sheer logical possibility, exist outside of the larger field of consciousness in which it actually occurs.[reference omitted] According to James, a person's psychical field at any one moment constitutes a non-decomposable unity; its several contents are to be viewed as 'aspects' mutually determining and interpenetrating each other rather than as 'parts' in any literal sense of this word. James' commitment to this view is implicit in his remark that 'each thought [a total moment of experience] is a fresh organic unity'.[reference omitted] The denomination 'organic unity' commonly refers to a totality whose parts are internally related, such that the nature of each essentially depends upon that of all others.
>
> For the sake of illustration, imagine what it would be like to drink a cup of coffee in Naples as opposed to drinking it in Edinburgh: isn't it plausible to think that the different atmospheres of the two cities (the characteristically different colours, sounds, flavours etc. one experiences there) would make a difference to the coffee's taste? The two tastes, as they occur in the total states 'Coffee-in-Naples' and 'Coffee-in-Edinburgh', would seem to be different – qualitatively, and therefore also numerically – experiential occurrences. (Basile 2010: 107–108)

Phenomenal holism is a way in which the emergence of qualities is intelligible. It seems at least partially intelligible how a novel experience such as "coffee-in-Naples" can arise from the experiences associated with being in Naples and the experience of drinking coffee existing within one unity of consciousness.

The account I have offered so far depends essentially on the view that total experiential fields are prior to their parts (as described in section 4.1.2 above) – I will call this experiential field holism. According to experiential field holism, the existence of the phenomenal parts depends on the phenomenal whole. According to phenomenal holism,

the *mode* of existence of the phenomenal parts, their character or how they feel, depends on the phenomenal whole. It seems to me these two principles not only go very well together; experiential field holism might entail phenomenal holism.

Assuming bottom-up phenomenal constitution, where phenomenal parts are prior to phenomenal wholes, it seems clear that not only is the existence of the whole dependent on the existence of the parts; the character of the whole is also dependent on the character of the parts. For example, if one of the parts of a red field is changed to a blue, the field as a whole will no longer be red field, but a red field with a blue dot. One cannot (usually) change any part without at least very subtly changing the whole.[31] Experiential field holism reverses one of these dependencies, the existential priority, by saying that the phenomenal whole is prior to the parts. It makes sense that one thereby also reverses the other kind of dependence; that the character of the parts should now depend on the whole as well, as per phenomenal holism. This would perhaps also make more sense of what existential priority for phenomenal entities really amounts to.

Now, this is just a speculation that clearly falls short of demonstrating that experiential field holism entails phenomenal holism, but it hopefully shows that the principles can at least be easily and naturally integrated.

Coleman has argued that phenomenal qualities, as opposed to subjects, can intelligibly combine, and that we witness this when colors blend in a painting and flavors blend in a meal. Blending, as he describes it, works holistically and not by aggregation. However, he thinks blending is still a kind of constitution, a non-aggregative kind. He writes:

> In painting, one deploys qualitative elements which can, in arrangement, alter one another's intrinsic character. This mutual conditioning is part and parcel of the integration of the qualitative elements into a whole, with systemic powers all its own: the overall phenomenological upshot of all the composing elements. Qualitatively distinct elements can combine into a smoothly variegated qualitative whole. This, a painted canvas, is the model I suggest we have in mind when we think of phenomenal combination.[footnote omitted]
>
> We are already perfectly well aware of phenomenal combination, if we care to think about it. One drinks a decent red wine with the Sunday roast beef because the flavours of

[31] One might think there can be multiple realization of phenomenal wholes, in the same way that the same color shade can be mixed from different sets of other shades. But this takes careful planning. Usually, when one changes a random part of a phenomenal whole, one gets a difference in the whole. There might, corresponding to this, be some ways of carefully altering a whole that leaves certain parts unchanged – multiple realization of parts by wholes.

> roast beef and red wine pleasingly interpenetrate. Each flavour in isolation is a distinct phenomenal element. Their coming together yields a whole qualitatively distinct from either of the parts, though the combinatory upshot of their properties. This happens through the two elements fusing together, forming a genuine phenomenal unity which is the logical product of its ingredients. This is the model for a phenomenally composite, but not merely aggregative, state. [footnote omitted]
>
> […] the phenomenal ultimates mutually condition one another, as they phenomenally fuse. They form a phenomenal unity, composed of a phenomenal multitude, where the quality of the whole is the logical product of the qualities of the ingredients.[footnote omitted] (Coleman 2012: 157–158)

In my view, colors and flavors blending is really an example of partially intelligible emergence, rather than fully intelligible non-aggregative but constitutive combination. It seems one cannot derive from having experience of, say, the taste of sugar and the taste of unsweetened cocoa how chocolate will taste. Hume claimed that one could derive the missing shade of blue in an otherwise continuous spectrum, and who knows whether one really could, but it sounds less plausible that one could derive what it is like to see green if one had only experienced yellow and blue. In other words, in between qualitative ingredients and their blended result, there is an epistemic gap that does not appear fully closable in principle. Therefore, if blending is indeed intelligible, it must only be partially so, and it therefore fits better into an emergent account.

By my account, which I now take to include both experiential field holism and phenomenal holism, the correct thing to say is that the qualitative unity of a painted canvas is a result of the color patches (*qua* phenomenal, not *qua* paint on the canvas) jointly transforming into it. Qualities similar, but not identical (because of holistic influence), to the transformed colors exist within the qualitative unity we experience from looking at the painting, and because the continuity is thereby obvious, the relation between them strikes us as intelligible.

In conclusion, then, holistic blending of qualities within a total experiential field is a mode of combination which can coexist with combination as causation, because it is actually an instance of it. The fundamental principles according to which blending works and is intelligible – phenomenal holism and continuity – are either already part of the account, or can be naturally integrated in it. Blending cases are therefore good illustrations of the partial intelligibility of combination as causation.

4.4 Conservation of Experientiality and Limits to Continuity

So far, I have offered an account of combination as a process similar to Parfitian psychological fusion. This account should be examined in more detail to see whether it really satisfies the hylomorphic criteria of causation and can be partially intelligible on that basis. The criteria are qualitative conservation, quantitative conservation and continuity.

In order to determine whether experiential matter is conserved we need a notion of quantity of experientiality. The basic idea is that a rich and complex experience is constituted out of a greater quantity of experientiality, or experiential energy-matter, than a simple experience. A human experience takes more experientiality than an insect experience, which again takes more experientiality than an electron experience. For the same quantity of experientiality, we can either get many simple experiential fields or a few rich and complex experiential fields. This is in analogy with physical matter, where we can get many small (or low-energy) things or a few big (or high-energy) things out of the same quantity of matter. Hopefully this is somewhat intuitive – I am not sure how it could be more precisely specified.[32]

In three figures below, I will illustrate some different specific models of an event of combination. As it turns out, all three models will involve either discontinuity or violation of the conservation of experientiality, and they seem to exhaust the basic options. According to my arguments so far, this indicates that combination must involve radical emergence after all. But discontinuity is only a defeasible indicator of radical emergence – sometimes discontinuity has other explanations. I will offer such an alternative explanation, and if it succeeds, the discontinuous models can be seen to not involve radical emergence and will therefore be acceptable.

In the figures, one experiential field will be represented by *a circle.* The *diameter* of a circle represents the quantity of experientiality that goes into the experience. The quantity of experientiality is proportional to the degree of complexity and richness of the experience. The circles are filled with different color patterns to visualize different levels of experiential complexity and richness and also to visualize how the experiences are continuous or discontinuous with one another. The *green arrows* connect experiences that are formally continuous with each other; where there is actually a sequence of

[32] Lewtas (2012) spells out what conservation of experientiality amounts to in a similar way (but for the purposes of a constitutive account of combination).

incremental qualitative changes in between them. The *red arrows* represent discontinuity between experiences that are supposed to be connected states of the same diachronic subject (in the reducible sense). The *black arrows* represent causal influence.

In the basic example of experiential fusion given in section 4.2 above, two simple streams of different color experiences joined to form a more complex, richer stream of color experiences. This involves a discontinuity. At the moment of combination, the moment of passing from two simultaneous experiences to one, there is also a sudden passing from two simple experiences to one with "twice" the complexity and richness (figure 1).[33]

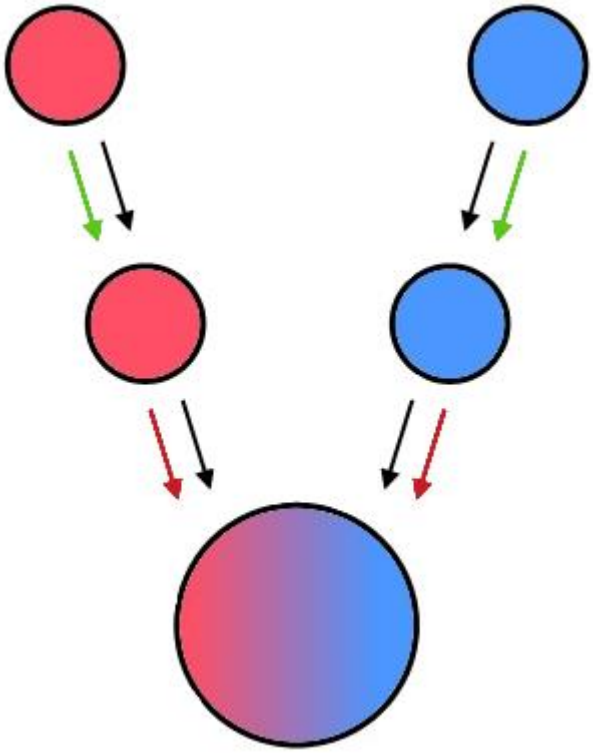

Figure 1: Symmetric and sudden combination

[33] In the example of red and blue combining into red + purple + blue, we have new qualities (purple) as well as more structure (a distribution of different colors, not just a homogenous field of one). Richness is meant to designate further qualities, not just a more complex combination of the same qualities. As I have just argued, purple cannot be derived from red and blue. Not being constituted by them, it cannot just be a matter of added complexity.

Given phenomenal holism, however, even if the combined experience were just red alongside blue with no purple and no fading into each other, this would actually be a brand new quality anyway, not just a repetition of the exact same red and blue of the simple streams. The redness of a red + blue experiential field is not identical to the redness of the pure red experience, because they are influenced by different phenomenal contexts. The rednesses will be similar but not exactly the same. The total combined experience would be one of fundamental, irreducible red-and-blue(-distributed-in-such-and-such-a-way)-ness. By phenomenal holism, complexity and richness are thus inseparable, as every time we get more complexity we also get new qualities.

In order for all qualitative transitions to occur gradually, complexity could gradually fade into both the streams. The process of combination would go as follows: experiences in two separate streams are causally interacting, causing each other to become gradually more and more similar. At a certain point they are just a step away from being qualitatively identical, and at that point they jointly cause the complex experience which is perfectly continuous with both of them (figure 2).

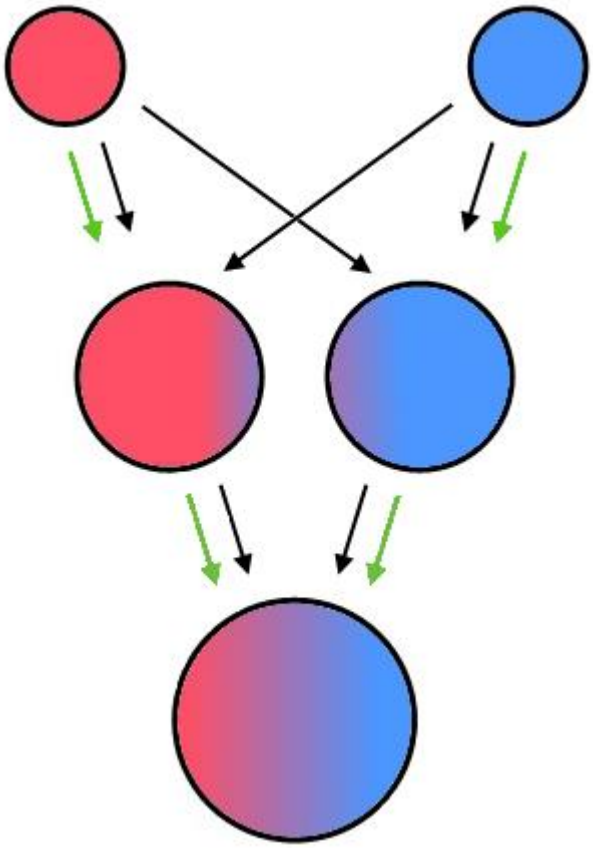

Figure 2: Symmetric and gradual combination

This model preserves continuity, but it violates conservation instead. At the moment where two almost identical experiences jointly cause one that is equally similar to both, we lose an amount of experientiality, because we replace two fairly rich experiences with one that is only slightly richer than each of them. If we replace two simpler experiences with a more complex one, it has to be "twice" as rich and complex in order for no experientiality to be lost. As the simple experiences gradually gain complexity before

combination, they will also gain experientiality from nowhere, which is another violation.[34]

The final possibility is thinking of combination as being asymmetrical. As one subject gradually gets more complex and richer, the other subject gets gradually less complex and more impoverished. The moment of combination is when the diminishing stream completely fades out, having transferred all its experientiality to the other stream (figure 3).

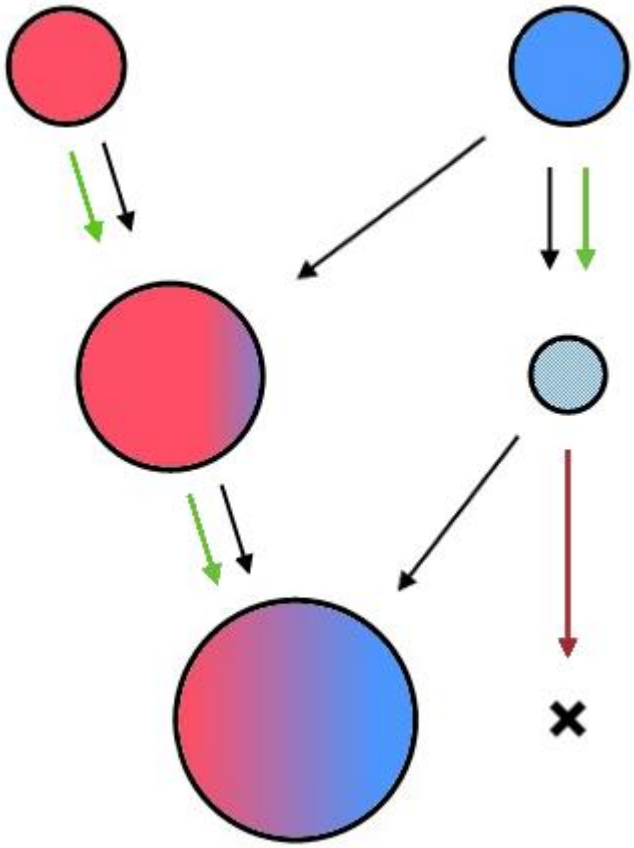

Figure 3: Asymmetric combination

In this figure, the middle experience on the right is supposed to be a dim version of the previous blue, so it is supposed to be qualitatively impoverished, but I could not make it less complex than a field of solid color.

This model gives as much discontinuity as the first model. The last experience in the stream that fades out will be a contributing cause to the combined experience, which is

[34] If the process of combination is a closed system. If we solve the problem by assuming that the experientiality lost and gained is transferred to and from other parts of a larger system, it seems we will not get continuity in these transferring processes, so it will just move the problem.

discontinuous with it. The next state of that subject is having no experience (meaning it no longer exists). But this transition, from a very impoverished and simple experience to no experience, will also be discontinuous. There is no continuity between even the dimmest experience and no experience – that is part of what makes consciousness radically emergent given physicalism, as argued above.

How can this be solved? The symmetric and gradual model in figure 2 clearly violates conservation and so must be rejected even though it gives perfect continuity. But there is hope for both the symmetric but sudden model and the asymmetric model, in figures 1 and 3 respectively, since discontinuity does not always indicate radical emergence.

A discontinuous transition does not indicate radical emergence if the discontinuous states are connected by a continuum of possible states. If an actual discontinuity could possibly be filled in with states that would make the transition continuous, it is not an indicator of a strict metaphysical binary like existence/non-existence, where there are not even possible states in between. In figures 1 and 3, every *individual* transition is possibly continuous in this way – there are possible states that could fill in the discontinuity. However, it is still not possible for all the transitions to be continuous *at the same time,* as in figure 2, where it leads to violation of conservation. When the process of combination is considered as a whole, then, there is a sense in which there are no possible states, or sets of states, that can fill in the discontinuity gap. This remains as an indication of radical emergence (in the form of radical annihilation).

