Univerzita Karlova

Filozofická fakulta

Ústav filosofie a religionistiky

Filosofie

Disertační práce

Consciousness in Nature. A Russellian Approach

Vědomí v přírodě. Russellovský přístup

školitel: doc. James Hill, PhD

2016

Mgr. Jakub Mihálik

Prohlášení:

Prohlašuji, že jsem disertační práci napsal samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 5. září 2016

.....

I would like to thank my advisor, James Hill, for his patient guidance as well as much inspiration throughout the years. I am also grateful to Philip Goff, Sam Coleman, Galen Strawson, Michaela Košová, Pavla Toráčová and Luke Roelofs for fruitful conversations about topics discussed in this thesis or insightful – written or spoken – comments on parts of the thesis. A special thank you goes to my wife, Sara Anne Mihálik, for proofreading the manuscript, continued support and so much more.

Table of Contents:

Foreword

The present volume has a two-part title. While the first part of its title, *Consciousness in Nature*, is fairly self-explanatory and will be further clarified in the first, introductory chapter of the volume, the subtitle, *A Russellian Approach* will not be explained until the penultimate chapter of the volume. It therefore perhaps deserves a few words of clarification in this foreword. First and foremost, it must be emphasised that the present volume is not a study in the philosophy of Bertrand Russell and many claims which it includes will be distinctly non-Russellian (for example, the two-dimensional semantic framework discussed in chapters 3 and 4, or the broadly Fregean view of truths or propositions presupposed throughout the volume). Still, Russell's thinking was an inspiration for the resulting position which this volume recommends as a way of integrating consciousness into the physical world.

This position, often called Russellian monism in the literature, has its roots in Russell's insights made in *The Analysis of Matter* and elsewhere.[1] In his work, Russell repeatedly dealt with the philosophical implications of modern physics and tackles the question as to what kind of knowledge of the physical world physics provides us with. His answer is that physics provides us merely with highly abstract, structural information about the physical world and the events which happen in it. There is then, arguably, much information about the physical world that escapes our observations, measuring devices and, as a result, our theories.

These epistemological limitations which, according to Russell, our physical theories exhibit, led him to the hypothesis that the physical events which constitute our brains have an intrinsic aspect, i.e. the aspect that escapes our physical theories, and that this intrinsic aspect consists in instantiations of qualities, such as the blueness of blue, which we are, he thinks, directly acquainted with from our perception. This hypothesis allows Russell to naturally integrate our percepts with their qualitative features, into the physical world. This hypothesis, argues Russell, allows us to answer the question what the hidden aspect of at least some physical events is, and the question what the place of percepts in the physical world is, both at the same time. It is this fascinating hypothesis which has much resonated in the contemporary debate of phenomenal consciousness and the prospects of its reduction and it is this hypothesis which I shall explore in the final two chapters of this volume.

---

[1] Russell (1954).

1. Consciousness in the Physical World

1. Two Images of Man

Owing to the immense progress of natural sciences triggered by the late 16th and 17th century scientific revolution, we have been learning a great deal about the physical, chemical and biological constitution of our universe and the laws which govern the processes and entities in it. We have, among other things, learned that all or most concrete things in the universe are ultimately – in a more, or less elaborate manner of organisation – composed of a set of fundamental entities which are governed by a relatively small number of relatively simple microphysical laws. Science then, generally speaking, reveals to us a picture of the universe, in which ordinary macroscopic things, such as chairs, trees or laptops, are nothing but complex systems consisting of smaller, fundamental, physical entities, or – to use J. J. C. Smart's phrase – the former are "nothing over and above" the latter.[2] This feature of the scientific picture is usually expressed in literature by saying that macroscopic things are *reducible* to the organisation of fundamental physical entities.

Given the reductive nature of natural sciences, we can say that they, generally speaking, aspire to provide an account of properties and relations of macrophysical things in terms of microphysical entities, properties and relations. We are, for example, able to explain macro-properties such as "having the flu", "being lightning" or "being nutritious" in terms of micro-processes and laws described and studied by biology, chemistry and physics. Given that the aspirations of science are universal and that we, i.e. thinking, judging, feeling, desiring, imagining etc. organisms, are among the macro-entities which exist in the universe, it follows that it should be a part of scientific aspirations to reductively explain the properties which we possess. It is, however, a subject of much controversy whether we comfortably fit into the scientific picture, i.e. whether natural science could even in principle reductively explain the above-mentioned properties.

Of course, we can only ask a question of this kind if we possess a conception of ourselves which is distinct from the scientific conception. If, that is, apart from being able to think about ourselves in scientific terms, roughly, as of complex physical systems produced over many generations by the workings of natural selection on random mutations, we also possess a different sort of conception of ourselves. I think it is hard to deny that, apart from the scientific conception, we also have a conception of ourselves as conscious and thinking persons with moods and feelings who are free in the sense that – on a given occasion – could have acted otherwise and are able to act upon reasons. An influential attempt to describe such a conception of man has been made by Wilfrid Sellars who

[2]Smart (2002, p. 61).

calls this conception, using a visual metaphor, the "manifest image of man-in-the-world".[3] Possession of the manifest image of one's self is understood by Sellars as an essential aspect of being human. At the same time, however, Sellars notices, the manifest image seems to clash with what he calls the "scientific image of man-in-the-world", i.e. with the idealised conception of man-in-the-world, which is revealed to us by natural sciences.

Given that the enquiry which I shall pursue in this volume presupposes that we have such two distinct images of ourselves, it will be worthwhile to say a bit more about Sellars's distinction, which I think captures the tension between science and ordinary thinking sketched above really well. Before I do so, however, let me state that my discussion of Sellars's distinction does not aspire to be a significant contribution to the blossoming literature on the philosophy of Wilfrid Sellars. It is rather meant to shed some light on the above-sketched general philosophical problem of the apparent clash between ordinary thinking and science and to show – in general terms – what I take to be the appropriate philosophical response to this problem. I take it that the more specific problem of the existence of consciousness in nature, the main topic of this volume, is one important part of this more general problem. While I shall say much more about consciousness in the second part of this chapter as well as in the chapters that follow, I shall now introduce the general problem using the contrast between the manifest and scientific image of man as articulated by Sellars.

It is important to notice that Sellars carefully distinguishes the manifest image from a primitive, pre-scientific conception of ourselves and the world which he calls the "original image" and which, he suggests, amounts to a sort of animism, a view according to which the conceptions of inanimate things, such as clouds or rocks, include attributes of persons such as free will, character, beliefs, desires, etc.[4] The manifest image has, according to Sellars, developed from the original image by means of much refinement, both empirical (by inductive inference) and conceptual. The manifest image is then, despite its name, fairly sophisticated. At the same time, however, Sellars insists that the manifest image is crucially different from the scientific image of man in that the latter, but not the former, features as its main components imperceptible entities postulated by physics and other sciences, such as atoms, electrons, strings, fields, etc.[5] Sellars notices that the two images seem to clash with one another as each of them involves key elements which seem incompatible with certain key elements of the other. While the manifest image views man as a person with thoughts, feelings and desires, the scientific image depicts man as a complex system of imperceptible physical entities.[6] While the manifest image involves a conception of things around us as possessing "true" colours, according to the scientific image they merely possess surface-reflectance properties.

[3]Sellars (1991a).
[4]Sellars (1991a, p. 6–7).
[5]Sellars (1991a, p. 7).
[6]Sellars (1991a, p. 25).

The apparent clash between the manifest image and the scientific image is viewed by Sellars as one of the central problems of contemporary philosophy. In view of this, Sellars sets it as a crucial task for philosophers to develop what he calls a "stereoscopic view" which can be done by fusing or synthesizing the manifest and the scientific image into a single unified conception. Here, one's view of the world is in Sellars's sense stereoscopic if one (1) accepts that both the manifest and the scientific image are in their main elements true and reveal to us what the world is really like, and (2) manages to clarify how this is possible, i.e. how is it that both the manifest and the scientific image are true and reveal to us what the world is really like.[7]

There are two main ways philosophers can fail to even address the challenge of formulating a stereoscopic view envisioned by Sellars. Both of these failures consist in not meeting condition (1). Some philosophers fail to meet condition (1) because they accept only the manifest image as one which reveals to us the true nature of reality. They thus view the manifest image as the only reliable guide to ontology.[8] The scientific image is seen by these thinkers as merely a different, more abstract, systematic and mathematical, kind of description of the true reality revealed to us in the manifest image. Sellars suggests that the one-sidedness of this kind is integral to the broadly understood Platonist stream in philosophy and much of the continental philosophical tradition as well as the philosophies of ordinary language and common sense.[9] Historically, a view of this kind is associated with the works of George Berkeley for whom the theoretical posits of physics (such as the Newtonian gravitational attraction or submicroscopic particles) are merely useful fictions which help the scientist make more accurate predictions and offer better explanations. The corresponding physical terms, e.g. force-terms, then, according to Berkeley, do not denote any supposed imperceptible entities.[10] Physical science is then seen by Berkeley as a useful instrument for making predictions and calculations but not as a reliable guide to ontology. The Sellarsian scientific image then, according to Berkeley's conception of science, describes the realm of abstract, quasi-fictional entities rather than hidden, imperceivable realities.

Other philosophers are, according to Sellars, unable to even embark on the task of constructing a stereoscopic view because they accept only the scientific image as revealing to us the true nature of reality. Owing to this they also fail to meet condition (1). They thus view the scientific image as the only reliable guide to ontology, the manifest image is seen by them as a realm of mere appearance and illusion which does not reveal to us anything true. Sellars suggests that a view like this can be found in Spinoza, who in Sellars's interpretation, viewed science to be the only source of true

[7]Sellars (1991a, p. 5).
[8]In this volume I shall use the term "ontology" to denote, roughly, the set of entities, properties and relations which the universe fundamentally consists of. While this usage is arguably etymologically misguided, I shall adopt it as it is common in analytical (or mainstream Anglo-American) philosophy and helpful for my purposes.
[9]Sellars (1991a, p. 19).
[10]See e.g. Berkeley (1721/1965, § 39, p. 262).

information about man and the world. I myself have some doubts about this understanding of Spinoza and think that some of the recent proponents of scientism and eliminativism, such as Patricia Churchland, are rather closer to a view of this kind.

The present volume can be seen as an attempt to construct what Sellars calls a stereoscopic view, i.e. a view which would meet conditions (1) and (2). The project pursued here, however, will also significantly differ from that envisioned by Sellars in that it will involve much zooming in – it will focus almost exclusively on constructing a stereoscopic view of consciousness. Still, its general shape, if perhaps not its result, will be distinctly Sellarsian. Before I focus on this more specific, project, I shall now briefly describe two attempts to construct a stereoscopic view in the general sense envisioned by Sellars. While I shall focus mostly on the stereoscopic view introduced by Sellars himself, I will also briefly mention Descartes, who is viewed by Sellars as his notable predecessor in this kind of project.[11] Apart from their own interest, these two views will provide me with an opportunity to illustrate what a stereoscopic view is supposed to amount to.

For Descartes, the characteristics of human persons which are included in Sellars's manifest image, such as that we are conscious, thinking beings, provided a reason to supplement the physical world, which he conceived of in terms of the extended substance (*res extensa*) and its modes, with instances of immaterial, thinking substance (*res cogitans*).[12] According to Descartes, while the structure and properties of the extended substance are revealed to us by science, the workings of the immaterial souls, which we can learn about by means of philosophical meditation, escape – even in principle – the grasp of science. At the same time, however, the existence and nature of immaterial souls account for many of the features which the manifest image attributes to us, such as the fact that we are thinking, conscious beings endowed with free will. We can see then that, in the Cartesian picture, each image reveals a part of the truth about the universe.

Descartes's attempt to fuse the manifest and the scientific image left many unpersuaded, the crucial worry – famously brought to Descartes' attention in a letter from Elisabeth, Princess of Bohemia – being the question of how an immaterial and non-extended soul could causally interact with the physical world, given their radical difference. His system, nevertheless, is a good attempt to provide what Sellars later called a stereoscopic view of man-in-the-world. Notice that, according to the Cartesian picture, it is not surprising that the scientific and the manifest image seem to clash – they, after all, describe, at least in part, ontologically distinct substances. This picture then, in effect, offers us a natural account of the conceptual dualism, expressed in the form of the two images of man-in-the-world. The conceptual dualism has, given the Cartesian view, its roots in ontological

[11]Sellars (1991a, p. 26).
[12]Descartes (1641/1996), see esp. 6th Meditation.

dualism and the view can then be seen as an attempt – if an early and ultimately unsuccessful one – at constructing a stereoscopic view of man-in-the-world. Descartes, after all, attempts to meet both of the above-mentioned conditions, holding (1) that both the scientific view and the manifest view reveal to us some truths about the world and that (2) we can make sense of how both of these views can be true at the same time.