I offer the following explanation: the impossibility of joint continuity has its source, not in radical emergence, but in the fact that experientiality comes in discrete quanta, i.e., in the form of fundamentally unified whole experiential fields. It is mathematically impossible to transition gradually from there being a certain *whole number* of experiences at one time to there being a different whole number of them at a subsequent time, if we have to add or take away one whole experience at a time – which we do because the idea of, say, half an experience does not make sense. In order to avoid discontinuity altogether we would have to accept an aspect of Leibnizian monadism, namely the idea that all subjects are indestructible, or that all streams of experience continue forever – they can fade to a minimum level of experiential complexity and richness, but never completely fade out. I see no reason to posit the conservation of the number of simultaneous experiential fields, as opposed to just the conservation of the quantity of experientiality. If the joint but not individual impossibility of continuity indicates only violation of

conservation of the number of simultaneous experiential fields, it would therefore be acceptable.

4.5 A Continuum of Qualities

In the figures above, colors represent experiences and it is clear how one can always fill in a transition between color combinations with possible in-between states. The color spectrum is a continuum that we can move around in without making any leaps. But are there always possible states between any experiences? The sensory modalities may appear to mark absolute distinctions between groups of qualities. How could you go gradually from, e.g., sight to smell? There are also other phenomenal groupings between which it is even harder to see that a gradual transition is possible, like sensory and cognitive phenomenology. It is clear that we cannot positively *imagine* a lot of such in-between states. If unimaginability suggests impossibility, it is even less plausible that there are possible in-between states that would allow us to pass gradually from basic electron experience to human experience, which would be even further apart.

McGinn gives a version of the combination problem formulated in terms of qualities, the problem of how we get the rich qualities of human experience from the basic qualities of ultimate particles:

> We cannot […] envisage a small number of experiential primitives yielding a rich variety of phenomenologies; we have to posit richness all the way down, more or less. An easy way to see this is to note that you cannot derive one sort of experience form another: you cannot get pains from experiences of colours, or emotions from thoughts, or thoughts form acts of will. There are a large number of phenomenal primitives. Accordingly, we cannot formulate panpsychism in terms of a small number of phenomenal primitives. (McGinn 2006: 96)

Richness all the way down is highly implausible,[35] so the worry then is that qualitative richness radically emerges. This has become known as the palette problem.

The account I have offered enables a reply to the palette problem: rich qualities are caused by basic qualities, not constituted, and therefore they need not be derivable. But in the absence of continuity between all possible qualities, it is not clear why the relation

[35] The main reason is that there is a limited number of fundamental particle types, which is much smaller than the vast number of primitive phenomenal types we would need according to the argument. This leads to a structural mismatch between the mental and the physical at the fundamental level, which is incompatible with Russellian monism and its solution to the problem of mental causation (which will be discussed in more detail shortly).

between types of qualities would not have to be classified as radical emergence; that is, it might appear that different phenomenal groupings constitute fundamentally different kinds of experiential matter, the forms of which cannot change into each other. Can the continuity of all possible states be defended, or would one have to accept discontinuity and attempt to explain why it does not indicate radical emergence?

In defense of continuity, the first thing to point out is that lack of imaginability or positive conceivability is not a good indicator of impossibility in the case of phenomenal qualities. In general, we have a very limited capacity for imagining determinate experiences that we have not had already. All the time we have experiences that we had no chance of imagining before they actually occurred.

In chapter 4, I mentioned Hartshorne's view that all possible qualities form a continuum. Charles S. Peirce also defended the same view:

> Of the continuity of intrinsic qualities of feeling we can now form but a feeble conception. The development of the human mind has practically extinguished all feelings, except a few sporadic kinds, sounds, colors, smells, warmth, etc., which now appear to be disconnected and disparate. In the case of colors, there is a tri-dimensional spread of feelings. Originally, all feelings may have been connected in the same way [...] (Pierce, cited in Hartshorne 1934: v)

Coleman (2013a) has highlighted the relevance of this view for contemporary panpsychism and neutral monism, and has pointed out how, surprisingly, there is in fact a state in between sensory modalities that we can experience. In hearing deep bass tones we are sometimes not sure whether we are hearing or feeling. This makes it seem more plausible that there could also be an experience that is between, e.g., sight and smell, even though we cannot imagine what it would be like, and our brains are perhaps not even capable of getting into a state of having such an experience.

Furthermore, there is nothing that rules out a state that is halfway in between modalities or other groupings of qualities, in the same way that there is something that rules out a state halfway in between consciousness and non-consciousness, or existence and non-existence in general. There is no metaphysical or conceptual problem with qualities in between modalities. Hence panpsychists can assume that it is just a problem of imagination, and that states that would fill in gaps in our phenomenology are possible.

In defense of the alternative view, that there is discontinuity but that this does not indicate radical emergence, it is perhaps arguable that all qualities can be recognized as

forms of the same experiential matter in spite of being divided into discontinuous groupings. Intuitively, discontinuity between phenomenal groupings does not seem as bad as the discontinuity between mental and physical properties, so maybe, like the discontinuities discussed in the previous section, it too can be explained in terms of something else than radical emergence. For example, one might think that the kinds of phenomenology we actually experience and can faintly imagine or conceive of can at least be seen to form a fine-grained spectrum. But the gaps between some kinds of phenomenology are perhaps so great and uneven that the spectrum will not be fine-grained enough to strongly indicate absence of radical emergence. If so, more must be added to the explanation of the gaps, and it is not clear what it could be. Additionally, there is a further consideration in favor of the continuum: if, as I have claimed, it could turn out that all possible physical states actually form a continuum, a mental continuum would perhaps be required anyway in order to preserve structural match between the mental and the physical, in accordance with the requirements of Russellian monism.

According to Russellian monism, the physical is the structural or relational aspect of mental intrinsic properties, and the structure of reality as revealed via the physical "from the outside" aspect must therefore match the structure as revealed from the mental "from the inside" aspect. The panpsychist solution to the problem of mental causation is premised on the relation between the mental and the physical being Russellian – if mental properties are the categorical grounds or realizers of physical dispositional structure, we can see how the physical and the mental can causally complement each other in such a way that the mental is neither redundant nor in competition with the physical.

It is now time to consider whether combination as a causal process can respect all the requirements of Russellian monism, and thereby preserve the advantage it has over dualism with respect to mental causation. This is the empirical aspect of the combination problem.

6

Combination as Causation: the Empirical Problem

Emergent panpsychism's empirical problem is its apparent conflict with at least one of these principles:

(v) The mental is not epiphenomenal.

(vi) There is no systematic overdetermination.

(vii) The physical is causally closed.

In the previous chapter, I made a number of claims about the structure of macroexperiences and the process by which they are formed. In this chapter I will show how this is compatible with these three principles.

I will begin by considering a question which is indirectly related to this, namely whether the part–whole priority structure of macroexperiences is compatible with what science tells us about the brain and its structure. I argued that macroexperiences are wholes that are prior to their parts. Brains, on the other hand, seem to be wholes whose parts are prior. I will argue that this problem can either be relatively easily solved, or be seen to more or less reduce to the problem of macromental causation, i.e., the problem of avoiding conflict with the three principles above.

The problem of macromental causation looks different depending on whether one accepts a strong or a weak reading of (vii); that is, whether one subscribes to the principle of microphysical causal closure or just physical causal closure. I will consider whether (v) and (vi) can be respected within the constraints of first a strong and then a weak reading of (vii), i.e., whether macroexperiences can be causally efficacious in a non-redundant way within microphysical and physical causal closure, respectively. I will conclude that this is possible in both cases. However, if microphysical causal closure is true, it is not possible in a very simple and elegant way, and this *prima facie* looks like a problem. In response, I will argue that if microphysical causal closure is true, then it is necessary for any theory to make a compromise between (macro-)mental causation and elegance, and that emergent panpsychism offers a better compromise than either physicalism, dualism and constitutive panpsychism.

If only physical causal closure is true, on the other hand, macroexperiences as I have construed them can be causally integrated in a very simple and elegant way. After having explained what the difference between these principles amounts to, I will argue that there is more evidence for physical causal closure than there is for microphysical causal closure. Then I will show how macroexperiences would have the same structure as the kind of emergent properties that are more likely to be found in biological organisms assuming microphysical causal closure is false, and that they are therefore perfectly suited for being the Russellian grounds of such properties.

1 OBJECTS AND SUBJECTS

Macroexperiences, according to the account I have offered, are fusions.[1] Entities that go into a fusion lose their individual identities. They exist as proper individuals (with both form and matter intact) only in the causal history of the fusion. They also *potentially* exist in the future caused by the fusion, since fusions can be undone.[2] They can, in a manner of speaking, actually exist within the fusion, if the fusion has parts or sections that resemble the entities that went into it. But these parts do not exist in same the fundamental sense that the fusion itself exists. A fusion is a whole that is existentially prior to its parts.

The brain, or any brain area that could support our consciousness, does not look like a fusion. The microphysical particles that go into the brain do not seem to lose their individual identities. They seem to exist simultaneously with the brain as constituting it, not only in its causal history. Microphysical particles are normally regarded as more fundamental than macroscopic objects like brains, and if so, the parts of the brain would be prior to the whole.

If we are correct to regard particles in the brain as prior to the wholes they constitute, and if a combined experience is, as is natural to think, the intrinsic nature of a relatively large area of the brain, one single experience will be the intrinsic nature of a big collection of different physical objects, i.e., particles. This would constitute a mismatch between the mereological structure of the physical and mental aspect of reality. What counts as an object from the mental point of view would not count as an object from the physical point of view – from there it would count as an aggregate of other objects.

[1]The kind of fusion I have proposed roughly fits the notion of fusion developed by Humphreys (1997). Seager has also argued that mental combinations are fusions, as discussed in chapter 1 (p. 49).

[2]As would be the case in the death of a macrosubject, according to panpsychism, where death would be decombination.

That the whole is prior to the parts in an experiential field is essential to avoiding the radical and brute emergence of a subject that would otherwise have to be invoked in order to account for the unity of consciousness (see chapter 5, section 4.1.2). Also, in order for non-constitutive combination to be partially intelligible, microexperiences must transform into the macroexperience and thereby lose their fundamental individuality. Therefore, in order to restore a match in mereological structure between the mental and the physical, it will have to be questioned whether the parts must really be regarded as prior to the whole in the case of macroscopic physical objects like the brain.

It is an open question in metaphysics what physical things are proper objects in their own right, and which ones are mere aggregates that do not fundamentally speaking exist. Some philosophers are mereological nihilists, claiming that the only proper objects are subatomic particles or simple substances (e.g., Sider forthcoming; Cameron 2002). Tables, chairs, brains, and other macro-objects do not fundamentally exist on this view. What fundamentally exist are just particles or simples with properties of being arranged table-wise, chair-wise, brain-wise and so on. Others are existence monists, claiming that only the whole universe fundamentally speaking exists, so that there are neither non-cosmic macro-objects nor subatomic particles (Horgan and Potrč 2000). Then there are mereological universalists, who claim that every macroscopic aggregate we can think of is a proper object in its own right (e.g., Lewis 1986: 211–213). Others again seek the more intuitive middle position, according to which some macro-objects (e.g., organisms, persons or brains only, or most things we ordinarily regard as objects, such as tables) are also proper objects while others (e.g., undetached rabbit parts, or the sum of my nose and the planet Venus) are not (e.g., Merricks 2001; van Inwagen 1990). However, all proposed criteria for objectivity that would secure a restricted middle position are very controversial.

That this discussion is going on in philosophy indicates that it is not a purely empirical question what counts as an object, fundamentally speaking, or in other words, that mereological structure is not something science unequivocally reveals. As discussed in chapter 1, science should, on the Russellian view, be taken to reveal at most spatiotemporal and causal structure, and it could be held that this does not determine mereological structure. If so, mereological structural match will perhaps not, strictly speaking, be a requirement of Russellian monism – but at the same time, it would be relatively easy to restore it. Physical mereology, not being fully constrained by science, could just be revised in order to make it fit emergent panpsychism. Emergent

panpsychism could be taken to entail the view that the particles constituting the brain area that supports our consciousness are not really objects in their own right. They *were* objects before they formed a new object, and they are still *potentially* objects in their own right, insofar as the brain will someday decompose back into separate particles. The conscious brain area, on the other hand, will *actually* be an object in its own right, simply in virtue of having a unified intrinsic mental nature. Having a unified intrinsic nature should be a respectable candidate for being the criterion of physical objectivity. Strawson (2010) argues that being a subject is a necessary and sufficient condition for being an object, on grounds independent of emergent panpsychism.

However, even if the spatiotemporal and causal structure that science reveals should not be taken to fully determine mereological structure, it is normally taken to constrain or partially determine it. To be is to have causal powers, said Plato's Eleatic Stranger.[3] Trenton Merricks has defended the claim that this is true at least as far as macroscopic objects go (2001: 81) – macroscopic proper objects should not be causally redundant. Regardless of how this criterion fares in metaphysics, in practice, at least, we tend to regard something as an object if it appears to *act* as an object. An important reason why we typically regard particles as proper objects without question, but wonder whether the brain or organism is just an aggregate (after reflection on information from science, that is, not pre-theoretically), is that it is not clear that the brain or organism acts as an object, in the sense of having causal powers that it does not possess merely in virtue of the causal powers of the particles that constitute it. In fact, it is not clear that any things outside the domain of *physics* possess causal powers that they do not possess in virtue of constituents that belong to the domain of physics. It is natural to think that fundamental laws relate fundamental objects, and insofar as the laws of physics are the only fundamental laws and they quantify over the subatomic parts of the brain, the brain is not a fundamental object.

Many think the empirical premise on which this reasoning is based is clearly true. It expresses the principle of microphysical, as opposed to merely physical, causal closure, i.e., the strong reading of principle (vii). If it is true, it leads to a strong analogue of dualism's empirical problem for emergent panpsychism: the problem of accounting for the causal relevance of the macromental to a distinct and apparently causally closed non-macromental realm. If this problem can be solved by showing that the macromental has causal efficacy, and objecthood is tied to causal efficacy, the problem of mereological

[3] At least according to some translations.

structural match will be automatically solved along with it. If macroexperiences are fundamentally causally efficacious – which they have to be, if efficacious at all, given that on my account they have no constitutive parts in virtue of which they can be derivatively or non-fundamentally efficacious – they will ground the fundamental causal efficacy of the macro-objects of which they are the intrinsic nature. These objects can then be regarded as fundamentally existing in virtue of their fundamental causal efficacy.

2 MICROPHYSICAL CAUSAL CLOSURE

The principle of physical causal closure can be formulated as follows (according to, e.g., Stoljar 2009):

> *Physical causal closure:* Every physical event (that has a cause) has a sufficient physical cause.

The parenthetical clause makes sure that the principle does not entail determinism. Indeterministic events are arguably uncaused.

Assuming that the physical can be defined without reference to physics (which I will discuss in section 3 below), the principle of physical causal closure is compatible with the strong emergence of causally relevant macrophysical properties, i.e., properties which are physical but do not belong to the domain of physics (i.e., microphysics). It allows that some macrophysical objects, objects that belong to the domain of physical special sciences such as chemistry or biology, have irreducible causal powers or are the relata of fundamental and irreducible emergent laws. The principle of microphysical causal closure, or the closure of physics, on the other hand, rules this out.

> *Microphysical causal closure:* every physical event (that has a cause) has a sufficient microphysical cause.

This principle entails that all laws or dispositions of the physical special sciences are grounded in and in principle derivable from the laws of (micro-) physics, or that all physical dispositions are grounded in and in principle derivable from microphysical dispositions.

The laws of physics can be regarded as the laws that are derivable from studying the simplest physical entities as they behave outside of complex wholes,[4] and microphysical causal powers or dispositions can be regarded as those dispositions of the simplest physical entities that are detectable from studying them as they behave outside of complex wholes. If the world obeys the principle of microphysical causal closure, then the laws of physics or the dispositions catalogued by physics, together with initial conditions, in principle predict the behavior of all complex wholes (insofar as their behavior *is* predictable in principle, i.e., is not indeterministic). It means that looking at an electron as it behaves in a particle accelerator and other relatively isolated situations should in principle be sufficient for finding out how it behaves when it is no longer isolated, like when it forms part of a brain or another macroscopic object. The behavior of macroscopic objects, like brains, may not ever actually be predictable from physics, but supporters of microphysical causal closure would regard this unpredictability as merely epistemic, a matter of macroscopic objects being so complex that we cannot see how their behavior is fully accounted for by the laws of physics, even though it really always is.

Whether empirical evidence more strongly supports microphysical or merely physical causal closure is debatable – I will discuss this in section 3.1 below. But if microphysical causal closure is true, what are the consequences for macromental causation? I will argue that microphysical causal closure does not threaten macromental causation given emergent panpsychism and combination as fusion in the same way that physical or microphysical causal closure threatens mental causation given dualism. Microphysical causal closure does not preclude non-overdetermining macromental causation; it only precludes the possibility of having non-overdetermining macromental causation in a highly elegant way. Then I will argue that the moderate inelegance of this hypothesis is acceptable in view of the advantages it offers.