Sellars himself, dissatisfied with dualism, provides a very different attempt at constructing a stereoscopic view. At first sight, it may not be quite clear that Sellars's view is truly stereoscopic in the sense of meeting both of the above-mentioned conditions. He, after all, argues for the primacy of the scientific image which he takes to be a guide to ontology, telling us which entities and properties really exist. He indeed writes, paraphrasing Protagoras, that "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not".[13] Despite such claims which almost hint at scientism, it would be wrong to view Sellars as suggesting that the manifest image is illusory and that we therefore need to give it up or at least radically revise it. Indeed, as already mentioned, the possession of the manifest image is for Sellars an essential feature of ourselves as human beings. Moreover, according to Sellars the activity of describing and explaining the world mentioned in the quoted passage is only one aspect of our activities in the world; equally importantly we, normative beings, also prescribe and proscribe.[14] Even more crucially, the key features of the manifest image can be in Sellars's view shown to be compatible with the scientific image.

We can now see why Sellars characterises his view as stereoscopic. He thinks that both the scientific view and the manifest view reveal to us some truths about the world and that we can make sense of how both of these views can be true at the same time. If his project succeeds, it thus meets both of the above-mentioned conditions. This is not the place to discuss Sellars's proposal in all its details. Still, I shall now briefly mention why Sellars thinks that the existence of coloured objects and conceptual thought, as they are revealed in the manifest image, are compatible with the scientific image. Thereafter, I shall consider Sellars's suggestion that the existence of our conscious sensations is also compatible with the scientific image.

While the manifest image reveals to us objects around us as coloured, the scientific image characterises them, roughly, as swarms or clouds of tiny entities which are, on their own, imperceivable by the naked eye and colourless. Given that Sellars rejects the option that colours could somehow "emerge" in systems whose parts lack colour, he adopts the view that the coloured objects revealed to us in the manifest image are "'appearances' to a human mind of a reality which

[13]Sellars (1991b, p. 173).
[14]deVries (2015).

is constituted by systems of imperceptible particles".[15] With the help of the appearance / reality distinction then the seemingly conflicting claims about objects can, according to Sellars, be shown to be compatible in the end as each of the two claims is about a different sort of objects.

The manifest image further reveals ourselves as beings capable of conceptual thought. Moreover, Sellars notices, we are unable to analyse the activity of conceptual thought in terms of any more simple elements which would not themselves have a conceptual nature – analysis never takes us any further than to individual concepts. The scientific image, on the other hand, tells us, as we saw, that we are nothing but complex physical systems. Sellars, once again, thinks that these seemingly contradictory claims can be shown to be compatible in the end. The main motivation for rejecting the claim that conceptual thought is a physical process in the brain, is, according to Sellars, the view that conceptual thought is essentially qualitative – just like sensations and mental imagery. He considers this view, however, to be unjustified. According to Sellars, conceptual thought should instead be viewed as a system of abstract, functional roles, multiply realisable by systems which feature the required level of complexity and the required kind of organisation. Once we adopt this functionalist understanding of thought, it becomes an open question what the nature and constitution of the realising system is. It certainly seems possible that the realising system could be carbon-based, such as in the case of our brain, but, for Sellars, it is also an open possibility that the system could be a silicon-based super-computer.[16] Even in the case of conceptual thought then, Sellars argues, the manifest and the scientific image turn out to be compatible in the end.[17]

Finally, let me investigate Sellars's attempt to incorporate into the stereoscopic view conscious sensations which are revealed in the manifest image but which seem to be absent from the scientific image. Sellars's discussion of sensations is particularly interesting from our point of view, as we can witness here Sellars attempting to tackle an embryonic version of the problem which this whole volume is dedicated to.

The problem with sensations, according to Sellars, is that they exhibit what he calls "ultimate homogeneity".[18] To understand this feature of sensations, recall Sellars's treatment of coloured objects. We saw that the colours of objects, as revealed in the manifest image, belong in Sellars's view not to the objects themselves but rather to their appearances in human minds. These perceptible appearances of objects which are the true locus of colours, are, according to Sellars, placed in our sensations rather than in the objects themselves as revealed in the scientific image.[19] Considering the characteristics of these perceptible properties, Sellars finds that they are revealed to

[15]Sellars (1991a, p. 26).
[16]For more on the functionalist reduction of mentality see chapter 2.
[17]Sellars (1991a, p. 34).
[18]Sellars (1991a, p. 35).
[19]Sellars (1991a, p. 36).

us in the manifest image as ultimately homogeneous in the sense that any region of, say, a coloured field will still be a coloured field, any part of a heard sound will be a heard sound, any part of a smell will be odorous, etc. As he puts it, "colour expanses in the manifest world consist of regions which are themselves colour expanses, and these consist in their turn of regions which are themselves colour expanses, and so on".[20]

Ultimate homogeneity is, however, nowhere to be found in the world revealed to us in the scientific image. There, Sellars writes, "the state of a group of neurons, though it has regions which are also states of groups of neurons, has ultimate regions which are *not* states of groups of neurons but rather states of single neurons".[21] Here the thought is that in the physical world, by investigating the composition of a particular region of matter, say a group of neurons, we discover individual parts, say neurons, which do not have on their own the character of a group of neurons but rather the character of individual neurons. By investigating the composition of individual neurons we discover their individual parts, say electrons, which do not have the character of individual neurons but rather the character of parts of neurons, etc. The scientific image then reveals to us a certain kind of heterogeneity – groups of neurons consist of individual neurons, these consist of electrons, etc. This kind of heterogeneity is, as we saw, lacked by the manifest world where coloured fields consist of coloured fields, etc. and can therefore be described as 'ultimately homogeneous'.

It will not work as an objection against Sellars's view to say that surely perceived colour fields rarely or never feature the same colour shade across their whole area. Sellars does not, after all, speak about homogeneous fields in the specific sense of featuring the exact same colour shade across their whole area but rather in the general sense of being coloured. He thus appeals here to the "manifest" intuition that perceived fields could not have regions which would utterly lack the property of being coloured. This intuition, I believe, is rock-solid and is close to being a conceptual truth.

Working with the notion of ultimate homogeneity, Sellars presents the following argument for dualism:

1. Conscious sensations feature ultimate homogeneity.

2. Neural processes do not feature ultimate homogeneity.

———

3. Conscious sensations are not neural processes.

[20]Sellars (1991a, p. 35).
[21]Sellars (1991a, ibid.).

The conclusion of this simple argument is that sensations are not identical with any neural processes going on in our brains. However, a dualism of conscious sensations on the one hand, and neural processes on the other hand is rejected by Sellars.

We can see why dualism is unacceptable for Sellars, if we notice that our sensations are an essential part of a correct explanation of how we construct the manifest image of the world.[22] Think, for example, of how it can be explained that we perceive objects around us as colored, as we do according to the manifest image. It seems that whatever the actual details of the explanation will be, this explanation will crucially invoke our sensations of these objects. If we, however, consider what a correct scientific explanation of how we arrive at our conceptions, including the conception which is the manifest image, would look like, we shall see, Sellars suggests, that it would be an explanation in terms of a complex physical (or physically-realised) process. The scientific image, furthermore, tells us, according to Sellars, that the material world which the given neural process is a part of, is a closed system of explanation, i.e. all the processes happening in it are explained by appeal to other material processes – no other, non-material explanatory posits are needed.

These considerations lead Sellars to the following dilemma. If we accept the dualism resulting from the above-sketched argument, and view sensations as non-material, they will, given the "explanatory closure" of the material world, necessarily lack any explanatory relevance. That, however, goes directly against the intuitive view discussed above that our sensations are a key part of the explanation of how we have formed the manifest image. The alternative looks hardly more promising: if sensations are identified with material, neural processes, their ultimate homogeneity will – for the reasons discussed above — turn out to be mere appearance or illusion; notice that here Sellars clearly cannot opt to transfer the fields of ultimate homogeneity into the realm of the mind, as he did in the case of the colours, attributed to things around us by the manifest view. Here, after all, he already is in the realm of the mind.

Faced with these unattractive options, Sellars offers an interesting attempt at finding a way out. Namely, he argues that the above-sketched anti-materialist argument ultimately fails because its premise (2) can be rejected. According to this premise, neural processes do not feature ultimate homogeneity. As we saw, this premise is justified by the empirical observation that these processes consist of smaller-scale processes, featuring smaller-scale entities, and that these smaller-scale processes consist of yet smaller-scale processes featuring yet smaller-scale entities, etc. Sellars, nevertheless, in the following passage, suggests that premise (2) can be questioned:

> *[I]f it should turn out that particles instead of being the primitive entities of the scientific*

[22]Sellars (1991a, p. 36).

> *image could be treated as singularities in a space-time continuum which could be conceptually 'cut up' without significant loss* – in inorganic contexts*, at least – into interacting particles, then we would not be confronted at the level of neurophysiology with the problem of understanding the relation of* sensory consciousness *(with its ultimate homogeneity) to systems of particles.*[23]

In this intriguing and difficult passage Sellars speculates that the universe may not at its fundamental level have a nature of distinct particles but rather of a space-time continuum. Its description in terms of distinct particles thus leaves out the crucial feature of continuity. While this omission in inorganic contexts has no serious consequences, this continuity cannot be omitted in organic contexts as it is precisely this continuity which is to account for the above-discussed experienced ultimate homogeneity of conscious sensations. How can continuity explain homogeneity? The idea here seems to be that if physical processes are at their fundamental level continuous, they will consist of no further, smaller-scale particles and we thus get fields which are homogeneous. This sort of reasoning then seems to lead Sellars to the view that we can reject premise (2) of the anti-materialist argument as well as that whole argument.

Will these homogeneous fields in the physical world possess colours and other sensed properties which, according to Sellars, feature homogeneity? It looks like something like this is Sellars's view. He writes, for example, that once we penetrate to the non-particulate foundation of the scientific image, we find that "[…] in this non-particulate image the qualities of sense are a dimension of natural process which occurs only in connection with those complex physical processes which, when 'cut up' into particles […] become the complex system of particles which, in the current scientific image, *is* the central nervous system".[24] In this highly condensed passage Sellars seems to suggest that the experienced homogeneous qualities, such as colours or sounds, are real aspects of the physical world.

The existence of sensations, which are a part of the manifest image, is thus, for Sellars, compatible with what the scientific image tells us. Still, it is clear that sensations have a special place in Sellars's system because their existence gives us reasons to supplement the scientific image – clearly, after all, the current scientific image does not tell us that there are coloured fields in the material world.[25] Sensations then seem to be one aspect of the manifest image which serves as a guide to ontology – a role which, for Sellars, normally belongs to science. We can thus see that sensations are treated by Sellars quite differently from conceptual thought or colours of objects. While the manifest image reveals to us objects around us as coloured, the scientific image corrects

[23]Sellars (1991a, p. 37).
[24]Sellars (1991a, p. 37).
[25]See eg. deVries (2015).

this assumption and places colours (whose existence and nature the manifest image correctly informs us about) in mental sensations. While the manifest image reveals to us thought as irreducibly conceptual, the scientific image corrects this assumption and leads us to a view of thoughts as multiply realisable functional states. In the case of sensations, on the contrary, the manifest image motivates us to add something – irreducible sensory fields – to the material world described by science. Sensations then seem to have a special position in Sellar's thinking as they supplement, although do not revise, the scientific image.

Is, however, Sellars' treatment of sensations plausible? It is beyond the scope of this volume to provide a definitive answer to this question. Admittedly, Sellar's proposal is speculative and it is far from clear that physics will end up discovering continuity at the fundamental level. Still, one might perhaps take it that the discoveries of quantum mechanics concerning quantum holism give us some reason to take the continuum hypothesis seriously. However this may be, there is, I believe, an important challenge which any proposal along these lines needs to address, a challenge resulting from the fact that Sellars is concerned with *conscious* sensations. We will be able to see what this challenge consists of once we clarify the phenomenon of consciousness – a task I shall take on in the following section.

2. The Many Concepts of Consciousness

The words “consciousness” and “conscious” are commonly used in everyday communication. We say, for example, that we were or were not conscious of a particular fact or that someone lost or regained consciousness. The words are also frequently used in humanities – there people sometimes talk about class or gender consciousness, of the literary method of the stream of consciousness, etc. – these “expert uses”, however, are, as will become apparent, rather tangential to our topic. Leaving them aside, it seems clear that even in everyday communication, the word “consciousness” is used to express multiple different, although related, concepts. It will be useful for our purposes to distinguish between these concepts.

Firstly, we say that someone is conscious if they respond to us, their eyes are open, simply, they are awake. We use “conscious” in this sense in contrast to “asleep”, “knocked out” or “comatose”.[26] Consciousness in this sense is attributed to subjects on the basis of behavioural criteria, i.e. judging from how they behave or act and we can therefore usefully call this concept of consciousness *behavioural*. The conditions which a subject must satisfy in order to be behaviourally conscious are then purely of a third-person nature – whether an organism is conscious in this sense is a fact

[26]The property which a given organism must exhibit in order to count as conscious in this sense has been called “creature consciousness” by Rosenthal (1993, p. 355).

directly cognitively accessible to external observers.