2.1 Macromental Causation Within Microphysical Causal Closure

On the account I have offered, macroscopic objects like brains have a single macroexperience which supplants the individual microexperiences that belonged to the parts of the brain before combination. Let us say that the macroexperience of the brain results directly from the fusion of the microexperiences of subatomic particles, although

[4] Except wholes that are, for some reason (which I will not try to identify), regarded as microphysical even though relatively complex, like sets of entangled particles.

there could be intermediate levels of combination in between. Subatomic particles would normally, when not forming a brain, have their own individual microexperiences. As these particles end up spatially arranged and interacting as a brain, their microexperiences will jointly cause a single macroexperience which belongs to all of the particles as a whole. But if microphysical causal closure is true, this change in the experience of the particles will make no difference to their behavior. They will still go on behaving as they always did, in accordance with the laws of microphysics. According to microphysical causal closure, everything always behaves according to the laws of physics, in complex wholes that result in combination as well as outside of them. Macroscopic brains with a single macroexperience will behave in the exact same way as the subatomic particles that went into it would *if* they had not produced a macroexperience but rather went on having their own individual microexperiences. Does this mean that macroexperience is epiphenomenal?

Macroexperience is clearly not epiphenomenal or an overdeterminer in the way that a mental substance would be given dualism and either microphysical or physical causal closure. According to dualism, mental substances or properties do not replace physical substances or properties and it is the competition with the physical that gives rise to the dilemma between epiphenomenalism and overdetermination. With panpsychism and combination as fusion, there is no such competition. Macroexperience has to be doing non-overdetermining causal work because it has supplanted the microexperiences. When the brain has a macroexperience, there will no longer be any microexperiences belonging to individual particles of the brain that could be the proper categorical grounds for the dispositions of the brain, or realizers of its causal structure.

However, there is still a sense of redundancy. It appears that the causal work of the brain *could* in some sense have been done by the parts of the brain instead, *if* their experiences had not combined. If the streams of microexperiences belonging to each individual subatomic particles that went into the brain had not at some point evolved into a single stream of macroexperience belonging to a whole brain area the physical behavior of the brain would have been no different. But the sense of "could" and "would" whereby this is true is not merely counterfactual, but *counternomic*. The fundamental causal laws dictate that the simple streams *will* combine into a complex stream as the brain forms. These would admittedly be some odd causal laws; they are not very simple and elegant. They result from the odd fact that the same physical structure or behavior can be grounded in two kinds of mental natures, amounting to a kind of inverted multiple

realization or multiple categorical grounding. A particle behaving in a certain way outside a brain will be realized by a microsubject; a particle behaving in the exact same way inside a brain will be realized by a part of a macrosubject.[5]

This would mean that the structure of reality as it can be discerned from the mental aspect is not isomorphic to the structure of reality as it can be discerned from the physical aspect. Still, the structure that is discernible from the physical aspect is embeddable in the structure that is discernible from the mental aspect, so they are fully compatible, as Russellian monism requires (see chapter 1, p. 10). It would still be true that science reveals the structure of reality, it only turns out that it cannot reveal its complete structure – that science constrains but underdetermines mental structure. Only via phenomenology can we know that the brain area that supports consciousness is really in a sense a simple thing (in virtue of having a single unified experience as its intrinsic nature), as its behavior does not reveal it.

In summary, my account of combination in conjunction with microphysical causal closure entails something very far from both epiphenomenalism and overdetermination. There is no question that macroexperience is causally efficacious and non-redundant. But its efficacy is secured by an inelegant and odd set of causal laws which result from the multiple categorical grounding of physical dispositions or multiple realization of physical structure.

In chapter 1 (p. 20), I explained the sense in which physical or microphysical causal closure can be respected in spite of having non-physical metaphysical underpinnings, which should perhaps be reviewed here. It is only the empirical content of the principle of microphysical causal closure that Russellian monism needs to respect. The empirical content is here assumed to be that physical causal structure is most systematically interpreted as being closed under microphysics. The "causal" in "causal structure" in this context should be taken to refer only to certain kinds of dependency relations – relations that we can use for prediction, manipulation and control – without implying (much) about their metaphysical underpinnings, because such implications would go beyond the domain of science. If emergent panpsychism, combination as fusion and microphysical causal closure are all true, the metaphysical underpinnings of physical causal structure are

[5] The particle inside the brain which is realized by the part of the macrosubject is not *really* a particle anymore: it has lost its status as an object because it is no longer a subject (it is now rather a part of a subject). But this is not detectable from the outside – it will still look like an object. Also, note that the part of the macrosubject that realizes it is a derivative part, not a constituent.

not microscopically (or micromentally) closed, but via multiple categorical grounding or realization they still result in a physical causal structure that is most systematically interpreted as being closed under microphysics.

Emergent panpsychism, with combination as fusion, can thus respect principles (v), (vi) and the strong reading of (vii). But is the posit of multiple categorical grounding or realization, and the resulting odd laws, so inelegant that it is not worth these advantages? After all, an important part of the motivation behind principle (vi) – no systematic overdetermination – and perhaps to some extent (iv) – no epiphenomenalism – as well, is that theoretical virtues such as elegance and parsimony are to be given much weight. Therefore, it could look like emergent panpsychism and its level of inelegance is incompatible with the underlying spirit these principles. However, I think this will look different when put in a broader perspective. There are reasons to believe that no theory can secure macromental causation, in every desirable respect and in a perfectly elegant way, if microphysical causal closure is true. Therefore, the solution proposed is as elegant as it is reasonable to demand.

2.2 Mental Causation and Elegance: A Necessary Compromise

It might seem that the problem of *inelegant* mental causation does not fundamentally result from emergent panpsychism, but rather from the conjunction of two theses that every main theory of macromental (i.e., what non-panpsychists would normally just call mental) causation under microphysical causal closure assumes:

(1) The microphysical is causally closed.

(2) The macromental (human and other animal consciousness) is located in macrophysical objects.

These two theses seem to entail either that epiphenomenalism is true of at least some aspects of our minds or that macromental causation must occur in an inelegant way, no matter what theory of consciousness they are combined with. If our minds are macroscopic, and the macroscopic does not have fundamental causal powers or are not relata of fundamental laws, then our minds cannot be efficacious in the same way that fundamental microphysical particles would be. Either our minds must be non-fundamentally causally efficacious, or there must be some extra structure that secures its

fundamental efficacy, structure which is compatible with microphysical causal closure, but also makes the world more complicated than it would otherwise be.

If this is correct, then every theory of the mind that does not question these theses would have to reflect the inelegant state of affairs they prescribe in one of these ways. No theory could make these facts go away nor trivialize the situation they entail without implicitly trivializing the theses. In view of this, emergent panpsychism's inelegant way of incorporating mental causation can be seen as an appropriate way of reflecting an assumed fact about reality, rather than as a theoretical vice. It secures fundamental (macro-) mental causation, but in an inelegant way. There are other theories that are more elegant, but they cannot then secure fundamental mental causation – at least in some sense of fundamentality.

I will now argue that all the main competitors of emergent panpsychism are not only affected by the dilemma, they also come out of it worse than emergent panpsychism, all things considered. Other theories that secure fundamental macromental causation under microphysical causal closure – mainly overdeterminist dualism – do so even more inelegantly or unparsimoniously. Other theories that secure macromental causation under microphysical causal closure more elegantly – physicalism and constitutive panpsychism – cannot secure fundamental macromental causation. In chapter 1 (p. 18 and 48), I pointed out that physicalism arguably faces an empirical problem in the form of the exclusion problem, and that constitutive panpsychism might face an analogue of it. It is on this basis I will argue that these views do not secure fundamental macromental causation.

The literature on the exclusion problem is vast and highly intricate, and it would be much too ambitious to attempt to establish this while considering the whole debate. I will argue for the more limited theses that given certain assumptions about agency, assumptions that will be familiar from the previous chapters, it would follow that neither physicalism nor constitutive panpsychism can secure fundamental macromental causation, in a certain sense, and that fundamental macromental causation, in this sense, is much more important than theoretical elegance. The assumptions I will argue from include: (1) there is such a thing as phenomenology of agency, (2) the core aspects of it must be veridical in order for agency not to be illusory, and (3) these core aspects put constraints on the character of mental causation.

I will not base the argument on the further assumptions that in agency we experience something in virtue of which causation is intelligible, or something that reveals it as

essentially mental, i.e., the most controversial sub-premises of the argument from causation defended in chapters 3 and 4. Therefore, this solution to the combination problem will still not presuppose acceptance of the argument from causation.

2.2.1 Physicalism and the Exclusion Problem

According to Kim's exclusion argument, if we reject downward causation from the macrophysical to the microphysical (which, according to Kim, is possibly incoherent, and in any case incompatible with microphysical causal closure), and hold that macrophysical properties are not type-identical to microphysical properties, then macrophysical properties will either be epiphenomenal or vicious overdeterminers. Physicalists ground mental properties in macrophysical properties, and hence the same must be concluded about mental properties. Physicalist responses to the exclusion argument divide (according to Robb and Heil 2013) into three main types.

The first type of response is to argue that macrophysical properties, at least mental macrophysical properties, are after all type-identical to microphysical properties. This would avoid the problem but faces the very influential objection from multiple realization, as well as the charge that the exclusion problem will reappear at the level of aspects of properties: mental properties will be excluded *qua* mental (Robb and Heil 2013: sect. 6.5).

The second type of response is to accept overdetermination, but claim that it is innocuous. Overdetermination results if the macrophysical and the microphysical are both fundamentally efficacious. One reason for saying that there is fundamental macrophysical causation is to observe that there seem to be *epistemically* irreducible laws that hold for the special sciences, and that macrophysical properties will be subsumed by them *qua* macrophysical. It is then argued that the overdetermination that results is not problematic in the way the overdetermination that would result from substance dualism is (Robb and Heil 2013: sect. 6.3).

The third type of response is to accept something some philosophers might call epiphenomenalism about the macrophysical, but also claim that it is innocuous. It is an option for those who hold that macrophysical properties are not fundamentally efficacious in the same strong sense as fundamental particles are – e.g., because subsumption under special science laws is not sufficient, maybe because these laws are not as strict as the laws of physics, or not irreducible in principle. However, macrophysical objects can still be efficacious *via* their microphysical grounds, in such a way that causal efficacy is

somehow inherited by the macrophysical from the microphysical. If some philosophers want to label this epiphenomenalism, then so be it, one might say, because it is not the kind of epiphenomenalism that matters. Inherited or derived causal efficacy should be good enough both for the special sciences and mental causation, proponents of this view would say – there is no reason why we would need the macrophysical to be efficacious in the absolutely fundamental sense (Robb and Heil 2013: sect. 6.4).

Now, if we take phenomenology of agency seriously, difficulties appear for both of the two latter types of responses. On the basis of phenomenology of agency, it can be argued that when it comes to mental causation in particular, leaving special sciences aside, there *is* good reason to want fundamental and not merely inherited causal relevance. It can also be argued that there is good reason to want the kind of causal efficacy that *would* generate vicious overdetermination if it were to exist alongside fundamental microphysical causation.

The importance of securing mental causation does not only, or even mainly, come from the desire to avoid inelegant nomological danglers, or from an inference to the best explanation for the predictive successes of folk psychology. In large part it comes from the fact that mental causation is a necessary condition for agency. As discussed in chapter 3 (p. 108), many take it as a necessary condition for agency that our phenomenology of agency is veridical. Phenomenology of agency, if we assume that it is veridical, imposes two demands on the character of mental causation that do not seem compatible with the two latter responses to the exclusion problem just mentioned. The problem for the innocuous overdetermination response is that our phenomenology presents mental causation as irreducible productive causation, and given such a view about causation overdetermination is very hard to construe as innocuous. The problem for the inheritance response is that phenomenology presents mental causation as agent-causation, which is causation that has its immediate source in the agent, and which is incompatible with the source fundamentally being the agent's constituents.

As already discussed, much support can be found for the view that our phenomenology of agency presents mental causation as irreducibly productive, as involving power exertion or "bringing-about". Esfeld argues that: "we need a more substantial metaphysics of causation than Humean regularities or counterfactual dependence if we take our experience as agents to be veridical" (Esfeld 2007: 212). In chapter 2 (pp. 86–87), I quoted Armstrong, Ginet, Mumford and Anjum, and Bayne and Levy as expressing the view that phenomenology at least seems to represent aspects of

productive causation. Kim also claims that "[…] agency requires the productive/generative conception of causation" (Kim 2007: 14), and O'Connor seems to agree: "we seem to directly observe […] causal connectedness – […] in the (putatively agent-causal) case of my own deliberate formation of intentions, the event […] seems to be my directly exerting causal control in bringing it about" (O'Connor 1996: 11).

On a reductive view, overdetermination is not necessarily vicious, because the discernment of higher-level causes is not a matter of positing new concrete features of reality – irreducible powers or laws – we are rather just identifying additional abstract regularities or dependency relations among the concrete properties that the world (or set of all possible worlds) already contains. Esfeld notes how, "in recent years, however, a number of authors have claimed that it is reasonable to abandon (4) [the principle of no overdetermination], provided that one endorses a Humean theory of causation," (Esfeld 2010: 100) and mentions Karen Bennett (2003), Barry Loewer (2007), Ausonio Marras (2007: 318–319), Thomas Kroedel (2008) and Jens Harbecke (2008: ch. 4) among such authors.

On a non-reductive view, on the other hand, where causation is a matter of concrete production or power exertion, overdetermination is clearly more of a problem. If both a macrophysical agent, understood as an organism or an area of the brain, and the aggregate of particles constituting the agent so understood, exert sufficient power to produce or bring about my bodily movements, it seems we have twice the power exertion or energy transfer that is needed. On a non-reductive view, overdetermination involves the multiplication of concrete properties such as causal powers, not merely making new distinctions and identifying abstract dependency patterns. It generates systematic redundancy in the same way as overdetermination by pairs of mental and physical Cartesian substances would. Loewer is one who supports this conclusion: "My diagnosis of what is going on in discussions of the exclusion argument is this: if causation is understood as production then it does seem that causal exclusion is, as Kim says, 'virtually analytic'" (Loewer 2007: 253). Esfeld is another: "[…] commitment to a production view of causation simply rules systematic overdetermination out" (Esfeld 2010: 101).

The inheritance view, on the other hand, which accepts what some may call innocuous epiphenomenalism, because mental causation will be derived rather than fundamental, does not have a special problem with irreducible productive causation. The

causal powers of an agent could simply be the sum of the causal powers of the agent's microphysical realizers. This is Horgan's view:

> First, mental properties and facts are determined by, or supervenient upon, physical properties and facts. Second (and contrary to emergentism), physics is a causally complete science; the only fundamental force-generating properties are physical properties. More specifically, the human body does not instantiate any fundamental force-generating properties other than physical ones. Third, mental properties nonetheless have genuine causal/explanatory efficacy, via the physical properties that "realize" mental properties on particular occasions of instantiation. (Horgan 1996: 498)

There is reason, however, to doubt whether the possession of such an inherited set of causal powers is sufficient for agency as we experience it. It seems that we as agents experience ourselves as complex in the sense of having many motives and mental states at the same time, but also as unified: firstly, in virtue of the unity of consciousness, which I have argued should be taken as prior to the unity of its parts (chapter 5, section 4.1.2), and secondly, in virtue of exerting only one single effort at a time (chapter 4, section 2.5). This experience cannot be veridical if the powers of macrophysical objects are always constituted bottom-up. A number of philosophers have claimed that unless there is fundamental macrophysical top-down causation, where the agent (organism or brain) determines its parts, rather than *vice versa*, as in constitution, there can be no such thing as *free* agency (Searle 1984; O'Connor 2000; Merricks 2001). Steward has recently defended the more sweeping claim that this is necessary for agency of any kind:[6]

> There are those who will say that to talk of epiphenomena here [in the face of the exclusion problem] is to make a premature and rather silly mistake. My activity, they will say, is just the same thing as the activity of my parts; there is no question of my agency having been usurped. We have simply said something about what constitutes that agency and have thereby reduced rather than eliminated it. It does not follow from the fact that my actions are constituted by neural firings, etc. that I do not really raise my arm, bend my leg, and so on, any more than it follows from the fact that my washing machine's activities are constituted by various movements of the drum, switch, pump, etc. that it (the machine) does not really wash my clothes. But merely to say this is, I believe, to say much too little to answer the concerns that the reductive picture tends to engender. We do

[6] Steward does claim that a certain kind of libertarian freedom, incompatible with not only bottom-up determinism but also past-future determinism, is also necessary for agency. However, her argument for the necessity top-down causation in agency are arguments that a compatibilist with respect to past-future determinism can also accept.