Another concept of consciousness – and the concept I shall be primarily concerned with here – has been brought to the attention of contemporary philosophers by Thomas Nagel.[27] According to Nagel's famous definition, an organism is conscious if and only if there is something it is like to be that organism for that organism.[28] Nagel's definition of consciousness suggests that an organism is conscious if it possesses what can metaphorically be described as its own *subjective point of view*. Intuitively, a subjective point of view is possessed by e.g. rabbits, dogs, apes and humans but not by e.g. chairs, mountains or dandelions. It is natural to say that the organisms who are conscious in Nagel's sense have, when they are conscious, experiences. It seems to be, for example, a part of what it is like to be me now, that I am now experiencing the sound of water in the nearby fountain, the smell of coffee and the visual contrast of black letters on a white computer screen.

Often, organisms which are conscious in Nagel's sense are also conscious in the behavioural sense. An organism which is behaviourally conscious and thus e.g. reacts to external stimuli and sensibly answers our questions, normally simultaneously undergoes certain conscious episodes or experiences, which presumably include the experience of someone talking to them, the experience of thinking about the best way to reply, etc. In such cases, the organism satisfies both of the two mentioned concepts of consciousness since it is both awake and undergoing conscious, or experiential, episodes. Still, Nagel's concept is importantly different from the behavioural concept because an organism does not satisfy it in virtue of behaving or reacting in certain ways, as in the case of the behavioural concept, but rather in virtue of undergoing experiential episodes. We can say then that the conditions of satisfaction of the second concept are not behavioural but rather experiential or phenomenal. Given that they have such a nature, it is reasonable to suggest that it is the organism itself, at least when it comes to grown up humans without significant cognitive impairments, who is best able to establish whether she herself is or was conscious. The conditions which an organism must satisfy in order to count as conscious in the second sense then arguably have a first-person character as they are not directly cognitively accessible to external observers but rather only to the experiencing organism itself. As such, this concept importantly differs from the above-introduced behavioural concept whose conditions of application, as we saw, have a third-person nature.

Given the difference between the two concepts, could a subject actually be conscious in one sense without being conscious in the other? It is reasonable to suppose that in most cases a subject who satisfies one of the concepts also satisfies the other. At the same time, a dreaming organism or – to

[27]Nagel (1974).
[28]Nagel (1974, p. 436).

take a more extreme case – a patient suffering from the locked-in syndrome, will fail to satisfy the behavioural concept while having conscious experiences and thus satisfying Nagel's first-person concept.

The property of organisms picked out by Nagel's concept of consciousness can be seen as a problem for Sellars's treatment of sensations introduced earlier in this chapter. More importantly, it can also be seen as a challenge for any attempted materialist reduction of the mind. Let me now briefly sketch why, for Nagel, consciousness is a serious obstacle for materialist reductionism. If we accept Nagel's definition of consciousness, it is arguably hard to conceive of how consciousness could be identical to a physical brain process. Any such brain process seems, after all, perfectly objective in the sense that it could be observed, studied and described at least in principle by any intelligent being. This is according to Nagel not true about consciousness, as his famous example with a bat illustrates. Nagel argues that some kinds of conscious states are in principle unknowable by us, using the conscious experiences of a bat as his example. There is no way, Nagel argues, that we, as humans, could know what it is like to be a bat, as we are a very different kind of organisms from bats who perceive their environment by means of the faculty of echolocation.[29] All information about the physical processes in a bat's brain seems, on the contrary, at least in principle cognitively available to us. Claiming that consciousness is identical with a particular physical brain process amounts, Nagel thinks, to claiming that a state which is knowable essentially only from a single point of view, the point of view of the experiencing organism, is identical with a process which is in principle knowable by any intelligent being, as physical processes seem to be. Such an identity, however, seems utterly mysterious.

Nagel illustrates this sense of mystery by emphasizing that if consciousness is indeed identical with a physical process, we are as far from understanding what that identity amounts to as a pre-Socratic thinker would have been from understanding the claim that matter is energy, if someone were to tell him that.[30] It may be due to remarks of this sort that Nagel gets sometimes called a “mysterian” with respect to consciousness. Making progress in understanding the psycho-physical identity of consciousness and the brain may, according to Nagel, require devising a wholly new method of understanding our consciousness which he calls “objective phenomenology”.[31] Nagel's scepticism with respect to the prospects of materialist reductionism is thus attenuated and leaves space for possible discovery of a radically new method of investigation. Such a method, he believes, could help us make sense of the mind-brain identity by enabling us to think of the mental realm in a more objective manner. Nagel then does not rule out that consciousness could ultimately turn out to be

[29]Nagel (1974, p. 439).
[30]Nagel (1974, p. 447).
[31]Nagel (1974, p. 449).

physical but thinks that understanding the identity would require a revolution in our understanding of consciousness.

While Nagel's argument is admittedly controversial, his definition of consciousness has resonated among many. It is this definition of consciousness which I shall adopt in this volume, unless specified otherwise. Interestingly, defining consciousness in this way leaves it open as to how far down the phylogenetic tree consciousness goes. It opens up the possibility that consciousness could be a much more widespread phenomenon in the universe than, e.g., the linguistic ability, conceptual thought or higher levels of intelligence. It is, I believe, reasonable to attribute consciousness in this sense to at least many species of higher animals as the members of these species seem to feature subjective points of view. It is, however, difficult to find justification for this suggestion. The problem is that, as we saw, the criteria of application of Nagel's concept of consciousness have a first-person nature – the only subject about which I *directly* know that it is conscious, is myself. I may be able to reasonably judge that other people and animals are conscious too, but my knowledge that they are conscious is indirect – I infer that they are conscious from their behaviour, including linguistic behaviour. While this inference will be close to certain in the case of human beings and, arguably, of other higher mammals, the level of certainty will grow increasingly lower as we move down the phylogenetic tree (think of snakes, snails or flies). While many will doubt that some of these organisms are conscious in Nagel's sense, it is important to notice that nothing we know about them rules out that they could possess at least a primitive form of consciousness.

One sometimes hears the objection that Nagel's definition of consciousness does not capture the specifically human kind of consciousness which is, of course, complex and which arguably essentially involves some sort of self-consciousness or self-awareness as well as the capacity for conceptual thought. My reply here is that while it is plausible that human minds involve these features, to include them in a general definition of consciousness would mean that simpler animals without conceptual thought or self-awareness could not by definition be conscious. Nagel's definition is, in my view, best seen as expressing not a sufficient but rather a necessary condition for being conscious in the human sense, while expressing a sufficient condition for being conscious in a more general sense.

Nagel's definition of consciousness pertains to conscious *organisms* rather than to conscious *states* of these organisms. It can, however, be straightforwardly modified to characterize conscious states. We can say that a mental state is conscious iff there is something it is like to be in that mental state (for the organism who is in that state).[32] I do not want to presuppose here that there are unconscious mental states but if, as Freud and others believed, there are such states, e.g. unconscious desires or

[32]Throughout this volume I shall use "iff" to mean "if and only if".

beliefs, then there is nothing at all it is like to be in these states for the organism which is in them, even though these unconscious states may well exercise influence on our behaviour and on our other, conscious or unconscious mental states. Given that in the case of conscious states there is something it is like to have them or to be in them, we can say that conscious states feature instances of qualitative what-it's-likeness, qualitative feels or simply *qualia*[33]. For a mental state to feature a quale simply means that there is something it is like to be in that state for the given organism.[34]

This definition corresponds with David Chalmers's characterisation of qualia as "those properties of mental states that type those states by what it is like to have them"[35] It is also consistent with other mainstream characterisations of qualia. Michael Tye, for example characterizes qualia as "introspectively accessible, phenomenal aspects of our mental lives".[36] According to Ned Block, qualia include "the ways things look, sound and smell, the way it feels to have a pain, and more generally, what it's like to have experiential mental states" and can be defined as "experiential properties of sensations, feelings, perceptions and, more controversially, thoughts and desires as well".[37] In the literature, examples of qualia typically include the specific blueness we experience when we look at the sky or the specific sweetness we experience when we eat milk chocolate. The fact that these are examples of sensory states is no coincidence since, as suggested by Block, sensory states are paradigm qualia-involving states. That does not mean that other conscious states could not feature specific qualia. It seems to me that there is something it is like for me to do mathematical equations in my head or to think of my plans for the following day and these are states which are not normally viewed as sensory but rather as instances of thinking. While this seems to be the intuitive view I do not wish to take a stance at this point as to the question whether thought always or essentially involves qualia.[38]

Since I view conscious states as involving qualitative feels or qualia, these states, as I see it, amount to what Ned Block, who has introduced an influential classification of types of consciousness,[39] calls phenomenally-conscious or simply P-conscious states. For Block, P-conscious states are those mental states which have experiential properties where a totality of experiential properties constitutes what it is like to have that state.[40] Block contrasts the concept of P-consciousness with the concept of access consciousness or simply A-consciousness. A mental state is A-conscious, according to Block, roughly, if its content is available for use in reasoning and in rational control of

[33]"Qualia" is a plural form of the singular form "quale".
[34]The expression "quale" was allegedly first used by C. I. Lewis (1929).
[35]Chalmers (1996, p. 359).
[36]Tye (2009).
[37]Block (2004).
[38]See Strawson (2011) for an argument for the view that cognitive states are qualitative.
[39]See Block (1995).
[40]Block (1995, p. 230).

behaviour and speech.[41] Here Block works with the idea that many, if not all, mental states have content. That some mental states have some content is particularly clear in the case of beliefs and desires, sometimes called the "propositional attitudes". If I, for example, believe that it is raining, then my mental state can be characterized as a belief with the propositional content "it is raining".[42] The belief will count as A-conscious iff its propositional content can – in conjunction with the content of my other beliefs – lead me to believe that I will get wet unless I bring an umbrella or if its content can – in conjunction with my desire not to get wet – motivate me to bring an umbrella.

In many cases A-conscious states are also P-conscious. If, for example, I have a visual perception that it is raining outside, then this mental state is plausibly both A-conscious and P-conscious since there is typically something it is like to have a visual perception that it is raining outside (which means that the state is P-conscious) and, at the same time, the propositional content of the state ("it is raining outside") is available for me to use in my reasoning and rational control of my behaviour and speech (which means that the state is A-conscious). Block's concept of A-consciousness is also clearly distinct from the above-discussed behavioural concept of consciousness. A patient suffering from the locked-in syndrome, for example, may presumably have A-conscious states, at least as long as the patient is able to reason, without satisfying the third-personal, behavioural concept.

Phenomenal consciousness, our main concern here, also needs to be distinguished from what Block calls *self-consciousness* and *monitoring consciousness*. Both of these concepts, he suggests, most naturally apply to organisms rather than to mental states. He views self-consciousness as a possession of the concept of the self.[43] A sign that a given organism possesses this concept is, Block suggests, that it is able to recognize itself in the mirror. According to Block, this ability is possessed by higher primates, such as chimps, but not, for example, by monkeys, dogs or human babies younger than 18 months.[44] I agree with Block that self-consciousness needs to be distinguished from P-consciousness and that, moreover, an organism can have P-conscious states without being self-conscious – here, dogs, rabbits and other animals come to mind.

An organism has monitoring consciousness, according to Block, iff it exhibits some sort of internal monitoring of its conscious states.[45] This monitoring can take a form of thoughts about one's own conscious states or some sort of inner perception or scanning. Importantly, this concept of consciousness should, once again, be kept separate from that of P-consciousness. Arguably, after all,

[41]Block (1995, p. 231).
[42]It may be that some mental states, such as sensations, have further, non-propositional content in addition to their propositional content.
[43]Block (1995, p. 235).
[44]There are experiments which give us reason to think that chimps (between 7 and 15 years of age) possess self-consciousness. When these chimps, who while anaesthetised had painted coloured spots on their foreheads and ears, wake up and look in the mirror, they will usually wipe the spots off. See Block (1995, p. 235).
[45]Block (1995, p. 235).

even, e.g., a lap-top computer has some sort of primitive internal scanning (thanks to which it knows, e.g., when to start its cooling mechanism) without being P-conscious.

It has not been my aim in this section to discuss all the possible distinctions when it comes to types of consciousness but rather to elucidate the type of consciousness which I think is the most relevant to this whole volume. The type of consciousness in question is one of phenomenal consciousness as it has been defined by Nagel. In this volume I shall take phenomenal consciousness to be a property of certain mental states as well as a property of some organisms (or perhaps even some non-organic entities).[46] For the sake of brevity, I shall often use the term "consciousness" to mean phenomenal (or P-) consciousness.