> not generally think that we relate to our parts in the way that a washing machine relates to its parts. We are really in charge of at least some of our parts, we tend to feel – the ones that respond to our voluntary control – and we initiate, orchestrate, and organize their movements into the patterns that are demanded for the execution of our plans in a way to which nothing corresponds in any inanimate entity. We are happy to concede, I think, that nothing is really up to the washing machine; its activities really are the sum total of the connected activities of its parts and there is no further sense in which it – the whole machine – can make anything happen. But with us, it is different. We feel that certain things are truly up to us; that our input can genuinely settle which way things will go. It is this idea that is so difficult to square with the picture according to which our activity is simply constituted by a maelstrom of neurological processes. (Steward 2012: 227–228)

She concludes that agency requires top-down causation:

> If there is no irreducibly top-down causation, it is utterly puzzling how I am supposed to be anything other than a place where certain lower-level events produce others, how anything can ever really count as having been 'up to me'. (Steward 2012: 233)

Top-down, or downward, causation entails strong emergence and is thereby incompatible with microphysical causal closure. Steward bases the conclusion that this is required for agency on analysis and empirical studies of the concept of agency, but (as noted in chapter 3, section 3.4) it is likely that this concept in large part has its source in the phenomenology of agency.

The prospects of physicalism getting past the exclusion problem thus look much worse if it is granted that phenomenology of agency must be veridical in order for agency to be real, and if securing agency is indeed essential to a satisfactory account of mental causation.

2.2.2 Constitutive Panpsychism and the Agent-Exclusion Problem

Does constitutive panpsychism face an exclusion problem if physicalism does? One might think that it helps for the purposes of preserving agency as we experience it if the constituents that are alleged to exclude the agent are themselves mental. It is not obvious that my agency would be threatened by my individual mental states, such as my beliefs and desires, being fundamental causes of my actions. Given constitutive panpsychism, it must be granted that the efficacy of beliefs and desires is derived from the efficacy of the basic microexperiential constituents of beliefs and desires, whatever they may be, but this would arguably not make a difference if it is in principle not so that mental states causally exclude their subject. However, if phenomenology of agency is taken seriously, as well as

the agent-causal picture which is derived from it, there is evidence that mental states do actually causally exclude their subjects. As Horgan argues:

> Experiencing one's behavior as produced by oneself is fundamentally different from experiencing it as caused by internal states of oneself [...] Hence (one might well think), if the behavior is really state-caused, then it is not a piece of genuine action at all; the phenomenology of agency is illusory and non-veridical. The real source of one's behavior is not really oneself, but instead is a state of oneself (or a combination of such states). (Horgan 2007: 190–191)

If any collection of our mental states is experienced as the direct causes of our behavior, then we also experience that the agent is excluded as a cause.[7] Horgan calls it the agent-exclusion problem. The following example by Donald Davidson is a paradigmatic case of agent-exclusion:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might lead him to form the intention to loosen his hold on the rope, and this intention might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. The climber experiences the loosening of his hold as caused by his intention to loosen his hold, but he does not experience the loosening of his hold as an action. (Davidson 1973).

This example has been taken by agent-causalists (but not by Davidson[8]) to illustrate how any kind of causation by mental states excludes rather than constitutes agency: there is no agency because the agent's mental states (i.e., his intentions) directly caused the behavior and the agent himself did not. Hence, exclusion would seem to be even more of a problem for constitutive panpsychism than for physicalism. If the constituents of agents are physical, we can theoretically derive via metaphysical reasoning that there would be exclusion. If the constituents of agents are our own mental states, as per constitutive panpsychism, we can also positively experience exclusion.

---

[7] Horgan thinks physical state-causation is compatible with mental agent-causation on reflection; however, the reconciliation depends on a kind of contextualism about causation that seems to me incompatible with the commitment to non-reductionism about causation that comes with taking agent-phenomenology seriously.

[8] Davidson himself concluded only that intentions must cause behavior "in the right way". Agent-causalist have argued that "the right way" is the agent-causal way, i.e., with the agentive subject as a causal intermediary.

Emergent panpsychism has no exclusion problem. Macroexperiences do not have constitutive parts that may exclude them. Macroexperiences have individual mental states as derivative parts, which will then be only derivatively causally active. They cannot be efficacious without the total macroexperience also being efficacious.[9] Moreover, macroexperiences can clearly be productive causes, as I have already argued at length, but independently of the arguments discussed it chapters 2–4, it is an integral part of the Russellian view that experiences are the categorical grounds of dispositions, and with this an irreducible causal role is already in place. But can macroexperiences be agents? I have argued that emergent panpsychists should take macroexperiences to be identical to subjects, and an agent need not be regarded as anything more than a subject of experience with causal powers. If macroexperiences have causal powers and are identical to subjects, it follows that they are agents.[10]

Emergent panpsychism thus secures mental causation in every way arguably required to preserve agency, or basic agency, as we experience it. It should also be noted that overdeterminist dualism does the same. It posits no physical constituents of the mental agent and need not posit any mental constituents either – it is open to a dualist just as much as a panpsychist to say that the agent or the total experience is prior to its parts. On the other hand, emergent panpsychism and overdeterminist dualism are both inelegant theories compared to physicalism and constitutive panpsychism. What is more important: mental causation – in the form of agent-causation – or elegance?

Anti-physicalists often argue that it is more certain that seemingly irreducible features of consciousness such as *what it is like*-ness and subjectivity exist, than that

[9] According to the view I have defended, in the experiences of agent-exclusion it cannot be the case that we actually experience the efficacy of mental states, because on my view mental states cannot possibly be efficacious except via the agent or total experience that it is a part of. Also, according to my argument in chapter 3, we only directly experience mental causation that has its source in the agent. When experiencing mental state-causation, then, the experience must be taken as an indirect experience of causation, in which we make some instinctive or automatic judgment about causation and perhaps also have some special causal phenomenology, but without this directly revealing natural necessity. This would be analogous to the experience that comes with seeing typical mechanical interaction, such as things pushing each other, where we feel we can see the causation, but can still conceive of the causal sequence as having been otherwise. Furthermore, I would have to say that we cannot be correct if we judge the mental states themselves to be direct causes *qua* our mental states. When the climber experiences his nervousness to cause him to loosen his hold, the feeling of nervousness must be taken as a representation of, and another of the effects of, the actual cause of the loosening of his hold, perhaps some physiological state of nervousness constituted by microsubjects (i.e., microagents).

[10] At least in a minimal sense sufficient for the views I am defending. Not everyone would be satisfied with agents that fundamentally speaking only last as long as a moment of experience.

various premises in arguments for reductionism are true. Similarly, it seems we can be more certain that we are agents in the basic way we experience ourselves to be than that the world is an elegant place. If so, we should think that a metaphysical theory that is compatible with the metaphysical structure of agency as we experience it is more likely to be true than one that is not fully so compatible, but is highly elegant otherwise. There would perhaps be a point where elegance trumps agency, if the theory that secures it is highly inelegant. Many find that this is the case with overdeterminist dualism. I will now argue that emergent panpsychism is at least significantly more elegant than overdeterminist dualism, and another version of dualism that could also secure mental causation within microphysical causal closure, so it could be within the acceptable limits of inelegance even if dualism is not.

2.2.3 The Inelegance of Dualism and Emergent Panpsychism

I will take overall elegance to be affected by a number of theoretical virtues such as systematicity, parsimony and absence of *ad hoc* maneuvers. In what ways is emergent panpsychism more elegant than the kind of dualism that can secure mental causation within microphysical causal closure?

Some would perhaps argue that emergent panpsychism is less elegant because it is a virtue to be more conservative with mentality, as dualism is. But this is not clear cut. It seems qualitative parsimony is important, but it is less clear that quantitative parsimony, with respect to a fundamental kind of property that both theories must posit anyway, is important. Arguably, it is more systematic to have mentality be ubiquitous, as with panpsychism, as opposed to having mentality show up some places and not others for no clear reason, as with overdeterminist dualism.

In order to settle which is the more elegant and virtuous theory, one should rather look at how *ad hoc* the extra metaphysical structure the respective theories posit is. How much of a fortunate coincidence is it that our kinds of minds can have causal influence in a way that is undetectable from the perspective of physical science, i.e., in a way that does not disrupt microphysical causal closure?

With emergent panpsychism, a macroexperience does the same causal work, or realizes the same structure, as something which is in many respects similar to it, namely a group of microexperiences. These two causes, or realizers of causal structure, would not only be of the same fundamental ontological category, they would also possibly have qualitative similarities. Let us say that a macroexperience of co-conscious red and blue

replaces and takes over the causal grounding work of one individual red experience and one individual blue experience, after they fuse into it. It is not altogether unreasonable that very similar sets of qualities (though not identical, given the holistic influence on the qualities in the macroexperience) would systematically ground the same dispositions or enter into the same relations if all that varies is their organization into wholes (and the holistic effect that come with this).[11]

With overdeterminist dualism, on the other hand, two fundamentally different properties or substances, a physical property or substance and a mental property or substance, systematically overdetermine the same effects. There is no reason to expect substances or properties with such radically different natures to converge in effects, even though there is nothing that rules it out. The overdetermination looks completely brute. For this reason overdeterminist dualism would be significantly more inelegant than emergent panpsychism.

One might also imagine a "covert replacement" Russellian dualism,[12] where, in analogy with emergent panpsychism, mental intrinsic properties or natures undetectably replace physical intrinsic natures in brains or organisms. Then there would be no overdetermination. However, that this replacement of physical intrinsic nature by a completely different mental intrinsic nature is not detectable by a difference in the behavior they ground would be unexpected and *ad hoc* for the same reasons as with overdetermination, and therefore equally inelegant.

In summary, here is how, according to my discussion, emergent panpsychism and its main competitors compare with respect to, on the one hand, mental causation in all desirable respects – which includes agency as we experience it – and theoretical elegance,

[11] Sometimes, however, two motives conceivably have a very different disposition combined than apart. For example, say that pain on its own disposes toward repulsion, pleasure toward attraction, while pain and pleasure at the same time could conceivable dispose toward indecisiveness, which would be different from both. Given microphysical causal closure, one could make the hypothesis that combination occurs only in situations when the combined disposition happens to be compatible with the sum of individual dispositions. Pain and pleasure would accordingly only combine to indecisiveness in situations where pain and pleasure belonging to respective microsubjects would, in the absence of combination, have resulted in a static situation resulting from an even struggle between these two microsubjects. For example, a microsubject in pain wants to repel an object, the microsubject feeling pleasure wants to attract an object, and they are about to enter a situation where these dispositions even each other out so that the object goes neither way. Combination then happens – they transform into a macrosubject feeling both pain and pleasure, which would also end up neither attracting not repelling the object, because it the pain and pleasure derived from the object is equally strong so it cannot decide.

[12] Thanks to David Chalmers for the objection that dualists might try the same "tricks" as the emergent panpsychists.

on the other, insofar as they accept microphysical causal closure (it includes epiphenomenalist dualism, even though it has not been discussed):

	Mental causation in every desirable respect	**Elegance (in other respects)**
Physicalism	No	Very elegant (+1)
Constitutive panpsychism	No	Very elegant (+1)
Emergent panpsychism	Yes	Moderately inelegant (-1)
Overdeterminist dualism	Yes	Very inelegant (-2)
Epiphenomenalist dualism	No	Very elegant (+1)

Given microphysical causal closure, it is necessary to make a compromise between mental causation and theoretical elegance, and it might look like emergent panpsychism actually offers the best one – at least if agency and phenomenology of agency are taken to be important.

But should we really be so sure about the principle of microphysical causal closure? I will now consider the evidence for microphysical causal closure, and how it measures up to the evidence for mere physical causal closure. If only physical causal closure is true, and there is strong emergence within the physical, then strongly emergent macroexperiences would have a perfect structural correlate.

3 PHYSICAL CAUSAL CLOSURE

The principle of physical causal closure is compatible with the strong emergence of causally relevant macrophysical properties. The question of the status of the evidence for physical causal closure compared to microphysical causal closure is therefore also a question of the status of the evidence for strong (but not radical) emergence of properties that are not epiphenomenal. C. D Broad claimed that strong emergence is when the properties of a whole are not predictable in principle from complete knowledge of the properties and configuration of its parts – that is, the properties of the parts as they manifest in isolation or in different kinds of wholes (Broad 1925: 61). If strongly emergent properties are causally relevant, i.e., if they make a difference to the behavior of the macrophysical objects to which they belong, it entails the falsity of the principle of microphysical causal closure. Strongly emergent behavior is in principle not predictable from a complete physics, given that the domain of physics is limited to isolated particles

and the simplest wholes they form. Strongly emergent properties would belong to the domain of special sciences such as chemistry, biology or neurology.

Strong emergence is only compatible with physical causal closure if the properties that emerge are *physical*, or the special sciences they belong to are physical sciences. How is the physical to be defined? Some think the physical must be defined with reference to physics, for example, as the kind of properties physics talk about and the properties that are grounded in them. On such a definition of the physical, physical causal closure would entail microphysical causal closure. But the physical does not have to be defined in this way, and in fact, there are notorious problems with doing so, starting from Hempel's dilemma. A viable option is to define the physical negatively, as the fundamentally non-mental, non-teleological, non-intentional, non-divine, non-vitalistic, etc. (Papineau 2001: 12–13).[13] One can also add positive attributes, such as having spatiotemporal extension or a spatiotemporal region of influence.

If the physical is causally closed, and the physical is characterized as fundamentally non-mental, non-teleological and non-intentional, it does not entail that the *nature* of physical causation involves no mentality, teleology or intentionality. If this were the case, physical causal closure could not be compatible with panpsychism, and especially not the animistic panpsychism defended in the previous chapters, according to which causation has all of these properties fundamentally.[14] It is, again, only the empirical content of the principle of physical causal closure (if true) that Russellian monism of any kind has to respect. The empirical content of physical causal closure, given that the physical is understood as the fundamentally non-mental, non-teleological and non-intentional, is that the causal structure of the world is most systematically interpreted as being closed under sciences whose fundamental concepts are not mental, teleological and intentional. In other words, no scientific causal explanations reference fundamentally mental properties, and no scientific explanations are irreducibly teleological or intentional by referencing irreducible purposes or final causes – and all physical events (that have a cause) have such a scientific explanation. This does not preclude that all physical events also have

[13] Papineau suggests we take the physical as the non-mental, and that we can add more special categories that we are interested in – I proposed the other categories. Stoljar objects that defining the physical as the non-mental makes the mind-brain identity theory incoherent (Stoljar 2010: 87), because the identity of the mental and the non-mental seems like a contradiction. By adding "fundamentally" to the negative definition we avoid this consequence.

[14] However, if the physical is defined negatively in this way, then *physicalism* would of course entail the falsity of panpsychism – which is in accordance with how I defined physicalism (chapter 1, section 2).

complementary non-scientific and metaphysical explanations that do reference irreducible mentality, teleology and intentionality, as per animistic Russellian panpsychism.

What is the evidence for physical causal closure, and how does it compare to the evidence for microphysical causal closure?

3.1 The Evidence for Physical and Microphysical Causal Closure

Physical causal closure is mainly supported by empirical evidence, or principles implicit in successful scientific methodology.[15] The principle of conservation of energy rules out all causation that is not conservative, such as free and spontaneous causation (Papineau 2001: 25), and therefore counts as evidence against some forms of causal influence that would be non-physical under the negative definition, e.g., mental forces *qua* spontaneous and creative, as they have traditionally been conceived (Papineau 2001: 19). Furthermore, science has not found anything going on in bodies and brains, places where non-physical influence would be expected to manifest, that requires explanation in terms of irreducible teleology, spontaneity, and so on (Papineau 2001: 30–32). The functioning of organisms and brains looks like it can be fully accounted for in terms of the mechanistic (i.e., lawful and non-teleological) interactions of their parts. From the empirical standpoint, then, there is no reason for, and good reason against, positing non-physical (by the negative definition) causation.

Are the empirical considerations in favor of microphysical causal closure equally strong? It appears not. The main general reason against positing some kinds of non-physical causation, the principle of conservation of energy, has no equivalent when it comes to non-microphysical causation. As discussed in chapter 5 (p. 157), the principle of conservation of energy plays a strong regulative role in science – whenever there is an apparent violation of the principle, scientists will set out to find the hidden source or destination of the energy that appears to either just show up or go missing. Proceeding in this way has been highly conducive to success.