A few further terminological remarks: I will take it that if an organism is phenomenally conscious, it experiences conscious states which are (partly) qualitative or, in other words, have qualia. I take it that the sum of all the qualia a given organism has at a given moment can be called the "phenomenology" which the organism has at the given moment.[47] I shall, however, also sometimes speak of more specific kinds of phenomenology, for example "visual phenomenology" or "audible phenomenology", by which I shall mean the typical kinds of qualia associated with the given sensory modality.

3. Sellars's Proposal and the Problem of Consciousness

Let us now return to the above-posed question of whether Sellars's attempt to construct a stereoscopic view of conscious sensations succeeds. We have seen that a stereoscopic view of sensations must accept that both the manifest image and the scientific image of sensations are correct in the sense that they reveal to us real features of sensations. As we saw, one obstacle on the road to a stereoscopic view of sensations has arguably been overcome by Sellars when he suggested that the ultimate homogeneity of sensations, integral to the manifest image, is compatible with the scientific view given that we reject the 'particulate' view of nature in favour of the continuity view. There is, however, another, and I think more serious – challenge to the stereoscopic view, namely the existence of consciousness in the sense defined in the previous section. Let me now clarify why consciousness can be seen as problem for Sellars's solution.

The problem, as I see it, can be expressed in the form of the following argument:

[46]By calling an organism conscious I do not mean to imply that it is phenomenally conscious all the time. My use certainly allows for periods of dreamless sleep when the organism has no phenomenally conscious states.
[47]I am aware that this way of using the term "phenomenology" is fairly uncommon in the continental philosophical tradition and that, moreover, it is somewhat etymologically misleading. I will, nevertheless, hold on to it in this volume as it is very common in analytical philosophy of mind.

1. At least some sensations are conscious.

2. Conscious sensations involve phenomenology.

3. Neural processes do not involve phenomenology

_____

4. Conscious sensations are not neural processes.

Here premise (1) is justified by reflection on the manifest image which reveals to us that most, perhaps all, of our sensations are conscious. I take it that (1) is likely to be accepted by most materialists as well as by their critics. Premise (2) is based on the intuitive claim that if a subject has a conscious sensation, there is something it is like for her to have this sensation which, given the terminological remarks made at the end of the last section, just means that sensations involve phenomenology. Some philosophers deny that consciousness involves phenomenology, one example here being Daniel Dennett who views the notion of phenomenology and qualia as suspect and misleading. I shall engage with Dennett's critique of the notion of qualia in the next chapter. Still, I think there is a very strong intuitive case in favour of premise (2).

Perhaps the most controversial premise of this argument is (3). It may seem simply to be common sense that neural processes do not involve phenomenology, but it is denied by many materialists who often hold that conscious phenomenology is a physical, perhaps neurophysiological process. Moreover, as we saw, even Nagel himself, does not accept (3), leaving instead open the possibility that conscious phenomenology could be identical to a physical process, although our current ways of thinking about the world do not allow us to understand how this identity could be true.

I shall argue in much detail for a thesis close to (3) in chapters 2 to 4. For now, let me just introduce a general consideration in favour of this premise. If we consider what we currently know about all the neural processes in our brains, we will see no immediate reason to think that any of these processes involve phenomenology. The kinds of facts which neurobiology tells us about the brain are, roughly, facts about different types of neurons and their intricately structured causal interactions. These facts, however, give us no immediate reason to think that the brain involves or gives rise to phenomenology. If the brain indeed involves phenomenology, it seems to be a further interesting fact about the brain, a fact distinct from all the neurobiological facts. This apparent distinctness of consciousness seems to make the phenomenon strangely immune to the standard methods of cognitive science which are used to explain mental capacities such as learning or appropriately reacting to stimuli. Standardly, these capacities are defined as functions and thereafter the neural mechanism responsible for their implementation is searched for. It is, however, far from

clear that consciousness with its phenomenology is definable in purely functional terms. It is this peculiar immunity of consciousness to standard methods of cognitive science which has led Chalmers to call consciousness the "hard problem" and classify mental capacities – which he views as susceptible to the standard methods of cognitive science – as "easy problems".[48]

Given that, as we saw, Sellars rejects the conclusion of this argument, as expressed by thesis (4) and that the argument looks valid, it is interesting to ask whether his position provides him with means to reject any of the premises. It seems to me that Sellars would view premises (1) and (2) as an integral part of the manifest image and, as a result, he would likely wish to reject premise (3). Such a step at least seems to be suggested by his general approach. How, however, could that be done? Surely, as we saw, he has more resources which could provide him with a reason to deny (4) than a mainstream materialist who, roughly, thinks that the micro-physical entities only have the properties which current physics informs us about, or properties which are very similar to these.[49] Unlike the mainstream materialists, Sellars thinks that the fundamental level of the physical world features ultimate homogeneity and sensory qualities.

Could not Sellars then simply claim that neural states, being physical but also having a sensory character, feature phenomenology and premise (3) should therefore be rejected? In view of Nagel's considerations about consciousness introduced above, such a proposal may seem unpromising as, surely, these neural states are in Nagel's sense perfectly objective, while consciousness is essentially subjective. Still, perhaps this proposal holds some promise, as I shall try to show in chapter 6 where I deal with some positions along similar lines. However, this may be, it has not been my aim to decide here whether Sellars's proposal can accommodate the challenge which I suggested consciousness brings about. Instead I only meant to suggest that consciousness is an important challenge for materialism as well as for any attempt to construct a stereoscopic view envisioned by Sellars.

In the following chapters I shall argue in much more detail for this last point, suggesting that the existence of consciousness indeed provides us with a reason to doubt that the universe is purely physical, at least if "physical" is understood in the mainstream way.[50] In the second part of the volume I shall explore the options open for those who try to account for consciousness in non-reductive ways.

[48]See Chalmers (1995).
[49]I shall say much more about my understanding of physicalism in the next chapter.
[50]Which is arguably not how Sellars understood it.

2. A Priori Physicalism

1. Two Varieties of Physicalism

In the previous chapter I tried to show why the phenomenal character of conscious states gives us at least a *prima facie* reason to doubt that mental states could be physical and why it can therefore be seen as an obstacle for the project of materialist reduction. Certainly, nothing we learn from physics, chemistry or biology suggests that some purely physical state is such that there is something it is like for an organism to be in it. Consider the intricate structure of neuronal firing in your brain, incorporating into your conception the sophisticated chemical and electrical processes in your brain, still, none of that seems to give you any reason to think that any state of your brain is phenomenally conscious. Think, for example, of the richly qualitative conscious state which one is in at a particular moment while eating vanilla ice-cream on a sunny day in May while walking down Champs-Élysées, having rich visual, auditory, olfactory and gustatory perceptions. Why think, *prima facie*, that any purely physical brain state could constitute, or produce that conscious state?

Despite questions of this sort, the prevalent view of consciousness in philosophy and cognitive science of the last roughly 60 years has been that some form of physicalism surely must be true. I take physicalism to be the view that consciousness is fundamentally physical, which means that truths about consciousness are wholly grounded in fundamental truths of completed physics.[51] We can say that truths about domain *A* ground truths about domain *B* iff truths about *B* hold in virtue of truths about *A* holding. Moreover, we can say that truths about domain *A* wholly ground truths about domain *B* iff truths about *B* hold solely in virtue of truths about *A* holding.[52] Physicalism then tells us, roughly, that fundamental truths of completed physics fix the truths about consciousness, or that phenomenal facts (facts about phenomenal consciousness) are ultimately physical facts. Using a theological metaphor, we can say that physicalism is true iff once God created all fundamental physical facts, he had no more work to do, as consciousness was brought into existence by that very creative act.[53]

A definition of this kind, which appeals to truths of completed physics, may seem to render the truth of physicalism difficult to assess at the present moment. We, after all, do not currently possess – and perhaps are quite far away from possessing – all the fundamental truths of completed physics – indeed there is no reason to think that our physical theory will not undergo significant, perhaps revolutionary, developments in the future. Should we then not replace the appeal to completed

[51]In this volume I use the terms "physicalism" and "materialism" and their derivatives interchangeably.
[52]See also Chalmers (2015, p. 248).
[53]Strictly speaking, he may have had more work to do with respect to some other properties if these are non-physical. Most people in the consciousness debate, however, assume that if there were anything non-physical then it could only be consciousness.

physics with an appeal to current physics?

To do that would be to misunderstand the spirit of physicalism. Notice, after all, that if the definition of physicalism appealed to truths of current physics instead of truths of completed physics, then new physical truths which future physics will no doubt discover, could by definition not ground truths about consciousness. There is, however, no reason why the physicalists should hold that. Clearly, physicalism is compatible with new physical discoveries being relevant with respect to the occurrence and nature of consciousness, although of course this appeal to future discoveries is not compulsory. Is then the appeal to truths of completed physics in the definition of physicalism feasible after all? I think so, as by adopting such a definition one is embracing the spirit of physicalism as a doctrine compatible with the discovery of new physical truths. Moreover, by adopting such a definition, the anti-physicalists make their job harder as the physicalists are then free to appeal to possible future discoveries in physics as explanatorily relevant with respect to consciousness. Moreover, we shall see that the dominant view among the proponents of materialist reduction of consciousness is that we already know enough to be justified in thinking that consciousness is physical and that we need not rely on future developments of physical theories when it comes to explaining consciousness.[54]

It is not my aim here to give a historical overview of the materialist attempts to reduce consciousness.[55] Instead, I shall merely introduce the two main branches of contemporary physicalism – a priori physicalism and a posteriori physicalism. The difference between these two kinds of physicalism roots from two quite different ways in which proponents of physicalism respond to the most influential anti-physicalist arguments: the knowledge argument, introduced by Frank Jackson, and the conceivability argument, discussed, among others, by Saul Kripke and David Chalmers.[56] It will therefore be useful to first take at least a cursory look at these arguments.

Despite their significant differences, both arguments draw their ontological conclusion of the falsity of materialism from epistemic premises, i.e. from premises concerning the logical relation between truths of completed physics, or simply physical truths, and truths about consciousness, or simply phenomenal truths, motivating us to embrace these epistemic premises by considering hypothetical scenarios.

The knowledge argument appeals to the hypothetical scenario of Mary, a super-intelligent and super-knowledgable colour-scientist, possessing complete physical knowledge of human colour-vision.[57] Mary then knows, unlike our best colour-scientists, everything physical there is to know

[54]That no reliance on developments is needed is clearly expressed, for example, in Dennett (2005, pp. 9–10).
[55]See Braddon-Mitchell – Jackson (2007) for a good overview.
[56]See e.g. Jackson (1982), Kripke (1980) and Chalmers (2010).
[57]This version of the scenario is introduced, for example, in Jackson (2003) or Jackson (1982).

about how visual information which enters our bodies via our eyes is processed by our brain as well as how the relevant brain processes lead to our sophisticated reactions (or lack of reaction) to the information. At the same time, Mary has spent her whole life in a black-and-white environment – living in a black-and-white room, connected with the outside world merely via black-and-white monitors. She has gained her extraordinary physical knowledge by reading black-and-white books and working with computers with black-and-white monitors. Also, all the food she has eaten, her clothes and her own body are also dyed black or white. As a result, Mary has never seen, for example, the colour red.

Confronted with this scenario, we are invited by Jackson to consider what happens on the day Mary is let out of her black-and-white room and is presented with a ripe strawberry. Will she, supposing that her colour vision works fine despite the lack of relevant impulses, learn anything new about the world or not? Jackson suggests that we should accept the intuitive conclusion that she does – she learns what it is like to see red and thus, we can say, gains phenomenal knowledge, i.e. knowledge of what it is like to be in particular phenomenally conscious states, in this case in a phenomenally red conscious state.

It seems then that upon leaving the black-and-white room, Mary gains new phenomenal knowledge, despite already possessing complete physical knowledge of colour-vision and having perfect reasoning skills as well as sufficient time for reflection. The claim that Mary gains new phenomenal knowledge, however, implies the epistemic thesis that complete physical knowledge does not include, nor does it a priori entail, even on ideal reflection, phenomenal knowledge (e.g. knowledge of what it is like to see red). If that is true, then there is an *epistemic gap* between physical truths and phenomenal truths.[58] This means that physical truths are somehow principally disconnected from phenomenal truths so that they neither imply phenomenal truths, even on ideal reflection, nor are they implied by them. The fact that physical knowledge is not complete knowledge and nor does it a priori entail complete knowledge should, however, if the knowledge argument is plausible, lead us to the rejection of physicalism.[59] We can say that according to the proponents of the knowledge argument, the existence of the epistemic gap between phenomenal and physical truths leads to an ontological gap between the physical world and phenomenal consciousness.

The other influential argument against physicalism, the conceivability argument, encourages us to conceive of a scenario in which all physical truths about the world, call their complete set P, obtain without an arbitrary phenomenal truth about our world, call it Q, obtaining.[60] Here Q can be, for example, the truth that John is phenomenally conscious or that John is currently experiencing the

[58]Chalmers (2010, 109).
[59]See Jackson (1982, p. 130).
[60]See Chalmers (2010, p. 141–205) for a detailed discussion of the argument.

colour red. The claim that such a scenario is conceivable leads, according to the proponents of the conceivability argument, to the claim that such a scenario is really possible.[61] This possibility, however, seems to be incompatible with the truth of physicalism: if, after all, it is possible that all physical truths obtain without an arbitrary phenomenal truth obtaining, how could the fact described by the phenomenal truth be nothing but some fact described by physical truths?

The conceivability argument then has the following general structure:

1. *P*&~*Q* is conceivable.

2. If *P*&~*Q* is conceivable, *P*&~*Q* is possible.

3. If *P*&~*Q* is possible, physicalism is false.

———

4. Physicalism is false.

One can make the conceivability argument more vivid by appealing to philosophical zombies. Here philosophical zombies are hypothetical creatures, which are physically and functionally just like ourselves (they are, so to say, our physical and functional replicas), but which, nevertheless, lack consciousness.[62] The argument will then have the following structure:

1. Zombies are conceivable.

2. If zombies are conceivable, zombies are possible.

3. If zombies are possible, physicalism is false.

———

4. Physicalism is false.

It is easy to see that the knowledge argument and the conceivability argument have a similar structure: they both use epistemic considerations as their starting point and have the metaphysical thesis of the falsity of physicalism as their conclusion. In the case of the knowledge argument the relevant epistemic thesis states the claim that there is an epistemic gap between physical and phenomenal truths. It is easy to see that the same epistemic gap is also expressed by premise (1) of the conceivability argument. If, after all, zombies are to be truly conceivable, then, at the very least, phenomenal truths cannot be a priori entailed by the complete set of physical truths. Both arguments then in effect tell us that (a) there is an epistemic gap, and that (b) the epistemic gap implies a

[61]Strictly speaking, the possibility in question is metaphysical possibility which is compatible with physical impossibility. While I shall disregard this complication at the moment, I shall return to it in the next chapter.
[62]See Chalmers (1996, pp. 94–95) for more on zombies.

metaphysical gap, i.e. the claim that phenomenal consciousness and the physical world are two metaphysically distinct domains.

Having identified the general structure of these arguments enables us to distinguish between the two main streams of current physicalism. Physicalists of the first kind simply reject that there is an epistemic gap while physicalists of the second kind accept that there is this gap but reject that the epistemic gap implies the existence of the metaphysical gap. The former strategy is adopted by those physicalists who hold that phenomenal truths are, at least in principle, deducible from, or a priori entailed by physical truths.[63] This branch of physicalism is therefore commonly called a priori physicalism in the literature. Clearly, if a priori physicalism is true, zombies are not even conceivable since the idea of my physical and functional replica without consciousness is contradictory. Similarly, according to a priori physicalism, if Mary knew all the physical truths about colour-vision, she would in principle be able to deduce from this body of physical knowledge all phenomenal truths about consciousness. As a result, she would not – despite our intuitions – learn anything new upon leaving her black-and-white room.

A different strategy is adopted by those physicalists who agree with the anti-physicalists that there is an epistemic gap but deny that the epistemic gap entails an ontological gap. In the key of conceivability, these physicalists hold that although *P&~Q* is conceivable, it is not the case that *P&~Q* is possible. There is thus, according to these philosophers, no a priori entailment between physical truths and phenomenal truths and, as a result, even if we knew all physical truths about the universe, we would need further information in order to also know the phenomenal truths about the universe. This version of physicalism is often called a posteriori physicalism in the literature. I shall discuss a posteriori physicalism in the following two chapters and focus on a priori physicalism in this chapter.

According to a priori physicalism, there is no epistemic gap between physical and phenomenal truths. The most serious intuitive obstacle to a priori physicalism is presumably the fact that we seem to possess concepts of conscious states which are not analysable in causal or physical terms, and, as a result, we seem to know truths about these conscious states which are not a priori entailed by any set of physical truths. Plausibly, for example, most of us, unlike Mary before she is released, possess a concept of experience of colour red, by means of which we grasp the phenomenal quality, or the quale of the experience which we undergo when we look at, for example, a ripe strawberry (supposing we are normal subjects). This phenomenal concept can be used in propositions, for example, in the proposition that our experience of red is very different from our experience of green. The phenomenal concepts involved in propositions of this kind do not seem to be analysable

[63]See e.g. Lewis (1966), Armstrong (1968).

in physical or causal terms – indeed the former seem utterly conceptually independent of the latter. This apparent conceptual independence of phenomenal concepts with respect to physical concepts explains why it seems to us that zombies cannot be a priori ruled out even if they are in fact not physically possible or why it seems plausible to us that even complete micro-physical knowledge of colour vision will not entail phenomenal knowledge, and perhaps even why we do not seem to be able to rule out the existence of disembodied conscious entities a priori.[64]

Given that this is the intuitive view, the task for a priori physicalists will be to show that there are in fact no phenomenal truths involving these kinds of unanalysable phenomenal concepts. The a priori physicalist will hold that even though it seems to us that certain truths about consciousness are unanalysable, this is in fact a false impression. How could truths about consciousness be analysable? Here perhaps the most promising proposal is analytic functionalism. According to analytic functionalism, sufficient reflection of our concepts of consciousness will reveal to us that truths about consciousness are really truths about functional states. Analytic functionalism is a version of the more general doctrine of functionalism. Let me therefore first say a bit about this more general view.

2. Functionalism

What is functionalism and why is it important for the materialists? According to Daniel Dennett, who has presented some of the most interesting functionalist proposals, the main idea behind functionalism can be expressed by the saying "handsome is as handsome does".[65] Functionalists hold that when it comes to explaining consciousness and other mental features, it is important to pay attention to what the brain does rather than what it, intrinsically, is. More precisely, it is important to pay attention to the functions implemented or realised by the brain rather than to its physical and chemical composition and nature. One can argue that the essence of functionalism is deeply ingrained in our thinking and language. In fact, many of our concepts have functional conditions of satisfaction – we classify something as a heart, an umbrella or food largely because of what the given thing does, i.e. which functions it is able to implement.

Despite these more general roots, the key inspiration behind functionalism with respect to the mental realm is the computer.[66] What makes computers unique and special is not what they are when it comes to their material composition but rather what they are able to do. Here computers are in fact, as Dennett emphasizes, quite mindlike. Dennett writes, for example:

[64]See Goff (2010).
[65]Dennett (2005, p. 17).
[66]Dennett (2005, p. 6).

> *Computers are mindlike in ways that no earlier artifacts were: they can control processes that perform tasks that call for discrimination, inference, memory, judgment, anticipation; they are generators of new knowledge, finders of patterns—in poetry, astronomy, and mathematics, for instance—that heretofore only human beings could even hope to find.*[67]

Despite the fact that computers are endowed with these extraordinary capacities or functions, their material constitution, what their hardware is made out of, is in fact relatively uninteresting. What makes computers interesting is their functional complexity. Similarly, functionalism in philosophy and science is the view that what renders a particular state mental is not its special material constitution but rather its role in a complex system of inputs, outputs and other functional states. Consider, for example, the belief that it is raining. According to the functionalist, this mental state is, very roughly, the state that is typically brought about by the perception that it is raining and the state which typically brings about the desire not to get wet and perhaps (in conjunction with particular beliefs about what umbrellas can do), the action of opening an umbrella when leaving the house. The given belief is then defined by the functionalists purely relationally, in terms of its function in the complex economy of perceptual inputs, other mental states (which are also defined relationally) and behavioural (including verbal) outputs.

This sort of relational nature of beliefs and mental states in general brings us to another interesting feature of functionalism. Just like, say, a chess computer programme can be run on a Mackintosh or an IBM computer, a given mental state, such as the above-mentioned belief that it is raining, could, according to functionalists, in principle be realised in a brain of a quite different physical composition, and perhaps even in an extremely complex future computer, as long as the realising system featured the required functional complexity. This is usually expressed by the functionalists in the form of the claim that functional states, including mental states, are *multiply realisable*.

What does functionalism have to say about consciousness? The difference between a non-conscious mental state and its conscious counterpart will, perhaps unsurprisingly, be a matter of the state's effects in a system of functional states, inputs and outputs which constitutes the mind, according to functionalism. Conscious perceptions, for example, will typically have rather rich effects, such as verbal reports and formation of corresponding beliefs and desires. All or most of these effects will, on the other hand, be absent in the case of non-conscious mental states. Dennett, for example, talks about the echo-making power of conscious mental states thanks to which these states are elevated to the status which he metaphorically calls “fame in the brain”. [68] We can see then that the fact that a particular mental state is conscious is, for the functionalists, a matter of what the state does, i.e. how

[67]Dennett (2005, ibid.).
[68]Dennett (2005, p. 165).

it affects our behaviour and other mental states rather than what it, in and of itself, is.

It is, I believe, reasonable to view Dennett, as some have suggested, as a proponent of a version of a priori physicalism, namely of analytic functionalism. If the truths about consciousness are truths about particular kinds of functional states realised by the brain in conjunction with truths about particular environmental inputs and behavioural outputs, it seems that truths about consciousness are a priori entailed by physical truths. We can see this if we, for example, suppose that we have a complete physical description of the workings of a clock. It seems that it will be fairly easy to deduce truths about the functional organisation of the clock's components from this description as the functional truths simply concern what each component does within the system. The analytic functionalist holds that the same is, at least in principle, true about the brain. That Dennett is a proponent of analytic functionalism is apparent from his insistence that zombies are upon sufficient reflection inconceivable.[69] Surely he would allow that they are at least negatively conceivable (i.e. their existence is not a priori ruled out) if he did not hold that our concepts of conscious states are ultimately analysable in functional terms.

Functionalism need not be analytic. Non-analytic functionalism, just like analytic functionalism, holds that consciousness can be accounted for in functional terms but, unlike analytic functionalism, allows that there are truths about consciousness which are not functional.[70] These truths involve phenomenal concepts by means of which we are, according to non-analytic functionalists, able to think about our conscious states in terms of their phenomenal character. Non-analytic functionalism is a version of a posteriori physicalism, a view I shall say much more about in the following two chapters.

How does analytic functionalism (from now on simply “functionalism”) cope when it comes to explaining consciousness? It will presumably be fairly easy for the functionalists to account for the property of A-consciousness, introduced by Block and characterized in the previous chapter.[71] There we saw that a mental state is A-conscious if its content is available for use in reasoning and in rational control of behaviour and speech. Clearly this concept of consciousness is relational and it is even natural to view it as functional.[72] Similarly, Chalmers describes the state of *awareness* which we are in whenever we have access to some information which we can use in the process of controlling our behaviour.[73] Once again, we are dealing with a functional notion and the state of awareness, as understood by Chalmers, will plausibly be susceptible to functionalist treatment.

[69]See e.g. Dennett (2005, p. 15).
[70]I understand truths here as, roughly, true Fregean propositions. Here, for example, the truth that John has met Superman and the truth that John has met Clark Kent are two different truths, although truths describing the same fact, i.e. the fact of John meeting the actual person known by some as Clark Kent and by others as Superman.
[71]Block (1995, p. 231).
[72]Block (1995, p. 232).
[73]Chalmers (1996, p. 28).

It is much less clear that the property of phenomenal consciousness is susceptible to a functionalist explanation. It certainly seems that possessing that property is not (merely) a matter of having states with particular functional links in one's brain. In other words, it is far from clear, unlike in the case of many other properties of organisms, such as photosynthesis or digestion that the property of phenomenal consciousness is functionalisable. One attempt to demonstrate that phenomenal consciousness isn't functionalisable is, of course, the zombie thought experiment. A zombie, after all, is supposed to be physically and functionally just like me but without consciousness. Certainly, if zombies are possible, the functionalist approach to consciousness is under threat as it will then be unable to show us why we are phenomenally conscious, given that we are functional twins of zombies. Another way in which this worry can be expressed is to say that functionalism offers us no solution to Chalmers's hard problem: i.e. to the question why some of the various functions implemented in the brains are accompanied by phenomenology.[74]

Despite the intuitive power of this worry, the functionalists have attempted to show that it is in fact misguided. In what follows I shall try to explore and evaluate perhaps the most interesting and systematic proposal along these lines, which has been introduced by Dennett.

3. Heterophenomenology

Dennett finds the supposed "inner" domain of qualia, phenomenology or phenomenal character, which the critics of functionalism like to appeal to, highly suspicious and has dedicated much of his philosophical work to arguing that philosophers' appeals to it are fraught with controversy. He in fact calls his position with respect to qualia *eliminative materialism*; in his view then notions of qualia or phenomenal character are misleading and confused and should be eliminated or removed from philosophical and scientific theories of the mind.[75] Only if we dispose of the notion qualia, shall we be, according to Dennett, able to advance in our pursuit of a scientific explanation of consciousness. The belief that qualia or phenomenal character exist is then, for Dennett, not merely a philosophical error but also a crucial obstacle in the way of the scientific study of the relation between consciousness and the brain, as it leads to a wrong interpretation of empirical data concerning neural processes in the brain. Qualia, Dennett thinks, need to be eradicated in the name of scientific progress just as we once eradicated the appeals to witches, *élan vital* or phlogiston from our scientific theories. Dennett writes, for example:

> *[i]t is not enough to withhold our theoretical allegiances until the sunny day when philosophers complete the tricky task of purifying the everyday concept of qualia. Unless we*

[74]Chalmers (1995).
[75]Dennett (2002, p. 244, fn. 2).