It is not clear that the ideal of microphysical ontological reductionism plays a similar regulative role. Kant argued that the unity of science was an essential regulative idea. However, philosophers of science such as Nancy Cartwright and John Dupré have argued

[15] There is also some empirical reason to think that physical causal closure is false, mainly that there is an interpretation of quantum mechanics which fits the observational data according to which consciousness collapses the wave function (Chalmers 2003: 125–127). But until this has been developed further and, if possible, tested, it seems the evidence for physical causal closure is stronger so far.

that scientific practice does not in fact proceed on the assumption that everything can in principle be reduced to levels below, and moreover, that it is very undesirable that they start doing so. Cartwright claims that "[...] belief in the unity of the world and the completeness of theory can lead to bad methodology [...] (1999: 13). She argues that the most successful science, e.g., applied physics such as engineering of lasers, actually proceeds as though phenomenological laws – higher level descriptive laws – are the true laws, while fundamental laws are treated as just one tool among many:

> Fundamental laws are supposed by many to determine what phenomenological laws are true. If the primary argument for this view is the practical explanatory success of the fundamental laws, the conclusion should be just the reverse. We have a very large number of phenomenological laws in all areas of applied physics and engineering that give highly accurate, detailed descriptions of what happens in realistic situations. In an explanatory treatment these are derived from fundamental laws only by a long series of approximations and emendations. Almost always the emendations improve on the dictates of the fundamental law; and even where the fundamental laws are kept in their original form, the steps of the derivation are frequently not dictated by the facts. This makes serious trouble for […] the view that fundamental laws are better. When it comes to describing the real world, phenomenological laws win out. (Cartwright 1983: 127)

Dupré claims, similarly, that:

> [...] the disunity of science is not merely an unfortunate consequence of our limited computational or other cognitive capacities, but rather reflects accurately the underlying ontological complexity of the world, the disorder of things. (Dupré 1993: 7)

He argues that in biology reductionism is often an inappropriate strategy for understanding the determinants of evolutionary change (Dupré 1993: 139), and that the reductionist strategy of population genetics, in particular, has led to few, if any, theoretical and practical benefits (Dupré 1993: 141). If Cartwright and Dupré are correct, we cannot say that successful scientific investigations in general proceed *as though* everything is in principle reducible to microphysics.

Is there a microphysical counterpart to the second kind of evidence for physical causal closure, the evidence from the actual explicability of brains and organisms in (negatively defined) physical terms? Answering this requires considering in more detail what a violation of microphysical, but not physical, causal closure would amount to. Some think it would amount to a kind of downward causation which is logically or metaphysically incoherent. As Kim has objected (1999), if macrophysical wholes

somehow exert influence on the microphysical parts that constitute them, it might seem this would result in paradoxical self-causation. If so, i.e., if physical but not microphysical causal closure is paradoxical, then all evidence for physical causal closure must be evidence for microphysical causal closure, which would be the only tenable version of the principle.

Irreducible macrophysical causation, which would be a form of strong emergence, can be conceptualized in various ways. Here is a simple way in which it might occur that involves no paradoxical causal loops. One might think that fundamental entities have certain dispositions that remain latent, i.e., are not triggered to manifest, in conditions that can occur in physics laboratories and other places where particles are relatively isolated. Rather, these latent dispositions are triggered by membership in certain complex wholes. For instance, when a particle ends up in certain close interactions with several other particles, like when forming a cell, this will trigger the latent disposition, and this will manifest in behavior that is different from the behavior the particle exhibited as part of simpler wholes or when isolated. Shoemaker (2002) explicates emergence in terms of latent dispositions.[16]

If emergence is the manifestation of otherwise latent dispositions of fundamental particles, macroscopic wholes that are loci of emergence would still clearly be causally relevant. They are the contexts that trigger the latent dispositions, and if they were absent, different behavior would manifest. There would be no way of avoiding reference to these wholes in explanations of emergent behavior or when formulating the laws that govern it.

Even if irreducible macrophysical causation is metaphysically coherent, another consideration against it is that it might seem to entail that macrophysical wholes have the power to violate the laws of physics, which science clearly tells us is not possible. Would the laws of physics be violated when latent microdispositions are triggered?

If there is strong emergence of this kind, it does not entail that the laws of physics are not true. As Cartwright says: "to grant that a law is true [...] is far from admitting that it is universal, that it holds everywhere and governs in all domains" (Cartwright 1994: 281). As already suggested, the laws of physics could be taken to inform us about how particles

[16] McLaughlin (1992) also argues that what he takes to be the mistake of the British emergentists was empirical, not philosophical, i.e., that the concept of strong emergence contains no contradiction. O'Connor (1994) argues that conceiving of emergence in terms of latent dispositions is *ad hoc*, and not necessary to make downward causation coherent in any case, so fundamental emergent dispositions can instead be located directly in macroscopic wholes themselves.

will behave when they are relatively isolated or in simple contexts, without also being taken to tell us, at least not *exactly,* how they will behave in more complex contexts. There could be some laws of physics, which, like the laws of conservation of energy, would hold in any context, but for such laws, there are many ways in which particles can behave within their constraints. There would then be room for irreducible macrophysical laws (grounded in latent dispositions) to determine particle behavior precisely within these constraints. Therefore, we need not think of the laws of physics as being ever violated by strong emergence simply because there could be much about which they remain silent.

We can now return to the question of whether there is a microphysical counterpart to the second kind of evidence for physical causal closure. What is the evidence that brains and organisms are devoid of not only non-physical causation, but also irreducible macrophysical causation of the kind coherently explicable in terms of latent dispositions? While what is going on in brains and organisms is to a great extent explicable in terms of the mechanistic behavior of their parts, it is not currently to any similar extent explicable in terms of microphysics. That we are able to account for the behavior of brains and organisms in terms of the regular, conservative and non-teleological behavior of their parts is evidence for physical causal closure only. In order to support microphysical closure, we also have to establish that the parts follow the same patterns of mechanistic interaction as we find in the very simple contexts that constitute the domain of physics, i.e., that their behavior can be fully accounted for by the laws of physics, and this has not been shown. That is not to say that there are no reasons to think this will, or could in principle, be shown. But considering only what has actually been shown, this seems to support physical causal closure much more strongly than microphysical causal closure.

Typically, proponents of microphysical causal closure grant that we have not succeeded, and most certainly never will succeed, in epistemologically reducing all special sciences to physics. Carl Hoefer explains why he still favors microphysical ontological reductionism: "Why, then, do I think that physics today gives strong evidence for the existence of true universal and fundamental laws? [...] We have already found such mathematical regularities that are true or very close to true wherever we are able to check" (Hoefer 2008: 309) He offers an example of how the structure of the hydrogen atom falls out "beautifully and exactly" from the Schrödinger equation. "Where we are clever enough to be *able* to test this theory [quantum mechanics] and this equation [the Schrödinger equation], they seem to be correct." (Hoefer 2008: 317–318). This is

arguably a good inductive basis for concluding that everything is in principle reducible to physics, and that it is only the complexity of the higher levels that hinders us from seeing it.

Arguably, though, these cases are too few, or taken from too limited a domain, to constitute a solid inductive case. The clearest cases of derivability are at the atomic or molecular level – so maybe they just show that there is no strong emergence at the atomic or molecular level? If strong emergence is correlated with macroexperience, then the only place we should definitely expect to find strong emergence (because we know that there is macroexperience there) is in organisms. Here we have no such clear cases of actual epistemic reducibility.

The fact that the special sciences are to a large extent epistemically irreducible cannot be taken as a clear indication in the direction of either microphysical causal closure or strong emergentism. The case for microphysical causal closure gains plausibility for every new reduction or derivation that is made. But the strong emergentist need only point to one case where we have a complete overview of the microphysical facts, and where we can see that they do not entail the behavior of the whole, in order to falsify microphysical causal closure. Boogerd et al. have presented evidence for what could be such a case. They derive a notion of emergence, which I have already mentioned, from Broad and Ernst Mayr, who: "both think of systemic properties as emergent if they cannot be deduced, even in principle, from the behavior the system's components show within simpler wholes" (Boogerd et al. 2005). They proceed to present evidence for it: "We show in an explicit case study drawn from molecular cell physiology that biochemical networks display this kind of emergence […]" (Boogerd et al. 2005: 131).

The case they present is a system, a metabolic pathway, here referred to as A, which has an enzyme subsystem, here referred to as A_1, as a component:

> The behavior of A_1 in isolation is sometimes qualitatively different from the behavior of A_1 in A, and therefore, since the behavior of A is a function of A_1, understood as a component, the behavior of A cannot generally be derived from studies on simpler subsystems of A. In general, the (dynamic) behavior of A is not simply the superposition of the (dynamic) behaviors of its subsystems studied in isolation. Dynamic interactions can bring about qualitatively new behavior in complex systems. This is precisely where prediction of system behavior on the basis of simpler subsystems fails. We cannot predict the behavior of the components within the entire system and so cannot predict systemic behavior. This is emergence, with novel system behavior that cannot be predicted on the basis of the behavior of simpler subsystems. (Boogerd et al. 2005: 165)

If Boogerd et al. are correct, then the microphysical would not be causally closed; a Laplacian demon with complete knowledge of the laws of microphysics, as well as the initial and boundary conditions of the metabolic pathway, would not be able to predict its exact behavior. But whether Boogerd et al. are in fact correct in their account of the case is probably a matter of controversy,[17] or so I would assume, given the wide acceptance of microphysical causal closure among philosophers.[18]

In summary, speaking from the empirical point of view, it should be conceded that it is at least to some degree an open question whether the principle of microphysical causal closure holds. There is much room for disagreement about the direction in which the evidence most strongly points, and about what should count as evidence. But it seems clear that microphysical causal closure is not as strongly empirically supported as mere physical causal closure.

3.2 Strong Emergence as the Correlate of Combination

The possibility of the world being only physically, but not microphysically, closed fits very well with emergent panpsychism. Strongly emergent dispositions or behavior would be grounded in strongly emergent combined macroexperience. New dispositions would emerge on the basis of the emergence of new intrinsic natures, macroexperiences, that replace the microexperiences that grounded the earlier dispositions. Constitutive panpsychism could not easily explain emergence. On this view, strong emergence would seem to entail a change in dispositions without a change in categorical, intrinsic

[17] An objection to Boogerd et al.'s case, given by Chalmers (in conversation), is that while the system's properties are perhaps not predictable from biochemistry or chemistry, the levels immediately below, it might still be predictable from the very bottom level of microphysics. To rule this out, one would need a complete microphysical specification of the system, which is not something we can obtain.

Dupré's view (as expressed in conversation) is that the matter is not empirically decidable in principle: reductionists can always claim that emergentists have missed microphysical details that would predict the seemingly emergent behavior, while non-reductionists can always claim that reductionists have insufficient basis for their inductive case, since they can never actually catalogue all microphysical details and show that everything else follows.

[18] While many philosophers explicitly commit to physical causal closure, they tend to assume either that the physical is to be defined relative to physics, or that the evidence for physical causal closure is also evidence for microphysical causal closure. Chalmers defines materialism about consciousness as "the thesis that consciousness is physical: that is, that truths about consciousness are grounded in the fundamental truths of a completed physics" (Chalmers forthcoming-b). Philip Pettit (1994) argues that the physical should be defined relative to physics. Brian McLaughlin claims that there "seem not a scintilla of evidence that there are emergent causal powers or laws […]" (McLaughlin 1992: 55). In the introduction to an anthology on *System Modelling in Cellular Biology*, it is claimed (without any supporting arguments) that strong emergentism "is completely antithetical to materialism and remains as yet on the fringes of scientific thought" (Szallasi, Stelling, and Periwal 2006: 8–9).

properties. This would result in there being ungrounded dispositions, which is not compatible with Russellian monism.

If we accept that having irreducible causal powers is a criterion of objecthood, then the mereological problem, discussed in section 1 above, should be solved along with finding a causal correlate of combination. But if an account of emergence in terms of latent dispositions is accepted, it might seem to persist. In experiential fusion, a macroexperience becomes the direct source of causal exertion and a direct bearer of dispositions, and microsubjects only survive in a non-fundamental sense. When particles form a whole that triggers their latent dispositions, on the other hand, they clearly survive in the sense of remaining bearers of dispositions. However, I think the mismatch is only superficial. The sense in which microsubjects can survive fusion and retain dispositions is really no weaker than the sense in which particles survive when they form a disposition-altering whole, given some assumptions about the identity conditions of particles that would follow from the Russellian view of the physical.

When microsubjects, streams of momentary microexperiences, fuse into a macrosubject, or one stream of macroexperiences, the microsubjects survive according to criteria analogous to Parfit's psychological criteria for personal identity – criteria of non-exact similarity and causal connectedness. If one of the microsubjects consists of a stream of pure redness, which subsequently contributes to a macroexperiences of red + purple + blue, the redness can survive within the macroexperience, but not the exact same redness, given the holistic influence of the phenomenal whole on the phenomenal parts.

When particles, streams of momentary microphysical states or time-slices, fuse into an emergent macroscopic whole, or a stream of macrophysical states or time-slices, the particles also survive only according to criteria of non-exact similarity and causal connectedness. Like microsubjects, they do not survive according to criteria of exact similarity, because the behavior of the particles has been altered by the influence of the whole, by the whole triggering the manifestation of previously latent and unmanifested dispositions. On some metaphysical views, particles have a non-dispositional, quidditistic essence which would remain exactly the same and preserve their individuality in a strong sense through an altering of dispositions – which would result in a mereological mismatch with microsubjects which, after combination as fusion, do not survive in virtue of preserving any quidditistic essence. But according to the Russellian view, the physical is dispositional all the way down, and the identity conditions of physical objects such as particles must be based similarity of dispositions and the relations that result from them.

Therefore, particles can only survive disposition-altering fusion in the same weak sense that microsubjects survive fusion, namely via retaining imperfect similarity and causal connectedness.

As agents, macrosubjects exert unified powers or efforts. But these efforts are exerted in view of a multitude of motives that exist within the unity of the macrosubject. Motives *qua* experiences are also holistically influenced – pain alone motives avoidance, while pain together with other motives can motivate other efforts. The way in which all the motives act as a whole – unified in one experiential field – can be different from how they would act as an aggregate, if they were contained in separate experiential fields. In the same way, the way in which particles act as an emergent whole is different from how they would act as an aggregate. As an aggregate, they would act according to the laws of microphysics; as a whole, they act according to emergent laws. It seems the individual motives experienced by an agent correspond well to individual particles of an emergent whole – both have their own dispositionality, but the way their dispositionality is manifested depends on the whole they form part of.

Many panpsychists have already suggested emergent phenomena as correlates of combination. However, they have focused on emergence of more extreme kinds, where we have not only the emergence of new behavior belonging to what can still be regarded as the same particles, but rather complete transformations of matter into an entirely new state – a form, quality or structure which has very little resemblance with the forms, qualities or structures of antecedent states. For example, Keith Turausky (2012) has suggested Bose-Einstein condensates as the type of correlate we should be looking for. This is a so-called macroquantum phenomenon, where matter enters a completely novel state.

As mentioned in chapter 1 (p. 49), Seager (2010) suggests that combination is a matter of so-called combinatorial infusion. Combinatorial infusions are "large simples," which are "partless yet extended" (Seager 2010: 180). This is very similar to the way in which I have suggested that combination works. However, one important difference is that I have allowed that fusions or combinations have parts, as long as the whole is prior to them, in virtue of having a stronger unity than they have, and in virtue of holistically influencing them.

As Seager points out, there is "little evidence that the brain supports any processes that could count as combinatorial infusion at the physical level" (Seager 2010: 181–182). On the conception of combination where the result is not a partless whole but merely a

whole that is prior to its parts, it is more likely that we can find a directly corresponding structure in places like the brain. We do not have to look for very distinctive states of matter like Bose-Einstein condensates, quantum coherence or, as per Seager's suggestion, something analogous to classical black holes.[19] We only need the behavior or dispositions manifested by particles to slightly change according to what kind of whole they are parts of, which according to Boogerd et al. there is already evidence for at the biological level.

When we consider our own minds phenomenologically, they do not appear to have exactly the structure of a large simple; of being partless. Our minds clearly have parts in the form of individual feelings, thoughts and sensations. The unity of consciousness that encompasses all this variety has a much stronger and exact kind of unity than any of elements within it, but these lesser unities, individual qualities, still distinctly exist. Hence the conception of combined experience as a whole that is prior to its parts, as opposed to partless, seems both more phenomenologically adequate, and more likely to have physical correlate.

3.3 Strong Emergence and Dualism

If the microphysical is not causally closed, how does that fit with dualism? The principle of conservation of energy and the mechanistic explicability of organisms and the brain do not rule out the influence of mental substances or properties that act conservatively and non-irreducibly teleologically. It might seem that if the laws of physics leave the behavior of certain wholes undetermined, mental substances or properties can find a causal role in being what fully determines these wholes within the constraints of physics. Two problems remain for the dualist, however.

The first is empirical: the evidence we according to Boogerd et al. have for emergence is found already at the level of cells, and traditionally, dualists do not see mentality as extending across the entire biological and biochemical realm, but limit it to humans and (some) animals. This is not essential to dualism, of course, but if dualists bite this bullet, they no longer have the advantage of being able to respect common sense judgment about how far mentality extends.