> *take active steps to shed this source concept, and replace it with better ideas, it will continue to cripple our imaginations and systematically distort our attempts to understand the phenomena already encountered.*[76]

In this passage, Dennett is critical of the view that philosophers and scientists working on consciousness need to account for the existence of phenomenal character or qualia in the physical world.[77] Since, according to Dennett, qualia do not exist – and I shall discuss his arguments for this radical view shortly –, such a project would not make sense. What then is the form which the science of consciousness should, according to Dennett, take?

According to the method of the study of consciousness suggested by Dennett and called by him *heterophenomenology*, the data for the theory of consciousness are not my supposed inner, first-person states, but other people's reports about their conscious states.[78] These are arrived at by gathering recorded verbal reports of various subjects concerning their own conscious states. These recordings with their acoustic properties will then be the raw data of heterophenomenology out of which we can get the true data of heterophenomenology if we adopt what Dennett calls the *intentional stance* with respect to the raw data.[79] This interpretive stance amounts to, roughly, understanding these reports as expressions of beliefs, desires and intentions of rational subjects. Once we do that, we will see that the reports form a sort of folk theory of the mind and consciousness. The heterophenomenologist is then quite close to the anthropologist who records and reconstructs the folk-tales, or folk-theories of the world.

Adopting the intentional stance towards this data does not mean accepting the resulting reports of conscious states and goings on – the theorems of the folk theory – as true. Indeed, according to Dennett, the reports need to be bracketed for neutrality which means neither their truth nor their falsity can be presupposed on the outset.[80] Having bracketed the reports in this way, we need to ask what will be the standard by which we can decide which reports are true and which are false. Here the standard for Dennett are the scientific discoveries concerning the functions implemented in the brain. Put simply, only those reports will be taken to be true which report states and goings on whose existence is scientifically confirmed.

Clearly then, according to Dennett, it is not a task of the science of consciousness to explain the states and goings on reported by the experimental subjects – many, perhaps most of these will, after all, be viewed as purely fictional. Instead the science of consciousness should explain precisely the

[76]Dennett (2002, p. 238).
[77]For a proposal along these lines see e.g. Chalmers (2010, pp. 37–58).
[78]Dennett (2005, p. 38).
[79]Dennett (2005, p. 37).
[80]Dennett (2005, p. 39).

reports and the beliefs expressed by them.[81] At this point we should be able to see why Dennett thinks consciousness is not a uniquely difficult problem: if the data which the theory of consciousness needs to explain consists of verbal reports of conscious states and processes, all that the theory needs to explain is how and why a given organism produces these verbal reports. While we may still be far away from a definitive explanation of such reports, we can already see what the general shape of such an explanation would be. It would presumably consist in identifying the neural mechanism responsible for triggering the given report. Of course, Dennett thinks that these reports should be viewed by the heterophenomenologists as expressions of beliefs, desires and other propositional attitudes and thinks that heterophenomenology also needs to explain these, but these propositional attitudes, according to functionalism, correspond to particular functional states. While we may still be far away from a definitive explanation of such propositional attitudes, we can already see that their explanation would need to clarify what neural states are responsible for realising the functional states which, according to functionalism, are these propositional attitudes. For Dennett then the science of consciousness basically needs to explain how the brain realises particular capacities or functions.

It is a subject of much controversy whether heterophenomenology is an adequate method for the science of consciousness.[82] Many philosophers have objected against it that the method leaves out exactly what makes the problem of consciousness particularly interesting, as well as particularly difficult, namely one's own phenomenology. While heterophenomenology works with third person data, a complete theory of consciousness should, according to the objectors, also explain the first-person data – your or mine conscious states, as they appear to you or me, i.e. with their particular qualia or phenomenal character. Here, however, Dennett offers an interesting reply to his opponents, arguing that the existence of properties such as qualia or phenomenal character is highly dubious.[83]

4. Direct Apprehension of Qualia

Dennett attempts to cast doubt on the existence of qualia by trying to show that there are no properties of conscious experience which possesses the second-order properties usually attributed to qualia. According to Dennett, philosophers normally think of qualia as ineffable, private, intrinsic and directly apprehensible. If, however, we reflect on the properties of our experience, we shall find, Dennett argues, no such properties, i.e. we shall find no qualia. Here, one could reply to Dennett that proponents of qualia need not define them in terms of these four second-order

[81]Dennett (2005, p. 38–39).
[82]See Chalmers (2010, pp. 54–58) for a sceptical take on heterophenomenology. For a current fairly positive take, see e.g. Hříbek (2016).
[83]Dennett (2002).

properties. The definitions of qualia presented in the previous chapter, for example, do not appeal to these second-order properties. Still, intuitively, it seems that qualia should have these attributes or at least some attributes in the vicinity of these. However it may be, in what follows I shall presume that qualia need to possess these four attributes.

I shall first focus here on Dennett's critique of the view that there are properties of experience which we directly apprehend or are intimately acquainted with. This direct apprehension thesis is, Dennett argues, illusory as we are unable to even reliably establish whether or not our qualia, supposing that they exist, have changed over the course of time.

Dennett demonstrates this point using a scenario with two coffee tasters working for Maxwell House whose job is to ensure that Maxwell House coffee keeps a constant taste over time.[84] One day after six years of working for Maxwell House the two tasters get together and in the course of conversation it turns out that they both agree that when they started working for the company six years ago, they both liked the taste of the coffee, but that they both no longer like it. Each of them, however, explains this change in a different way. One of them, Mr. Chase, claims that Maxwell House coffee still tastes the same to him but he has become a much more demanding and sophisticated coffee drinker so he doesn't value the taste of Maxwell House as highly as before. Dennett characterizes Chase's description of the change he went through as the claim that the relevant taste qualia, i.e. the way coffee tastes to him, have not changed over time while his aesthetic judgments about the qualia, i.e. certain kind of reactions to these qualia, have shifted as he no longer enjoys the taste.[85]

Mr. Sanborn, the other taster of Maxwell House coffee also no longer likes the taste of Maxwell House which he once enjoyed. He, however, thinks that the reason for the change has something to do with a change in his taste apparatus. Sanborn claims that his taste buds, or perhaps the brain processes analysing the signals from the taste buds, have somehow changed or deteriorated which is why Maxwell House coffee now tastes different to him. The original taste of the coffee, before the bodily change, is, however, still valued as highly as it once was by Sanborn. According to Dennett, Sanborn is basically telling us that his taste qualia have changed because of certain physiological changes, while his judgments about the qualia, i.e. his reactions to these qualia, have stayed constant.[86]

Having collected these two hypothetical reports, Dennett suggests – distinctly in the spirit of heterophenomenology – that we cannot take it for granted that these reports are true.[87] We need to

[84]Dennett (2002, pp. 231–232).
[85]Dennett (2002, p. 232).
[86]Dennett (2002, ibid.).
[87]Dennett (2002, ibid.).

keep scientific neutrality and accept that people's reports about consciousness and qualia could be wrong, i.e. are not infallible. Moreover, as Dennett emphasizes, almost all participants in the qualia debate, including the proponents of qualia, would agree that it would be naïve to take the reports offered by Chase and Sanborn as certainly true.[88] If we approach their reports with scientific neutrality, we shall not be able to rule out that, despite believing the opposite, Chase may have had the same happen to him as Sanborn, i.e. his qualia may have changed while his aesthetic judgments may have not. Similarly, we cannot quite rule out that Sanborn is in fact in the situation described by Chase, or – for that matter – that both coffee tasters have undergone a combination of a change to their qualia and a shift of their evaluative judgments. Their beliefs about their situation will be, as Dennett puts it, just as fallible as, for example, our subjective judgments about the constancy or changes of temperature or lighting intensity in a room one is in.[89]

If, as most will agree, we cannot fully rely on the subjects' description of the relevant change, one could wonder whether there is a way to empirically, objectively verify these reports. One way to do this would be to identify a particular kind of brain property with qualia and to view a particular neural process as a realiser of the relevant judgments about qualia. The results of such empirical tests would, however, as Dennett explains, be of merely limited evidential value when it comes to deciding whether Chase's and Sanborn's descriptions of what happened are true.[90] These empirical tests are, however, as Dennett explains, still more useful than the relevant introspective judgments about qualia when it comes to checking the reliability of the subjects' reports.

Of course, once we endorse empirical testability of judgments about qualia, we approach the claims of the given subject about his or her qualia as merely fallible and corrigible *hypotheses*. If, however, judgments about qualia are corrigible in the same way in which we normally take subjective judgments about regular natural properties (think one's bodily temperature or weight of objects) to be corrigible and fallible, the claim of our direct acquaintance with qualia (or intimate direct knowledge thereof) is, according to Dennett, seriously undermined. This, of course, casts doubt on the concept of qualia introduced above, which, according to Dennett, reveals to us qualia as properties with which we are directly acquainted. As Dennett puts it, qualia

> *[…] far from being directly or immediately apprehensible properties of our experience, [...] are properties whose changes or constancies are either entirely beyond our ken, or inferable (at best) from "third-person" examinations of our behavioral and physiological reaction patterns […].*[91]

[88]Dennett (2002, p. 233).
[89]Dennett (2002, p. 236).
[90]Dennett (2002, p. 235).
[91]Dennett (2002, p. 236).

As we saw, the proponents of qualia do not usually endorse strong infallibilism about qualia, i.e. roughly the thesis that we simply cannot be wrong in our introspective judgments about qualia, but, according to Dennett, they underestimate the extent and implications of our fallibility about qualia. Namely, they fail to see, according to Dennett, that fallibility and empirical corrigibility disrupts the very concept of qualia.

Let me now sketch a way in which proponents of qualia could reply to Dennett's critique. They could emphasize that it is no coincidence that in the current debate about qualia almost no one holds that we possess knowledge of qualia which is in every way infallible or incorrigible. In the definitions of qualia discussed in the previous chapter, the infallibility thesis certainly does not appear and Block, for example, openly allows that some of our beliefs about qualia could be false.[92] The absence of incorrigibility talk from many of the discussions of qualia is, as I see it, no accident and should not be viewed as a mere *ad hoc* reaction to Dennett's critique.

The incorrigibility talk is absent, as I see it, simply because the reasons why qualia are evoked in this context are not epistemological, such as, for example, in the case of sense-data or similar philosophical posits. The motivation behind the qualia talk is not to find entities whose knowledge would be absolutely indubitable or infallible but rather to capture the nature of consciousness.

Proponents of qualia who draw attention to qualia in order to raise objections against materialist reductionism about consciousness, do not need absolute certainty about the nature of their qualia. They are just fine knowing, with a reasonable amount of certainty, that they now have mental states featuring particular qualia, knowing that there is now something it is like to have such states and perhaps also believing with a reasonable justification (consistent with a possibility of error) that they had certain qualitative states in the past, without any need for infallible knowledge that the past qualia were exactly the same or different from their present qualia. That itself seems to suffice to raise an objection against materialism, including, of course, Dennett's functionalism.

Merely allowing for the fallibility of some of our qualia judgments would, however, miss the point of Dennett's argument. The real problem for the qualia theorist, according to Dennett, is rather that the fallibility seems to undermine the traditional view that we somehow directly apprehend our qualia. Here Dennett's thought seems to be that given that, as his thought experiments emphasize, we can fairly easily make mistakes about our past qualia, it would be strange to hold that we are intimately acquainted with our qualia or that we directly apprehend them. Here, however, the proponent of qualia can simply insist that nothing Dennett says casts doubt on the claim that we are intimately acquainted with our *current* qualia. This seems to be true anytime we pay sufficient

[92]Block (1994).

attention to the qualia we are at the moment having – consider attending to your taste qualia when eating a mango or to your visual qualia when looking at the clear blue sky. It seems plausible to say that one is at that moment acquainted with ones' qualia and has direct knowledge of them.

Dennett anticipates that this sort of reply can be made by qualia proponents and rejects it. The problem with this sort of reply, he argues, is that "if absolutely nothing follows from this presumed knowledge – nothing, for instance, that would shed any light on the different psychological claims that might be true of Chase or Sanborn – what is the point of asserting that one has it?"[93] Here I think Dennett is right that, for example, Chase's knowledge of his current Maxwell House coffee quale will not help us answer the question whether his qualia have changed or stayed constant. Still, the proponent of qualia can insist that at the very least, Chase's knowledge of his current qualia indicates that more needs to be explained when it comes to one's consciousness than how the brain implements various mental capacities. Qualia after all seem to be something over and above mere capacities.

Dennett's remarks on fallibility of certain judgements about qualia could, I think, motivate proponents of qualia to formulate a limited infallibilism thesis which would be grounded in the fact that at a given moment we perhaps have some infallible knowledge of our current qualia – we know how our mind is right now, phenomenally speaking. Chalmers attempts to formulate such a thesis when he tells us that we have infallible knowledge only when it comes to what he calls "direct phenomenal beliefs". These are a mere subset of the beliefs which we have about the qualia we are currently experiencing. These beliefs attribute a certain qualitative nature captured by a direct phenomenal concept to a quale picked out via a demonstrative concept and have roughly the form "This state is such and such" where "this state" is a demonstrative concept and "such and such" is a direct phenomenal concept which captures a particular phenomenal character and is formed solely by attending to this phenomenal character.[94] It seems that Dennett's scenario described above gives us no reason at all to doubt such direct beliefs about our qualia. If, however, even such a small class of our beliefs about qualia is infallible, then we seem to have a strong reason to believe that qualia or something very much like them are real properties of experiences.

5. The Intrinsicality of Qualia

Another second-order property traditionally attributed to qualia is their intrinsicality. This property is usually understood in contrast with the purely relational character of certain properties. Here the thought is that there are two kinds of properties in the world: relational and intrinsic. Things have

[93]Dennett (2002, p. 233).
[94]Chalmers (2010, pp. 277-279).

relational properties in virtue of being related to other things in particular ways, the obvious examples being “to be 5 meters from Tom” or “to be Tom's sibling”. Intuitively, apart from relational properties, there are also intrinsic properties, properties which things possess in virtue of being what they are in and of themselves. Consider the phenomenal experience of red, for example. It certainly sounds plausible that the phenomenal character of this experience is not fully exhausted by its relations to other things, which means that the phenomenal character is arguably fully, or at least partially intrinsic to the experience, had by the experience, as it is in and of itself. This kind of consideration has resonated among many of the proponents of qualia. Chalmers, for example, writes that “[t]here is only one class of intrinsic, nonrelational property with which we have any direct familiarity, and that is the class of phenomenal properties”.[95] Similarly, Jaegwon Kim writes “it seems to me that the felt, phenomenal qualities of experiences, or qualia, are intrinsic properties if anything is”.[96]

These quoted views, of course, tell us that qualia are intrinsic properties, but they also allow that there may not be any other truly intrinsic properties. The question whether there are any intrinsic properties outside the realm of phenomenal consciousness, will not be of central importance here but is certainly an interesting one. At first sight, the properties of size, shape or mass can seem like paradigm cases of intrinsic properties, but upon reflection it is not clear that they are. When it comes to size and shape, current physics arguably tends to view them as rather dispositional, i.e. relational properties of things. As William Seager tells us, moreover, the property of mass may arise from interaction with the Higgs field which would arguably mean that mass is not intrinsic.[97] Seager considers the physical property of spin as a better example of an intrinsic property but this example will hardly help us really grasp the idea of intrinsicality. It is therefore unclear whether there are any intrinsic properties apart from, supposedly, qualia.

The question whether there are any properties of experience whose nature is not exhausted by their relations to other entities or properties, is of course, of crucial importance for Dennett. For him, as we saw, the fact that a particular state is conscious amounts to the fact that it has particular dispositions within a complex system of functional states, which, according to functionalism, constitute the mind, and behavioural outputs. If it turned out that conscious states have qualia and qualia are intrinsic, the task of reducing them to functional states, as envisioned by Dennett, would be hardly possible. Dennett therefore attempts to provide us with reasons to doubt the claim that conscious states have any intrinsic properties. Let us now take a look at his argument.

Firstly, Dennett emphasizes that philosophers who work with the notion of intrinsicality, fail to

[95]Chalmers (1996, p. 153).
[96]Kim (1998, p. 102).
[97]Seager (2010, p. 171).

provide us with its clear definition.[98] Here Dennett appeals to the fact that there is an ongoing debate about the notion of intrinsicality in philosophy which has not yet provided us with a satisfying definition of the notion, i.e. a definition against which there would be no counterexamples.[99] He suggests that we should therefore give up the notion of intrinsicality until an uncontroversial definition of intrinsicality is formulated, if that ever happens at all.

Here I think the proponent of qualia can reply that while Dennett is certainly right that we are not in possession of a satisfactory definition of intrinsicality, it is far from clear that this means that we should abandon the notion. That certainly does not seem to be a common practice in philosophy. Many of the terms philosophers use are taken by them to be primitive, i.e. such that they cannot be defined, only perhaps approximately characterised. What's more, philosophers often work with terms which they do not take to be primitive and yet for which we currently have no satisfactory definition – here the concept of knowledge comes to mind. In such cases philosophers typically work with approximate definitions, i.e. definitions of the relevant phenomena, which get most cases right, against which, nevertheless, counterexamples exist.[100] It seems to me therefore that Dennett's appeal that we should give up the notion of intrinsicality until its satisfactory definition is formulated, is ungrounded.

Moreover, as I see it, the concept of intrinsicality can, for the purposes of the consciousness debate, be defined negatively. Intrinsic properties are then those properties whose nature is not fully relational, i.e. whose nature is not reducible to relations to other properties. Here I think that the notion of a relation is very likely primitive and non-definable, or at least not definable in simpler terms. Notice that this negative notion of intrinsicality will still be an obstacle for the functionalist reduction of consciousness and it is far from clear that there will be clear counterexamples against such a purely negative definition.

Dennett's main argument against the claim that qualia are intrinsic is based on his claim that qualia, supposing that they exist, cannot be separated from our judgements about them and in general from our reactions to them.[101] These judgements, Dennett argues, co-constitute the qualia of our experiences which casts doubt on the supposed intrinsicality of qualia. Without intrinsicality, however, there are, for Dennett, no qualia, at least no qualia in the traditional sense.

Why think, however, that our judgements about our qualia co-constitute the qualia themselves? Here Dennett first invites us to think once again about Chase, the taster of Maxwell House, who no longer likes the coffee and explains this shift as a shift in his judgements about qualia, which have

[98]Dennett (2002, p. 240).
[99]See e.g. Langton – Lewis (1998).
[100]See Chalmers (2012, p. 17).
[101]Dennett (2002, p. 236–237).

supposedly remained constant. He used to like the Maxwell House coffee qualia but he no longer does. Dennett now brings into the story Chase's wife who, once Chase tells her about the change, responds to him "Don't be silly! Once you add the dislike you change the experience!"[102] Chase, according to Dennett, gradually realises that she is right about this. If Chase's wife is right, however, we should, Dennett thinks, accept that our aesthetic judgments about our qualia co-constitute our qualia.

Dennett subsequently introduces another thought experiment to cement this conclusion. This thought experiment works with the notion of an acquired taste. He invites us to think of an experienced beer drinker who did not enjoy his first sip of beer which he took many years ago but who, nevertheless, gradually worked his way towards the current state of really enjoying the taste of beer.[103] If we asked this person whether what he enjoys is the taste of his very first sip of beer, he would presumably laugh at us and tell us that nobody can enjoy *that* taste. If we did, beer would clearly not be an acquired taste. The taste which the beer drinker is currently enjoying is a taste, which he worked his way towards by means of regularly drinking beer. Indeed, precisely the fact that the beer drinker now enjoys the taste of beer guarantees, according to Dennett, that the new taste is not the same as the taste of the first sip.[104]

If we accept this conclusion, argues Dennett, we will have a good reason to doubt the intrinsicality of qualia. Clearly, after all, the aesthetic judgement about a supposed quale is a different mental entity than the quale itself, as it is a reaction to the quale. At the same time, however, as Dennett's scenario is meant to show us, the aesthetic judgement co-constitutes or influences the nature of the quale. What the quale is, is then at least partially a matter of the reactions which it triggers. If so, qualia can hardly be intrinsic, i.e. non-relational. That means, concludes Dennett, that we are not dealing with qualia or at least not with qualia in the traditional sense of intrinsic properties which the introduced argument targets.[105]

How effective is Dennett's argument against the intrinsicality of the supposed qualia? As I see it, the first thing that should be said about it is that it is not clear that it targets the notion of intrinsicality which is the most relevant in the context of the debate about the reducibility of consciousness. According to the negative definition of intrinsic properties provided above, intrinsic properties are those properties whose nature is not fully relational, i.e. not reducible to the property's relations to other properties. Call this notion of intrinsicality *weak intrinsicality*.

It is not clear that anything Dennett says casts doubt on this sort of weak intrinsicality of qualia.

[102]Dennett (2002, p. 236).
[103]Dennett (2002, p. 236).
[104]Dennett (2002, p. 237).
[105]Dennett (2002, ibid.).

Even if we allow, as Dennett thinks we should, that some, perhaps all, qualia are co-constituted by judgements about these qualia, it would still give us no reason to assume that the nature of qualia is fully exhausted by and reducible to these relations. Indeed, the way Dennett phrases things, i.e. that we sometimes, perhaps always, make judgements about qualia and these judgements change the nature of qualia, seems quite compatible with there being non-relational aspects of qualia. Importantly, this weak notion of intrinsicality can still be viewed as an obstacle for Dennett's functionalism, according to which conscious states are defined purely relationally, i.e. purely in terms of what they do, what causes and effects they have. Even weak intrinsicality of qualia will thus clearly be a problem for the project of the functionalist reduction of consciousness.

At the same time, it is not clear why the proponent of qualia could not hold that qualia are intrinsic only in this weak sense and thus arguably save the notion of qualia as intrinsic properties. It seems to me therefore that Dennett's considerations about intrinsicality do not show that qualia do not exist.

As I see it, the problem of Dennett's argument is that it targets a rather strong notion of intrinsicality and it is not clear that the proponent of qualia is committed to this notion. According to this notion, a property is intrinsic iff its instantiation and nature is not co-constituted by anything else, presumably any other properties or entities. Even though, as already mentioned, I think one can reasonably hold that qualia exist without holding that they are intrinsic in this stronger sense, it is, I think, interesting to ask whether Dennett's argument manages to cast doubt on the claim that qualia are in this sense strongly intrinsic.

I suggest that there are reasons to think that the argument is ultimately unsuccessful even when it comes to *strong intrinsicality* of qualia. As I see it, the proponents of qualia could insist that while our judgements about qualia effect the qualia we experience, this is not, upon reflection, in conflict with the intrinsicality of qualia. Namely, they could say that when we make a particular, e.g. aesthetic, judgement about a particular quale, call it Q1 which we experience at time T1, and accept at the same time, as I think we should, that this judgement can change the nature of the quale, it is reasonable that the act of judgement gives rise to a new quale, call it Q2. Moreover, given that our judgement is, as suggested by Dennett, a reaction to Q1, it is reasonable to think that Q2 comes into existence at a later time, call it T2.

Does this mean that either Q1 or Q2 are somehow co-constituted by something external to them, i.e. by the given aesthetic judgement? Consider Q1 first. The relevant aesthetic judgement is a reaction to Q1 which means that the nature of Q1 is *ex hypothesi* not co-constituted by the judgement. What, however, is the relation between the aesthetic judgement about Q1 and the new quale Q2? Here,

clearly the given act of judgement led to the change of Q1 into Q2 so, the judgement at the very least was a part of the cause why at T2 we are experiencing Q2 and not Q1.

It seems to me that if the judgement merely causally contributed to Q2's coming into existence, this, on its own, does not mean that Q2 is not strongly intrinsic. It is, after all, compatible with the claim that the nature of Q2 is not constituted by anything else than Q2 itself, that this nature is not constituted by a relation of Q2 to something else.

Is there any reason then to claim that the nature of Q2 is somehow constituted by Q2's relation to something else, merely the particular judgement about Q1? If so, then we would need to give up the thought that qualia are strongly intrinsic. It is not clear to me, however, that the proponent of qualia needs to hold that the nature of Q2 is constituted by its relation to the relevant judgement. The judgement, after all, already, so to say, did its job by giving rise to Q2 and there seems to be, as I see it, no reason to think that it would still need to somehow co-constitute Q2.

If these considerations are correct, then Dennett's argument does not manage to demonstrate that qualia are not strongly intrinsic. Even if they are not, however, all Dennett says seems to be compatible with weak intrinsicality of qualia. I believe then that Dennett's considerations about intrinsicality do not show that qualia do not exist.

In this chapter I have attempted to show that Dennett's arguments against the existence of qualia, although interesting, are ultimately unsuccessful. Supposing that my considerations were correct I suggest that we should take our first-person evidence that qualia exist seriously. Once we do that, however, we are back with the epistemic gap whose existence gives us strong reasons to reject a priori physicalism. Luckily for the physicalists, there is another established form of physicalism which they can appeal to, namely a posteriori physicalism, a view I shall discuss in the next two chapters.