[19] Seager claims that classically modeled black holes would have the structure of a combinatorial infusion, but he does not suggest them as actual correlates of mental combinatorial infusions. Rather, he argues, as discussed in chapter 1, that the existence of such models shows that the notion of a combinatorial infusion is intelligible.

The second problem is one of insufficient motivation. I have already argued that if dualism tries to mimic emergent panpsychism's account of mental causation within microphysical causal closure, it comes across as much more *ad hoc* than emergent panpsychism. This will also be the case if dualism tries to mimic emergent panpsychism's account of macromental causation within mere physical causal closure.

If there is emergent behavior on higher levels, this would not constitute any radical difference between the level of physics and the higher levels. Things at all levels behave according to the same basic principles of efficient causation, so the main difference between things with emergent dispositions and things with dispositions as described by physics, is their size and complexity, which are both matters of degree. If there were something like irreducible teleology to be seen in the behavior of things at higher levels, or some other novel characteristic rendering it discontinuous with lower level behavior, it would make sense to posit radically discontinuous mental properties or substances to explain such radically new behavioral principles. But why should an entirely new nature have to be introduced in order to explain the fact that the same kinds of things show the same kind of law-governed, efficiently caused behavior, the only difference being the exact pattern of the behavior?

In other words, dualism's problem is that emergent physical behavior is not radically discontinuous with microphysical behavior, but mental properties or substances dualistically conceived would be radically discontinuous with the physical. Emergence within a world closed under mechanistic, conservative principles for behavior is not special enough to warrant such a special explanation. It could be posited only on an *ad hoc* basis, for the sake of dualism itself.

One could also argue that given the similarities between higher and lower levels, if macrophysical behavior needs to be determined by a mental substance or property, then there is no reason why microphysical behavior would not need it too. The laws of physics may leave certain things open, but the laws of nature do not, and mental substances or properties must act in accordance with the laws of nature. If the laws of macrophysics need a mental explanation or underpinning, then the laws of microphysics should need one as well – both being brute patterns. That both equally need mental explanation or underpinning, in the form of Russellian grounding, is exactly what panpsychists think, and this seems to be the less *ad hoc* conclusion.

4 COMBINATION AS CAUSATION: SUMMARY

The two last chapters started out from a dialectical situation according to which panpsychism appears to have an advantage over competing views, physicalism, dualism and non-panpsychist Russellian monism, in avoiding their respective main problems. Panpsychism appears to avoid having to posit brute emergence, *a posteriori* identity, or other problematic relations between our mentality and its base, and to avoid epiphenomenalism and systematic overdetermination within the constraints of physical causal closure, all without appealing to unknown natures, properties or relations. It is up to defenders of competing views to either show that their respective problems can be solved, to show that panpsychism is an untenable position in and of itself, or to show that panpsychism, while tenable in and of itself, does not really solve the problems either, but only has them reappear in another guise. The combination problem has a good chance of achieving the latter result. It challenges panpsychists to close the epistemic gap between fundamental mentality and our own minds or, if this cannot be done, explain why our minds are not thereby to be regarded as inexplicable emergents relative to fundamental mentality. It also challenges the panpsychist to explain how our minds are not causally redundant once fundamental micromental causation is ubiquitous, and to do all this without appeal to unknown natures.

I have argued that even if there could be no fully intelligible constitutive relation between micromentality and our macromentality, panpsychists are not forced into accepting that macromentality is radically emergent and not intelligible at all, because these are not the only alternatives: it is open to the panpsychist to claim that macromentality is caused by microexperience, and that there is a sense in which the causal relation is partially intelligible. Then, I showed the ways in which macroexperiences could not only be caused but also themselves be causally efficacious, without violating the principles of microphysical and physical causal closure respectively. In this way, panpsychism – of the emergent type – avoids the problems of competing views after all. If this is correct, those who oppose panpsychism must find another argumentative strategy against it than posing the combination problem.

My general argument for this had two main parts, one concerning the intelligibility problem emergent panpsychism seems to inherit from emergentism, and the other concerning the empirical problem which it seems to inherit from dualism. My solution to the intelligibility problem began by proposing a distinction between causation and radical emergence. I argued that causation is partially intelligible in virtue of relating

determinates of the same fundamental determinable, or forms of the same matter, in ways that respect principles of conservation and continuity. I argued that radical emergence is unintelligible in virtue of not abiding by any such principles. From this, it could be concluded that the relation between our kind of mentality and its base could in fact be causal and partially intelligible given panpsychism, but could only be radically emergent and not intelligible at all given physicalism – only within panpsychism can our kind of mentality be a resultant of something which is a determinate of the same fundamental determinable, namely experientiality itself. That consciousness is not a form of non-mental matter, and thereby cannot be intelligibly produced by a fundamentally non-mental base or antecedent, is the fundamental problem that some well-known anti-physicalist arguments can be interpreted as pointing to, as opposed to the presence of an epistemic gap that is not closable in principle.

The solution to the intelligibility problem further consisted in showing how macrosubjects could be caused by microsubjects while respecting the criteria of conservation and continuity. *Prima facie,* the relation between different individual subjects seems not to qualify as being free of radical emergence, firstly, because a macrosubject is clearly not just a configuration of microsubjects, i.e., not a form of them, and secondly, because the coming into being of a new subject appears to involve discontinuity characteristic of radical emergence, since a subject seems to either exist or not exist, and cannot come into being gradually. I argued that the solution to this is to revise our view of what it is to be a subject of experience, along the lines proposed by Strawson and Parfit. We should think of experientiality as fundamental, as a kind of matter, stuff or determinable, and total experiential fields as the fundamental forms this matter can take. Subjectivity is then integrated into experientiality by seeing that both the subject of experience and the content of experience are inseparable aspects of experientiality itself. The unity of consciousness is identified with the unity of a total experiential field, where the whole is prior to its parts. The subject as something that is supposed to persist through time is reduced to a series of momentary total experiential fields connected by similarity and causation.

On this background, we can think of mental combination as a kind of Parfitian psychological fusion, where a group of microexperiences jointly cause a single macroexperience, which is equally similar and equally strongly causally connected to all of them, so that they all count as having survived in it. Such a process would involve no discontinuity of the kind that indicates radical emergence and no violation of

conservation. It would also be manifested in the *prima facie* intelligible phenomenon of holistic blending of phenomenal qualities. The conclusion to the first part of the argument is that mental combination can be construed in such a way that it is just as intelligible as any causal process – at least just as *intrinsically* intelligible, i.e., intelligible insofar it is not taken into account how the causal process as described could fit into the causal structure of reality as a whole.

The second part of the argument consisted in showing how the causal structure of combination as I construed it is compatible with the causal structure of physical reality as science reveals it, in order to secure the causal relevance of macroexperience. I first considered an indirectly related problem of mismatch between the part–whole priority structures of brains and macroexperiences: brains seem to be wholes whose parts are prior to them, while macroexperiences, on my account, are wholes that are prior to their parts. In order to restore structural match, it could be proposed that the brain can in fact be regarded as being prior to its parts, so that the particles within it are not actually objects in their own right, only potentially so – because the idea that all and only subjects are objects is, as Strawson has argued, a coherent solution to the open philosophical question of what is the true principle of mereological composition. However, if the criterion of objecthood is taken to involve causal efficacy, one cannot revise mereology as one likes. The problem of mismatch of part–whole priority structure would transform into the problem of macromental causation.

If the microphysical is causally closed, as many philosophers think, it might seem that the macromental must be causally redundant. However, I showed that when we look more closely, emergent panpsychism with combination as fusion actually gives us mental causation in every way that we could want, only at the price of the inelegant posit that some physical dispositions can be multiply categorically grounded or that some physical structure can be multiply mentally realized. Other theories, physicalism and constitutive panpsychism, are more elegant, but they cannot give us (macro-) mental causation in every desirable respect, given that these respects include agency as we experience it. Overdeterminist and Russellian dualism can also secure mental causation in all respects, but even less elegantly. A different solution to the problem of macromental causation would be possible if it turned out that causal closure holds merely for the physical and not for the microphysical – which is a possibility I showed that there is good reason to take seriously. If there is strong emergence within the physical, then mental combination as I construed it would be its perfect structural correlate. The conclusion to the second part of

the argument is that panpsychism with combination as (emergent[20]) causation gives us mental causation in the same robust way that interactionist and overdeterminist dualism would, but in a more elegant way.

The resulting picture might look complicated, but in a way, it is simple. Our experiences are just as fundamental as any other kind of experience. They cannot be derived from, nor be reduced to, anything else. Any total experience is a basic component of reality, and the relations they stand in are primarily causal and non-deductive. However, in a sense there is something that is more fundamental than any particular form of experience, namely experientiality itself. This is what all experiences have in common, and what conditions their existence. Thus, we end up with a kind of constitutive explanation of consciousness as we know it after all. Just as we ordinarily think the existence of a certain quantity of physical energy-matter or stuff explains how it is that particular physical things can exist, the fact that this energy-matter is fundamentally experiential explains how it is that our experiences can exist.

When leaving constitutive panpsychism behind, it may seem as though we also leave behind an orderly picture of the structure of the world, where order is constituted by hierarchy. However, as the combination problem appears to show, that is an order into which our experiences refuse to fit, even as we go so far as to allow that the hierarchy is mental throughout. The combination problem appears to show that the idea that the mental is fundamental is must be taken all the way in order to achieve its intended result. Emergent panpsychism abandons all reductionism and constitutive hierarchy, but we are still left with a conservative and continuous causal order. It is a picture where our experiences are fundamental, but not exceptional, and integrated in the causal order of nature, but not subordinated – so it should also be an order we can live with as causal agents.

[20] I defined the notion of emergent causation in chapter 5, section 2.2, as a relation which is partially intelligible, and therefore different from radical emergence, but less extrinsically intelligible, and therefore different from ordinary causation. If microphysical causal closure is true, what appear as highly extrinsically intelligible causal relations from the physical aspect are grounded in less extrinsically intelligible relations between macroexperiences that appear from the mental aspect. If only physical causal closure is true, macroexperience grounds emergent causation, i.e., less extrinsically intelligible relations, and this is apparent from both aspects.

7

Two Accounts of Causation: the Agentive and the Hylomorphic

I have defended two different accounts of the intelligibility of causation. According to the argument from causation, defended in chapters 3 and 4, we experience *fully* intelligible, *necessary* connections between motives and efforts – we understand how motives such as pain *must* lead to avoidance efforts, *ceteris paribus.* This is the agentive account. According to my proposed solution to the intelligibility aspect of the combination problem, defended in chapter 5, there are *partially* intelligible, *possible* connections between different forms of the same matter – we understand how one form of matter *can* result from another form of the same matter. This is the hylomorphic account. In this final, short chapter, I will show how these two accounts fit together.

1 IDENTITY AND DISTINCTNESS

According to the agentive account, we only experience necessary connections within one and the same subject. The connection between efforts and their results outside the subject (or their results within the same subject, for that matter, as in mental action[1]) is not directly experienced – as Hume and other critics of volitional accounts of causation have, in my view, rightly emphasized. The hylomorphic account makes partially intelligible how there can be connections between different forms of the same matter. I argued that we should take subjects to be the fundamental forms of experiential matter. Therefore, the hylomorphic account makes partially intelligible how there can be connections between subjects and fills in the gap in the agentive account. The accounts are not in competition, because the agentive account concerns immanent causation, connections within a subject, and the hylomorphic account concerns transeunt causation, connections between subjects.

[1] That the connection between efforts and results within the subject is as unintelligible as the connection between the efforts and results outside the subject can be explained by the identity view of subjects, experiences and contents. A mental action and the mental state it results in will be different experiences, and therefore different subjects in the fundamental sense (i.e., thin subjects). All connections between efforts and results are therefore transeunt, i.e., hold between different subjects in the fundamental sense, and therefore equally unintelligible.

The identity view of subjects, experiences and contents, defended by Strawson, is essential to the reduction of subjects to forms of experiential matter, which is required in order for mental combination to be intelligible according to the hylomorphic account. As discussed in chapter 4 (section 2.4), the identity view is also helpful for the agentive account because it lets it avoid an aspect of agent-causation that many find implausible, the fundamental persistence of agentive subjects. However, the identity view can also seem to be in tension with the agentive account. On the agentive account, the subject is the causal nexus between motives and efforts, and both the motives and the efforts must be the subject's own experiences. If the experiences of the subject are not really distinct from it, as per the identity view, it means that the subject in fact causally connects its own parts. This lack of distinctness gives rise to conflict with some traditional conceptions of causation and its relata.

Firstly, causes are traditionally conceived of as temporally prior to their effects. The identity view entails that subjects fundamentally speaking only last as long as an experience, that is, for a moment, and it is subjects in the fundamental sense (i.e., thin subjects) that connect motives and efforts. Motives and efforts connected by such as subject would therefore have to exist at the same moment, i.e., be simultaneous, which is in tension with the traditional view. One response to this is to argue that the traditional view should be abandoned; that causes and effects are in fact best conceived as simultaneous. This was Kant's view. He writes:

> A great majority of efficient natural causes are simultaneous with their effects, and the sequence in time of the latter is due only to the fact that the cause cannot achieve its complete effects in one moment. But in the very moment in which the effect first comes to be, it is invariably simultaneous with the causality of its cause. (Kant 1781/1929: A203)

> For instance, a room is warm while the outer air is cool. I look around for the cause, and find a heated stove. Now the stove, as cause, is simultaneous with its effects, the heat of the room. They are simultaneous and yet the law is valid […] If I view as a cause the ball which impresses a hollow as it lies on a stuffed cushion, the cause is simultaneous with the effect. (Kant 1781/1929: A203)

On the simultaneity view, there is no temporal gap between the cause and effect. Causation can still take time, because, as Kant points out, the effect can take time to be fully achieved. The room is not heated instantaneously – the stove must continuously heat

it for several minutes. Another reason why causation takes time is because the cause and the effect can both be temporally extended, and only the end of the cause must be simultaneous with the beginning of the effect. When the stove gets warmer than the air around it, heat will start transferring to the air immediately without a temporal gap in between. However, it takes time for the stove to get to the point of being warmer than the room, and it takes time for heat to spread in the room. Mumford and Anjum (2011b: ch. 5) argue at length that a dispositionalist view of causation not only allows but also requires that causation is simultaneous.

A further response to the problem is to point out that experiences, while short-lived, are still temporally extended. Strawson claims that:

> There's almost universal agreement that we don't experience the present of experience just as a (moving) point or front in time that is itself temporally dimensionless, but as something that has an intrinsically temporally extended phenomenological character. (Strawson 2009: 249)

Strawson thinks it can be taken for granted that "experience takes time: it can't exist or occur at an instant, where an instant is defined as something with no temporal duration at all" (Strawson 2009: 256). He still takes experiences to be very short and speculates about durations from just milliseconds and up to a few seconds. But however brief experiences are, as long as their duration is not infinitesimally short, there will be time for motives to temporally precede efforts – either in the sense that there is a temporal gap between them, or in the sense allowed by the simultaneity view, according to which the experience of the motive begins before the experience of the effort, so that the beginning of the effort is simultaneous with the end of the motive only, as opposed to both experiences beginning exactly simultaneously.

Secondly, causation is traditionally conceived of as a necessary connection between distinct existences – thereby violating what is known as Hume's dictum. If motives and efforts are not really distinct, but are rather posteriorly carved out parts of a strongly unified total experiential field, then how can the necessary connections between them be causal? There are different ways of reading the terms "necessary" and "distinct" in this claim about causation. I will show that on the most straightforward reading, the connection between motives and efforts cannot qualify as causal; however, on closer inspection this reading can be seen to be contradictory and will thereby render it an analytic truth that no connections could possibly be causal. On another reading, which

removes the contradiction, the connection between motives and efforts does qualify as causal.

What does it mean for two existences to be distinct? It seems that two things are distinct if there is some sense in which they can *possibly* come apart. If two things cannot possibly come apart, in any sense of possibility, they are not distinct. But what does it mean for two existences to be necessarily connected? It seems that two things are necessarily connected if there is some sense in which they *cannot* possibly come apart. If distinctness and necessity are taken to involve the same kind of modality, then the idea of a necessary connection between distinct existences would clearly be contradictory. For example, there could not be metaphysically necessary connections between metaphysically distinct existences, because then it would both be metaphysically possible and metaphysically impossible for these existences to come apart. The contradiction is removed if the sense in which the two things can come apart is different from the sense in which they cannot come apart. For example, if the laws of nature are regarded as metaphysically contingent, it will not be contradictory to say that there are *nomologically* necessary connections between *metaphysically* distinct existences. On this view, it will be metaphysically possible, but nomologically impossible, for causes and effects to come apart.

But this reading cannot account for motives and efforts being distinct and necessarily connected. The connection between them is experienced as metaphysically necessary – pain (or affective, painful, pain) and avoidance effort cannot come apart (in the absence of interference from other motives) in any possible world. In general, dispositional theories of causation, of which the agentive account is a species, are often necessitarian, i.e., they entail that the laws of nature are metaphysically necessary. So this problem affects a broad class of theories of causation. All necessitarian theories must answer the question: if the necessary connection is metaphysically necessary, in what sense can the existences thus connected be distinct?

My answer to this question must refer to the sense in which individual experiences are in fact distinct on the identity view. I have argued that individual experiences must be regarded as different aspects of a total experiential field, posteriorly carved out of it along no actually pre-existing joints, but still constrained by objective relations of similarity and difference (see chapter 5, section 4.1.2). In this way, individual experiences can be regarded as aspectually and/or conceptually distinct. The causal relation between

experiences of motives and efforts would accordingly be a *metaphysically* necessary connection between *aspectually* or *conceptually* distinct existences.

2 A UNIFIED ACCOUNT

The hylomorphic account makes transeunt causation partially intelligible, while the agentive account makes immanent causation fully intelligible. They are thus compatible, but do they also depend on each other? In chapter 4 (section 2.6, p. 134), I argued that with respect to the intelligibility of transeunt causation the agentive account is compatible with any combination of the following three options: (1) to take a (fully or partially) mysterian attitude toward transeunt causation, (2) to argue that we can infer a positive conception of the nature of transeunt causation from our acquaintance with the nature of powers in effort, and (3) to argue that we can abstract away from immanent causation a positive account of transeunt causation. As I will shortly defend, the hylomorphic account of transeunt causation fits option (3). Therefore, it does not have to be accepted along with the agentive account, because the agentive account is compatible with choosing option (1) or (2) (or both) instead. The argument for the hylomorphic account does not depend on the agentive account either; I only defended it on the basis that it seems implicitly presupposed in science, metaphysics and philosophy of mind, and to a certain extent on the basis that it makes intuitive sense. The accounts are therefore still independent even though compatible.

However, both accounts could also be unified into a single theory of causation, by seeing how the hylomorphic account fits option (3). As I will now show, the agentive account and the hylomorphic account share are very similar structure, and therefore the latter can plausibly be regarded as an abstraction of the former. Such a unified account would be more complete and systematic, and therefore more attractive, than each of the views in isolation or their brute conjunction.

In chapter 3 (section 3.7), I argued against the viability of abstracting away the metaphysical structure of agent-causation and taking it as a basis for a theory of non-mental causation. However, I did not argue that abstracting the metaphysical structure of agent-causation is in itself problematic – I only argued that it is not suitable for the purpose of arriving at an account of non-mental causation of the sort to which agent-causation can be seen to reduce. The hylomorphic account is not supposed to be a reductive basis for the agentive account, so my previous criticism does not prohibit regarding the former as an abstraction from the latter.

What structural similarities are there between the two accounts? Both accounts require that causes and effects have something in common that connects them: they must either be experiences of the same agentive subject or forms of the same matter. In both cases, the common nexus is a *totality*. An agentive subject is a concrete total experiential field. The matter that forms have in common – when the forms are subjects – is the abstract totality of all experientiality. In this way, on both accounts, causation relates parts or portions of the same experiential whole – a concrete experiential whole in the concrete agentive case and an abstract experiential whole in the abstracted, hylomorphic case. A difference between the accounts is that in the agentive case the whole is prior to and more strongly unified that the parts, but in the hylomorphic case, the opposite is the case: the parts or forms are more strongly unified than the whole, the totality of experiential matter. Individual subjects are substances that are composites of form and matter, while the totality of matter has no form of its own, its form is only an aggregate of the forms of its parts. However, in this respect, the hylomorphic account systematically inverts the structure of the agentive account, and they remain isomorphic. In virtue of this inversion, both accounts result in subjects being the strongest unities that exist, while the unity of individual experiences and the unity of all experiential matter are both weaker and more abstract.

In view of these similarities, is seems like the hylomorphic account could actually be derived by abstraction from the agentive account, and perhaps ultimately justified on the basis of it, as opposed to on the weaker basis of making intuitive sense and actually being widely presupposed. However, I am not confident about any specific deductive connections between them, and so I will leave this as a proposal for how things could turn out and refrain for now from speculating any further about it.

The unified account could perhaps be labeled the *hylozoistic* account. Hylozoism is the Greek term for a variety of panpsychism, the view that matter (*hylē*) has life (*zōē*). Pre-Socratic hylozoists regarded life as an animating principle, the property of being the source of one's own activity or somehow directly connecting to a source of activity. Hylozoism can therefore be regarded as differing from hylomorphism and non-animistic panpsychism by requiring that that the forms of matter are intrinsically active, and in virtue of being active they must also be mental, in accordance with the agentive view.

3 PANPSYCHISM AND CAUSATION: SUMMARY

In this thesis, I have defended a new (relative to the present debate) argument for panpsychism and proposed an account of mental combination, both of which on the basis of views about the nature and intelligibility of causation. They were both offered as responses to the most serious problems for the two main lines of argument for panpsychism, the argument from categorical properties from philosophy of science and metaphysics, and the Hegelian and anti-mysterian arguments from philosophy of mind. In chapter 1 (section 3.2), I showed how the former line of argument is in danger of being undermined by a challenge from irreducible dispositionality, as posited by dispositionalism and related views. I also showed (section 3.1) how the latter line argument is in danger of being undermined by the combination problem, on the basis of which panpsychism can be accused of only moving and repeating the problems it is argued to solve.

According to the argument from causation, everything physical must have irreducibly causal properties, and since the only irreducibly causal properties we know are mental properties, it follows, via some supporting premises, that everything physical must have mental properties. In chapter 2, I presented the rich history of this argument, and showed its importance for panpsychism as well as the metaphysics of causation and agency in general. In chapter 3, I showed how the argument can be formulated as an argument from dispositional properties, which can again be generalized into an argument from non-structural properties, which makes panpsychism immune to the challenge from irreducible dispositionality. I then defended the soundness of the argument, with most of the emphasis on the highly controversial premise that the only irreducibly dispositional (and causal) properties we know, or have a positive conception of, are mental properties. In chapter 4, I further defended the argument against both empirically and philosophically grounded objections.

According to my account of combination, macroexperiences are fusions of microexperiences, and processes of fusion are causal processes. In chapter 5, I argued that this enables a solution to the intelligibility aspect of the combination problem, because causation can be regarded as partially intelligible on the basis of a hylomorphic theory of change. With this account, panpsychism would after all succeed in rendering the appearance of *our* kind of consciousness more intelligible, as the argument from philosophy of mind promises. In chapter 6, I argued that the fusion account enables a solution to the empirical aspect of the combination problem. Macroexperiential fusions

supplant their microexperiential bases, and will therefore not be causally redundant. Fusions can fit into the structure of a world closed under microphysics via the posit of multiple mental categorical grounding of physical dispositions. It fits optimally into the structure of a world closed only under the physical sciences broadly construed. On this account, then, panpsychism would after all succeed in causally integrating *our* kind of consciousness into the causal structure of reality as science reveals it, as the argument from philosophy of mind further promises.

In this final chapter, I have dissolved some apparent tensions between the two accounts of causation that form essential parts of the argument from causation and the causal account of combination respectively, the agentive account and the hylomorphic account, and indicated how they can be systematically unified.

BIBLIOGRAPHY

Alter, Torin, and Yujin Nagasawa. 2012. What Is Russellian Monism? *Journal of Consciousness Studies* 19 (9–10): 67–95.

Anscombe, G. E. M. 1971. *Causality and Determination: An Inaugural Lecture*. Cambridge: Cambridge University Press.

Aristotle. 1984. *The Complete Works of Aristotle: The Revised Oxford Translation*. Vol. 1. ed. J. Barnes. Princeton: Princeton University Press.

Armstrong, David M. 1983. *What Is a Law of Nature?* Cambridge: Cambridge University Press.

———. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Basile, Pierfrancesco. 2010. It Must Be True – but How Can It Be? Some Remarks on Panpsychism and Mental Composition. *Royal Institute of Philosophy Supplement* 85 (67): 93–112.

Basile, Pierfrancesco 2009. Back to Whitehead? Galen Strawson and the Rediscovery of Panpsychism. In *Mind That Abides: Panpsychism in the New Millenium*, ed. D. Skrbina. Amsterdam/Philadelphia: John Benjamins Pub. Co.

Bayne, Tim, and Neil Levy. 2006. The Feeling of Doing: Deconstructing the Phenomenology of Agency. In *Disorders of Volition*, eds. N. Sebanz and W. Prinz. Cambridge, MA: MIT Press.

Beebee, Helen. 2006. Does Anything Hold the Universe Together? *Synthese* 149 (3): 509–533.

———. 2011. Necessary Connections and the Problem of Induction. *Noûs* 45 (3): 504–527.

Bennett, Karen. 2003. Why the Exclusion Problem Seems Intractable and How, Just Maybe, to Tract It. *Noûs* 37 (3): 471–497.

Bird, Alexander. 2007. *Nature's Metaphysics: Laws and Properties*. Oxford: Clarendon Press.

Blackburn, Simon W. 1990. Filling in Space. *Analysis* 50 (2): 62–65.

Boogerd, F. C., F. J. Bruggeman, R. C. Richardson, A. Stephan, and H. V. Westerhoff. 2005. Emergence and Its Place in Nature: A Case Study of Biochemical Networks. *Synthese* 145 (1): 131–164.

Bourget, David, and David J. Chalmers. 2013. What Do Philosophers Believe? *Philosophical Studies*: 1–36.

Broad, C. D. 1925. *The Mind and Its Place in Nature*. London: Kegan Paul, Trench, Trubner & Co.

Callender, Craig. 2001. Taking Thermodynamics Too Seriously. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32 (4): 539–553.

Cameron, Ross. 2002. Quantification, Naturalness and Ontology. In *New Waves in Metaphysics*, ed. A. Hazlett. Basingstoke: Palgrave Macmillan.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Clarendon Press.

———. 1994. Fundamentalism Vs. The Patchwork of Laws. *Proceedings of the Aristotelian Society* 94: 279–292.

———. 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chalmers, David J. 1995. Facing up to the Problem of Consciousness. *Journal of Consciousness Studies* 2 (3): 200–219.

———. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
———. 2003. Consciousness and Its Place in Nature. In *Blackwell Guide to Philosophy of Mind*, eds. S. P. Stich and T. A. Warfield. Malden, MA: Blackwell.
———. 2009. The Two-Dimensional Argument against Materialism. In *Oxford Handbook to the Philosophy of Mind*, eds. B. P. McLaughlin and S. Walter. Oxford: Oxford University Press.
———. 2010. *The Character of Consciousness*. New York: Oxford University Press.
———. forthcoming-a. The Combination Problem for Panpsychism. Pagination from draft, available from: http://consc.net/papers/combination.pdf.
———. forthcoming-b. Panpsychism and Panprotopsychism. Pagination from draft, available from: http://consc.net/papers/panpsychism.pdf.
Clifford, William K. 1874/1886. Body and Mind. In *Lectures and Essays*, eds. L. Stephen and F. Pollock. London: Macmillan.
Cohen, Marc S. 2012. Alteration and Persistence: Form and Matter in the *Physics* and *De Generatione Et Corruptione*. In *Oxford Handbook of Aristotle*, ed. C. Shields. New York: Oxford University Press.
Coleman, Sam. 2012. Mental Chemistry: Combination for Panpsychists. *Dialectica* 66 (1): 137–166.
———. 2013a. Neurocosmology. In *The Nature of Phenomenal Qualities*, eds. P. Coates and S. Coleman. Oxford: Oxford University Press.
———. 2013b. The Real Combination Problem: Panpsychism, Micro-Subjects, and Emergence. *Erkenntnis*.
Collingwood, R. G. 1937. On the So-Called Idea of Causation. *Proceedings of the Aristotelian Society, New Series* 38: 85–112.
Coopersmith, Jennifer. 2010. *Energy, the Subtle Concept: The Discovery of Feynman's Blocks from Leibniz to Einstein*. Oxford: Oxford University Press.
Dainton, Barry. 2010. Phenomenal Holism. *Royal Institute of Philosophy Supplement* 85 (67): 113–139.
Davidson, Donald. 1973. Freedom to Act. In *Essays on Freedom of Action*, ed. T. Honderich. London: Routledge and Kegan Paul.
Dowe, Phil. 1992. Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science* 59 (2): 195–216.
———. 2008. Causal Processes. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Fall 2008 ed. Available from: http://plato.stanford.edu/archives/fall2008/entries/causation-process/.
Dowell, J. L. 2006. Formulating the Thesis of Physicalism. *Philosophical Studies* 131 (1): 1–23.
Dupré, John. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA and London: Harvard University Press.
Edwards, Paul. 1967. Panpsychism. In *The Encyclopedia of Philosophy*, ed. P. Edwards. New York: Macmillan.
Ellis, Brian. 2002. *The Philosophy of Nature: A Guide to the New Essentialism*. McGill Queens University Press.
———. 2010. An Essentialist Perspective on the Problem of Induction. *Principia* 2 (1): 103–124.
Esfeld, Michael. 2007. Mental Causation and the Metaphysics of Causation. *Erkenntnis* 67 (2): 207–220.
———. 2010. Causal Overdetermination for Humeans? *Metaphysica* 11 (2): 99–104.

Euler, Leonhard. 1833. *Letters of Euler on Different Subjects in Natural Philosophy: Addressed to a German Princess*. Vol. 1. ed. D. Brewster. New York: J. & J. Harper.
Everitt, Nicholas. 1991. Strawson on Laws and Regularities. *Analysis* 51 (4): 206–208.
Fair, David. 1979. Causation and the Flow of Energy. *Erkenntnis* 14 (3): 219–250.
Fine, Kit. 2012. Guide to Ground. In *Metaphysical Grounding: Understanding the Structure of Reality*, eds. F. Correia and B. Schnieder. Cambridge: Cambridge Univsersity Press.
Fodor, Jerry A. 1989. Making Mind Matter More. *Philosophical Topics* 17 (1): 59–79.
Ford, Marcus P. 1981. William James: Panpsychist and Metaphysical Realist. *Transactions of the Peirce Society* 17 (2): 158–170.
Garber, Daniel. 2009. *Leibniz: Body, Substance, Monad*. New York: Oxford University Press.
Ginet, Carl. 1997. Freedom, Responsibility, and Agency. *The Journal of Ethics* 1 (1): 85–98.
Goff, Philip. 2006. Experiences Don't Sum. *Journal of Consciousness Studies* 13 (10–11): 53–61.
———. 2009. Why Panpsychism Doesn't Help Us Explain Consciousness. *Dialectica* 63 (3): 289–311.
———. 2011. A Posteriori Physicalists Get Our Phenomenal Concepts Wrong. *Australasian Journal of Philosophy* 89 (2): 191–209.
———. forthcoming. The Phenomenal Bonding Solution to the Combination Problem. In *Panpsychism*, eds. G. Brüntrup and L. Jaskolla. Oxford: Oxford University Press.
———. manuscript. Consciousness and the Limits of Science. http://www.philipgoffphilosophy.com/publications.html.
Grahek, Nikola. 2007. *Feeling Pain and Being in Pain*. Cambridge, MA: MIT Press.
Harbecke, Jens. 2008. *Mental Causation: Investigating the Mind's Powers in a Natural World*. Frankfurt am Main: Ontos Verlag.
Hartshorne, Charles. 1934. *The Philosophy and Psychology of Sensation*. Chicago: The University of Chicago Press.
———. 1954. Causal Necessities: An Alternative to Hume. *The Philosophical Review* 63 (4): 479–499.
———. 1977. Physics and Psychics: The Place of Mind in Nature. In *Mind in Nature: The Interface of Science and Philosophy*, eds. J. B. Cobb and D. R. Griffin. New York: University Press of America.
Hawthorne, John. 2001. Causal Structuralism. *Noûs* 35 (s15): 361–378.
Hill, Christopher S. 1997. Imaginability, Conceivability, Possibility and the Mind-Body Problem. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 87 (1): 61–85.
Hoefer, Carl. 2008. For Fundamentalism. In *Nancy Cartwright's Philosophy of Science*, eds. L. Bovens, C. Hoefer and S. Hartmann. London: Routledge.
Horgan, Terry. 1996. Reduction, Reductionism. In *The Encyclopedia of Philosophy: Supplement*, ed. D. M. Borchert. New York: Simon & Schuster Macmillan.
———. 2007. Mental Causation and the Agent-Exclusion Problem. *Erkenntnis* 67 (2): 183–200.
———. 2011. The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits. *Humana Mente* 15: 77–97.
Horgan, Terry, and Matjaž Potrč. 2000. Blobjectivism and Indirect Correspondence. *Facta Philosophica* 2: 249–270.

Hume, David. 1739–40/1995. *A Treatise of Human Nature*. In *The Complete Works and Correspondence of David Hume*, ed. M. C. Rooks. Charlottesville: InteLex Corporation.
———. 1748/1995. *An Inquiry Concerning Human Understanding*. In *The Complete Works and Correspondence of David Hume,* ed. M. C. Rooks. Charlottesville: InteLex Corporation.
Humphreys, Paul W. 1997. How Properties Emerge. *Philosophy of Science* 64 (1): 1–17.
Jackson, Frank. 1986. What Mary Didn't Know. *The Journal of Philosophy* 83 (5): 291–295.
James, William. 1890/1981. *The Principles of Psychology Vol. 1*. In *The Works of William James*, Vol. 8, eds. F. H. Burkhardt, F. Bowers and I. K. Skrupskelis. Cambridge, MA and London: Harvard University Press.
———. 1902/1987. The Varieties of Religious Experience. In *William James: Writings 1902–1910*, ed. B. Kuklick. New York: Library of America.
———. 1907/1975. *Pragmatism: A New Name for Some Old Ways of Thinking*. In *The Works of William James*, Vol. 1, eds. F. H. Burkhardt, F. Bowers and I. K. Skrupskelis. Cambridge, MA and London: Harvard University Press.
———. 1909/1977. *A Pluralistic Universe*. In *The Works of William James*, Vol. 4, eds. F. H. Burkhardt, F. Bowers and I. K. Skrupskelis. Cambridge, MA and London: Harvard University Press.
———. 1911. *Some Problems of Philosophy: A Beginning of an Introduction to Philosophy*. New York, London, Bombay and Calcutta: Longmans, Green & Co.
———. 1912a. *Essays in Radical Empiricism*. New York, London, Bombay and Calcutta: Longmans, Green & Co.
———. 1912b. The Experience of Activity. In *Essays in Radical Empiricism*. New York, London, Bombay and Calcutta: Longmans, Green & Co.
Jammer, Max. 1957. *Concepts of Force: A Study in the Foundations of Dynamics*. Cambridge, MA: Harvard University Press.
Kant, Immanuel. 1781/1929. *Critique of Pure Reason*. Translated by N. Kemp Smith. London: MacMillain.
———. 1786/2004. *Metaphysical Foundations of Natural Science*. Translated by M. Friedman. Cambridge: Cambridge University Press.
Kim, Jaegwon. 1989. Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives* 3: 77–108.
———. 1999. Making Sense of Emergence. *Philosophical studies* 95 (1): 3–36.
———. 2007. Causation and Mental Causation. In *Contemporary Debates in Philosophy of Mind*, eds. B. P. McLaughlin and J. D. Cohen. Malden, MA: Blackwell.
Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Kroedel, Thomas. 2008. Mental Causation as Multiple Causation. *Philosophical Studies* 139 (1): 125–143.
Ladyman, James. 2013. Structural Realism. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Summer 2013 ed. Available from: http://plato.stanford.edu/archives/sum2013/entries/structural-realism/.
Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Clarendon Press.
Langton, Rae. 1998. *Kantian Humility: Our Ignorance of Things in Themselves*. Oxford: Oxford University Press.
Leibniz, Gottfried Wilhelm. 1691/1965. De Primae Philosophiae Emendatione, Et De Notione Substantiae. In *Die Philosophischen Schriften*, ed. C. I. Gerhardt. Hildesheim: G. Olms.

———. 1695/1989a. A New System of the Nature and Communication of Substances, and of the Union of the Soul and Body. In *Philosophical Essays*, eds. R. Ariew and D. Garber. Indianapolis: Hackett Publishing Company.
———. 1695/1989b. A Specimen of Dynamics, toward Uncovering and Reducing to Their Causes Astonishing Laws of Nature Concerning the Forces of Bodies and Their Actions on One Another. In *Philosophical Essays*, eds. R. Ariew and D. Garber. Indianapolis: Hackett Publishing Company.
———. 1698/1908. On Nature in Itself. In *Philosophical Works of Leibnitz*, ed. G. M. Duncan. New Haven: The Tuttle, Morehouse & Taylor Company.
———. 1699–1706/1989. From the Letters to De Volder. In *Philosophical Essays*, eds. R. Ariew and E. Watkins. Indianapolis: Hackett Publishing Company.
———. 1704/1981. *New Essays on Human Understanding*. Translated by P. Remnant and J. Bennett. Eds. P. Remnant and J. Bennett. Cambridge/New York: Cambridge University Press.
———. 1704/1997. Letter to Damaris Masham, May 1704. In *Leibniz's 'New System' and Associated Contemporary Texts*, eds. R. S. Woolhouse and R. Francks. Oxford: Clarendon Press.
———. 1714/1925. *The Monadology and Other Philosophical Writings*. Translated by R. Latta. London: Oxford University Press.
Levine, Joseph. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64 (4): 354–361.
Lewis, David. 1986. *On the Plurality of Worlds*. Malden, MA: Blackwell.
Lewis, David, and Rae Langton. 1998. Defining 'Intrinsic'. *Philosophy and Phenomenological Research* 58 (2): 333–345.
Lewtas, Pat. 2012. Building Minds: Solving the Combination Problem. Paper read at the workshop: Panpsychism at the Reef, at Great Barrier Reef/Australian National University.
———. 2013. What It Is Like to Be a Quark. *Journal of Consciousness Studies* 20 (9–10): 9–10.
Libet, Benjamin. 1985. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *The behavioral and brain sciences* 8: 529–566.
Loar, Brian. 1997. Phenomenal States Ii. In *The Nature of Consciousness: Philosophical Debates*, eds. N. Block, O. Flanagan and G. Güzeldere The Mit Press.
Loewer, Barry M. 2007. Mental Causation, or Something near Enough. In *Contemporary Debates in Philosophy of Mind*, eds. B. P. McLaughlin and J. D. Cohen. Malden, MA: Blackwell.
Mach, Ernst. 1886. *Beiträge Zur Analyse Der Empfindungen*. Jena: Verlag von Gustav Fischer.
———. 1897. *Popular Scientific Lectures*. Translated by T. J. McCormack. Chicago: The Open Court Publishing Company.
Madden, Edward H., and Peter H. Hare. 1971. The Powers That Be. *Dialogue* 10 (1): 12–31.
Marmodoro, Anna. undated. *Power Structuralism in Ancient Ontologies Project* Available from: http://www.power-structuralism.ox.ac.uk/about_the_project [cited 01.11.2013].
Marras, Ausonio. 2007. Kim's Supervenience Argument and Nonreductive Physicalism. *Erkenntnis* 66 (3): 305–327.
Martin, C. B., and John Heil. 1999. The Ontological Turn. *Midwest Studies in Philosophy* 23 (1): 34–60.

Martin, C. B., and Karl Pfeifer. 1986. Intentionality and the Non-Psychological. *Philosophy and Phenomenological Research* 46 (4): 531–554.
McGinn, Colin. 1989. Can We Solve the Mind–Body Problem? *Mind* 98 (391): 349–366.
———. 2006. Hard Questions – Comments on Galen Strawson. *Journal of Consciousness Studies* 13 (10–11): 90–99.
McLaughlin, Brian P. 1992. The Rise and Fall of British Emergentism. In *Emergence or Reduction?: Prospects for Nonreductive Physicalism*, eds. A. Beckermann, H. Flohr and J. Kim De Gruyter.
Merricks, Trenton. 2001. *Objects and Persons*. Oxford: Clarendon Press.
Molnar, George. 2003. *Powers: A Study in Metaphysics*. Oxford: Oxford University Press.
Montero, Barbara. 2006. What Does the Conservation of Energy Have to Do with Physicalism? *Dialectica* 60 (4): 383–396.
Mumford, Stephen. 2004. *Laws in Nature*. New York: Routledge.
Mumford, Stephen, and Rani Lill Anjum. 2011a. Dispositional Modality. In *Lebenswelt Und Wissenschaft, Deutsches Jahrbuch Philosophie 2*, ed. C. F. Gethmann. Hamburg: Meiner Verlag.
———. 2011b. *Getting Causes from Powers*. Oxford: Oxford University Press.
Nagel, Thomas. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83 (Oct.): 435–450.
———. 1979. Panpsychism. In *Mortal Questions*. Cambridge: Cambridge University Press.
———. 1986. *The View from Nowhere*. Oxford: Oxford University Press.
———. 2012. *Mind and Cosmos*. New York: Oxford University Press.
Nahmias, Eddy A. 2002. When Consciousness Matters: A Critical Review of Daniel Wegner's the Illusion of Conscious Will. *Philosophical Psychology* 15 (4): 527–541.
Nahmias, Eddy, Stephen G. Morris, Thomas Nadelhoffer, and Jason Turner. 2004. The Phenomenology of Free Will. *Journal of Consciousness Studies* 11 (7–8): 162–179.
Newman, Max H. A. 1928. Mr. Russell's "Causal Theory of Perception". *Mind* 37 (146): 137–148.
O'Connor, Timothy. 1994. Emergent Properties. *American Philosophical Quarterly* 31 (2): 91–104.
———. 1995. Agent Causation. In *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, ed. T. O'Connor Oxford University Press.
———. 1996. Why Agent Causation? *Philosophical Topics* 24, pagination from penultimate draft, available from: http://www.indiana.edu/~scotus/files/Why_Agent_Causation.pdf.
———. 2000. *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
O'Connor, Timothy, and Hong Yu Wong. 2005. The Metaphysics of Emergence. *Noûs* 39 (4): 658–678.
O'Sullivan, Trish. 2010. Buddha-Nature. In *Encyclopedia of Psychology and Religion: Springer Reference*, ed. D. A. Leeming. Berlin and Heidelberg: Springer. Available from: http://www.springerreference.com/docs/html/chapterdbid/70271.html.
Papineau, David. 2001. The Rise of Physicalism. In *Physicalism and Its Discontents*, eds. C. Gillett and B. Loewer. Cambridge: Cambridge University Press.
———. 2002. *Thinking About Consciousness*. Oxford: Clarendon Press.

Parfit, Derek. 1971. Personal Identity. *The Philosophical Review* 80 (1): 3–27.
Perry, Ralph Barton. 1935. *The Thought and Character of William James: As Revealed in Unpublished Correspondence and Notes, Together with His Published Writings*. London: Oxford University Press.
Pettit, Philip. 1994. A Definition of Physicalism. *Analysis* 53 (4): 213–223.
Place, U. T. 1996. Intentionality as the Mark of the Dispositional. *Dialectica* 50 (2): 91–120.
Price, Huw. 1991. Agency and Probabilistic Causality. *The British Journal for the Philosophy of Science* 42: 157–176.
Reid, Thomas. 1788. *Essays on the Active Powers of Man*. Edinburgh: John Bell and G. G. J. & J. Robinson, London.
Robb, David, and John Heil. 2013. Mental Causation. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Spring 2013 ed. Available from: http://plato.stanford.edu/archives/spr2013/entries/mental-causation/.
Rosenberg, Gregg. 2004. *A Place for Consciousness*. Oxford: Oxford University Press.
Russell, Bertrand. 1912. On the Notion of Cause. *Proceedings of the Aristotelian Society* 13: 1–26.
———. 1927. *The Analysis of Matter*. London: Kegan Paul, Trench, Trubner & Co.
———. 1948/1992. *Human Knowledge: Its Scope and Limits*. London: Routledge.
Schaffer, Jonathan. 2009. On What Grounds What. In *Metametaphysics*, eds. D. J. Chalmers, D. Manley and R. Wasserman. Oxford: Oxford University Press.
Schiller, Ferdinand C. S. 1906. Humism and Humanism. *Proceedings of the Aristotelian Society, New Series* 7: 93–111.
———. 1907. *Studies in Humanism*. London: MacMillan.
Schopenhauer, Arthur. 1859/1966a. *The World as Will and Representation Vol. 1*. Translated by E. F. J. Payne. New York: Dover.
———. 1859/1966b. *The World as Will and Representation Vol. 2*. Translated by E. F. J. Payne. New York: Dover.
Schrenk, Markus. 2010. The Powerlessness of Necessity. *Noûs* 44 (4): 725–739.
Seager, William. 2006. The 'Intrinsic Nature' Argument for Panpsychism. *Journal of Consciousness Studies* 13 (10–11): 129–145.
———. 2010. Panpsychism, Aggregation and Combinatorial Infusion. *Mind and Matter* 8 (2): 167–184.
———. 2012. *Natural Fabrications: Science, Emergence and Consciousness*. Berlin and Heidelberg: Springer.
Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
———. 1984. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.
———. 1990. Consciousness, Explanatory Inversion, and Cognitive Science. *Behavioral and Brain Science* 13: 585–642.
Searle, John R., Daniel C. Dennett, and David J. Chalmers. 1997. *The Mystery of Consciousness*. London: Granta.
Shapiro, Stewart. 1997. *Philosophy of Mathematics: Structure and Ontology*. Oxford: Oxford University Press.
Shoemaker, Sydney. 1980. Causality and Properties. In *Time and Cause: Essays Presented to Richard Taylor*, ed. P. van Inwagen. Dordrecht: Reidel.
———. 2002. Kim on Emergence. *Philosophical Studies* 58 (1–2): 53–63.
Sider, Theodore. forthcoming. Against Parthood. *Oxford Studies in Metaphysics*.
Skrbina, David. 2005. *Panpsychism in the West*. Cambridge, MA: MIT Press.

Stephan, Achim. 2002. Emergentism, Irreducibility, and Downward Causation. *Grazer Philosophische Studien* 65 (1): 77–93.
Steward, Helen. 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.
Stoljar, Daniel. 2001. Two Conceptions of the Physical. *Philosophy and Phenomenological Research* 62 (2): 253–281.
———. 2009. Physicalism. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Fall 2009 ed. Available from: http://plato.stanford.edu/archives/fall2009/entries/physicalism/.
———. 2010. *Physicalism*. New York: Routledge.
Stout, George Frederick. 1931. *Mind and Matter: The First of Two Volumes Based on the Gifford Lectures Delivered in the University of Edinburgh in 1919 and 1921*. Oxford: Macmillan.
Stout, George Frederick, C. A. Mace, A. C. Ewing, and C. D. Broad. 1935. Symposium: Mechanical and Teleological Causation. *Proceedings of the Aristotelian Society, Supplementary Volumes* 14: 22–112.
Strawson, Galen. 1987. Realism and Causation. *The Philosophical Quarterly* 37 (148): 253–277.
———. 1989. *The Secret Connexion: Causation, Realism and David Hume*. Oxford: Oxford University Press.
———. 2006a. Panpsychism? Reply to Commentators with a Celebration of Descartes. *Journal of Consciousness Studies* 13 (10–11): 184–280.
———. 2006b. Realistic Monism: Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies* 13 (10–11): 3–31.
———. 2008a. The Identity of the Categorical and the Dispositional. *Analysis* 68 (4): 271–282.
———. 2008b. What Is the Relation between an Experience, the Subject of the Experience, and the Content of the Experience? In *Real Materialism*. Oxford: Clarendon Press.
———. 2009. *Selves: An Essay in Revisionary Metaphysics*. Oxford: Oxford University Press.
———. 2010. Fundamental Singleness: How to Turn the 2nd Paralogism into a Valid Argument. *Royal Institute of Philosophy Supplement* 85 (67): 61–92.
Suppes, Patrick. 1974. Aristotle's Concept of Matter and Its Relation to Modern Concepts of Matter. *Synthese* 28 (1): 27–50.
Szallasi, Z., J. Stelling, and V. Periwal. 2006. *System Modelling in Cellular Biology: From Concepts to Nuts and Bolts*. Cambridge, MA: MIT Press.
Tegmark, Max. 1998. Is "the Theory of Everything" Merely the Ultimate Ensemble Theory? *Annals of Physics* 270 (1): 1–51.
Tully, R. E. 2003. Russell's Neutral Monism. In *The Cambridge Companion to Bertrand Russell*, ed. N. Griffin. Cambridge: Cambridge University Press.
Turausky, Keith. 2012. Picturing Panspychism. Paper read at the workshop: Panpsychism at the Reef, at Great Barrier Reef/Australian National University.
van Inwagen, Peter. 1990. *Material Beings*. Ithaca: Cornell.
Ward, James. 1915. *Naturalism and Agnosticism*. London: A. & C. Black, Ltd.
Wegner, Daniel M. 2003. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.
———. 2004. Précis of *The Illusion of Conscious Will*. *Behavioral and Brain Sciences* 27 (5): 649–659.
Wilson, Catherine. 2006a. Commentary on Galen Strawson. *Journal of Consciousness Studies* 13 (10–11): 177–183.

Wilson, Jessica M. 2006b. On Characterizing the Physical. *Philosophical Studies* 131 (1): 61–99.

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. In *Oxford Studies in the Philosophy of Science*. Oxford: Oxford University Press.

———. 2007. Interventionist Theories of Causation in Psychological Perspective. In *Causal Learning: Psychology, Philosophy, and Computation*, eds. A. Gopnik and L. Schulz. New York: Oxford University Press.