3. A Posteriori Physicalism

1. Living with the Epistemic Gap

In the previous chapter I introduced the distinction between a priori and a posteriori physicalism. We saw there that these two branches of physicalism differ with respect to the way their proponents philosophically react to the apparent epistemic gap between physical and phenomenal truths. As explained, if there is an epistemic gap between physical and phenomenal truths, there is, even given ideal reflection, no a priori entailment between physical and phenomenal truths.[106] While a priori physicalists, whose views I discussed in the previous chapter, deny that there is an epistemic gap, a posteriori physicalists, whose views will be the topic of this and the next chapter, grant the existence of this gap, although they, unlike the anti-physicalists, deny that the epistemic gap has ontological implications. In the key of conceivability, we can say that while the a priori physicalists deny even the conceivability of $P\&\sim Q$ (where P is the complete physical truth and Q is an arbitrary phenomenal truth), the a posteriori physicalists allow for the conceivability of $P\&\sim Q$ but deny that the conceivability of this scenario implies the possibility of $P\&\sim Q$. The attractiveness of a posteriori physicalism, compared to a priori physicalism, then consists in the fact that, if successful, it shows us how the epistemic gap is consistent with the austere physicalist metaphysics.

In order to appreciate the attractiveness of a posteriori physicalism, one needs to understand, firstly, why it is plausible that there is an epistemic gap and, secondly, why it is plausible that the epistemic gap implies an ontological gap. Here, "ontological gap" simply means the (supposed) fact that consciousness is a non-physical property (or entity) in the physical world. It is, of course, the second point, i.e. the step from the epistemic and the ontological gap, which the a posteriori physicalists take issue with. Supposing that I provided sufficient motivation for the belief in the existence of an epistemic gap in the previous chapter, let me now focus on the second point.

Why should the epistemic gap imply the metaphysical gap? Here one can appeal to the overall plausibility of the general claim that if consciousness is a macro-physical property, there should be no epistemic gap between truths about consciousness and physical truths. One possibility is to appeal to the thesis that truths about macro-physical properties are arguably often entailed by truths about micro-physical properties and their organisation. Many properties of organisms, such as for example, being a gene, are, after all, naturally understood as functional properties and truths about functional properties are arguably a priori entailed by the truths about the microphysical

[106]In what follows I will, unless specified otherwise, use the term "epistemic gap" to refer to the epistemic gap between physical and phenomenal truths. The term can, of course, be also used more generally to refer to an epistemic gap between any two classes of truths.

mechanisms realising these functions.[107] Once we know, for example, that DNA realises the function of storing genetic information and transmitting it over to next generations, we will know a priori, that the property "being a gene" is realised. It is arguable that the same goes for the property "being alive" which can be understood as a functional property which an organism has iff it metabolises, reproduces, adapts to its environment, etc. One can appeal to these cases to lend inductive support to the thesis that if consciousness is physical, then truths about it must be a priori entailed by micro-physical truths, i.e. to the view that the epistemic gap speaks against the physical nature of consciousness.

A particularly vivid way of arguing for the step from the epistemic gap to the ontological gap is Jackson's knowledge argument, introduced in the previous chapter.[108] This argument basically tells us that if a macro-property is physical, then its nature is fully predictable, at least for an ideal reasoner, given complete physical knowledge. In other words, if a macro-property is physical, then all truths about this property are a priori entailed by complete physical knowledge. The thought behind the argument is then that if having a phenomenally red experience were a physical property then Jackson's Mary – given her ideal epistemological situation and superior reasoning abilities – would be able to deduce from this physical knowledge the knowledge of what it is like to see red. Intuitively, however, Mary is unable to do that and thus learns this phenomenal fact only once she is released from her black-and-white room and actually sees red for the first time. Applying the *modus tollens* to this reasoning, we will, it seems, need to reject the original claim that having a phenomenally red experience is a purely physical property. The epistemic gap then arguably implies the ontological gap. Of course, this reasoning is not uncontroversial, but I think it certainly provides some support to the thesis that the epistemic gap implies an ontological gap.

Finally, one can appeal to the conceivability argument to find support for the step from the epistemic gap to the metaphysical gap. Here the fact that there is the epistemic gap just means that $P\&\sim Q$ is conceivable and the premise that $P\&\sim Q$ is possible leads directly to the ontological gap. Why, however, should conceivability of $P\&\sim Q$ imply possibility of $P\&\sim Q$? This is a subtle issue which I cannot hope to do full justice here. Still, as I see it, there is a strong case for the view that (a) some kind of conceivability-possibility link (a C-P link for short), if perhaps highly attenuated, should be taken seriously, and (b) that this C-P link can get us from the epistemic gap to the ontological gap.

Why, however think that there is any C-P link at all? Here we can simply appeal to the intuitive thought that in many cases things and scenarios which do not actually exist, but which are

[107] Chalmers (2010, p. 7).
[108] See e.g. Jackson (1982).

conceivable, are also arguably possible, or at least there is no clear reason to deny their possibility. There is, for example, as a matter of fact not an exact, brick-to-brick replica of the White House in Prague, but it is certainly conceivable that Prague would feature such a replica and there is no reason to think that this conceived scenario does not correspond to a real possibility. There is, to use Chalmers's example, actually no one-mile-high unicycle, but it certainly seems conceivable that such a vehicle would exist and, once again, there is no reason to think that its existence is not possible.[109]

More generally, one can appeal here to the claim that it is hard to make sense of how we find out about what is possible if not from what is conceivable; we certainly could hardly find out about possibilities by merely investigating the actual world. The general claim that there is a C-P link is, of course, quite compatible with the view that this link is highly attenuated. Let me now bring to our attention some of the ways this link would need to be attenuated in order for its existence to be plausible.

Perhaps the most common objection against the claim that there is a C-P link of some sort is that some conceivable scenarios are just not physically possible. Although, for example, it seems clearly conceivable that something would travel at a speed higher than a speed of light, there may well be no logical contradiction in a description of a scenario of this sort, it is not in fact physically possible given the laws of nature.

Here we can say, using the logical apparatus of possible worlds, that the laws of nature place constraints on physically possible worlds. This objection then amounts to the claim that for some conceivable scenarios, such as the one in which an entity travels faster than light, there are no corresponding physically possible worlds and, as a result, the space of physically possible worlds does not mirror the space of conceivable scenarios.

A natural reaction to this sort of objection is to distinguish physical possibility from metaphysical possibility and hold that while there is no link from conceivability to physical possibility, there surely is a link from conceivability to metaphysical possibility. What, however, does metaphysical possibility amount to? The easiest road to grasping the concept of metaphysical possibility is to consider that it is conceivable that laws of nature could be different from the way they are in our world. Surely, one can conceive of possible worlds which are nomologically different from our world and in which thus physicists have been discovering laws of nature which are to a degree different from the laws of nature in our world. One can, for example, think of sci-fi short stories which describe such worlds and it is far from clear that the scenarios described in these stories are

[109]Chalmers (1996, p. 96).

inconceivable, i.e. logically incoherent.[110] Are such conceivable worlds also possible? Clearly, they will not be physically possible as they feature different laws of nature. It is here that the notion of metaphysical possibility will come in handy. We can say that metaphysically possible worlds include, apart from worlds which are nomologically equivalent to our world,[111] also conceivable worlds with laws of nature which differ from those which exist in our world. While the above-mentioned objection challenges the link from conceivability to physical possibility, it leaves the link from conceivability to metaphysical possibility intact.

Another objection which is sometimes raised against the claim that some C-P link is viable appeals to examples such as that it is conceivable that Goldbach's conjecture is true but it is also conceivable that it is false. Clearly, however, only one of these two conceivable scenarios corresponds to a true possibility.[112]

In response to this objection it will be necessary to distinguish what is conceivable for us, at the stage of mathematical knowledge we are at and what is conceivable given ideal reasoning abilities and unlimited time for reflection. We can say that it is *prima facie conceivable* both that Goldbach's conjecture is true and that it is false but only the true option, whichever it is, is *ideally conceivable.* Here the thought is that the ideal intellect will be able to figure out whether the conjecture is true or false and the other option will then no longer be conceivable for this intellect.[113] While this objection then questions the link from *prima facie* conceivability to metaphysical possibility, it leaves the link from ideal conceivability to metaphysical possibility intact.

Sometimes, moreover, the link from conceivability to possibility is rejected with appeal to the cases of a posteriori necessity discussed by Saul Kripke who famously argued that certain identity statements whose negations are conceivable, such as “Water is H_2O” or “Hesperus is Phosphorus”, are necessary, i.e. true in all possible worlds.[114] If Kripke is right, then even though it is conceivable that water is XYZ (where XYZ stands for a chemical substance which is just like water, i.e. is, very roughly, an odourless, drinkable liquid which exists in rivers, lakes and seas, but has a somewhat different chemical structure), it is not metaphysically possible and the link from conceivability to possibility seems, once again, under threat.

In reply the proponent of a C-P link can appeal to the fact that there is a clear sense in which “Water is XYZ” is not conceivable.[115] To see this, one needs only to notice that given that, if Kripke is right,

[110]Here the example of the scenario described in Walter S. Tevis's story “The Big Bounce”, in which the law of the conservation of energy is broken with respect to a specific ball whose behaviour is then governed by a different law, comes to mind.
[111]We can say that two possible worlds are nomologically equivalent iff they have the same laws of nature.
[112]See Chalmers (2010, p. 145).
[113]Chalmers (2010, p. 143).
[114]Kripke (1980).
[115]Chalmers (2010, p. 145).

"water" is a rigid designator, and given that the concept refers to H_2O in the actual world, it plausibly refers to H_2O in all possible worlds. In view of that, "Water is XYZ" looks, in the end, inconceivable. To think of conceivability in this sense is to take into consideration a posteriori facts about reference of our concepts in the actual world (and the Kripkean intuitions about the semantics of natural kind terms). We can, with Chalmers, call this kind of conceivability *secondary conceivability*. There is, however, another sense of conceivability, in which "Water is XYZ" is conceivable. This kind of conceivability results from considering what the concepts expressed by the statement reveal to us a priori. Clearly, a priori reflection on the concept "water" does not reveal to us that it refers to H_2O but rather that it refers to what we can call "watery stuff", i.e. very roughly, liquid, odourless, drinkable stuff that can be found in rivers, lakes, seas, etc. If this consideration is correct, there seems to be a clear sense in which "Water is XYZ" is conceivable. We can, with Chalmers, call this kind of conceivability *primary conceivability*.[116]

With this distinction in place, we can see that even the Kripkean cases do not quite sever the C-P link. Namely, these cases are compatible with the claim that secondary conceivability implies metaphysical possibility. Of course, this C-P link is, on its own, not very interesting from the point of view of the consciousness debate as in this debate the actual reference of phenomenal concepts is disputed. Is there any other C-P link which the Kripkean cases allow for? Here we can notice that even if we accept Kripke's reasons for holding that statements such as "Water is XYZ" do not describe metaphysical possibilities, there seem to exist possible worlds which are interestingly related to these statements, namely possible worlds in which the watery stuff is XYZ. Of course, if Kripke is right, then it would be wrong to describe these worlds as possible worlds in which water is XYZ – as long as we suppose that water is actually H_2O and as a result we view these worlds as counterfactual. If, however, we give up on this supposition and think of a possible world in which the watery stuff is XYZ as of the actual world, then this world can plausibly be correctly described as a world in which water is XYZ.

Some descriptions of possible worlds are then correct about these worlds if they are considered as actual but false about them if the worlds are considered as counterfactual. Take, for example, the XYZ world, i.e. the world in which the watery stuff is XYZ. If Kripke is right, then the statement "Water is XYZ" will be correct about this world if this world is considered as actual, but the statement will be false about the XYZ world if the world is considered as counterfactual and if the watery stuff in the actual world is, as a matter of fact, H_2O or another substance non-identical with XYZ. We can say, using Chalmers's terminology, that the XYZ world does not *satisfy* "Water is XYZ" (the statement is false about this world if the world is considered as counterfactual) but this

[116]Chalmers (2010, ibid.).

world *verifies* this statement (the statement is true about this world if the world is considered as actual).[117] Here the idea is that the state of the actual world sets certain a posteriori requirements on the reference of the given concept (the concept *water* in this case) across possible worlds and the particular possible world which we consider as counterfactual either satisfies or fails to satisfy these requirements.

This point can also be expressed as the thought that concepts have two dimensions of meaning. Here it is natural to think of meanings of concepts in terms of the concepts' intensions and to think of intensions as expressible in terms of requirements on the concept's reference expressed by the given concept. We have seen that, for example, the concept *water* will have different requirements with respect to its reference in a given possible world if the world is considered as actual, and if the world is considered as counterfactual. We can call the set of requirements on reference in the possible worlds considered as actual, which the concept expresses, the *primary intension* of the concept. In the case of the concept *water*, for example, its referent in the possible world considered as actual needs to meet the condition of being the watery stuff. We can call the set of requirements on reference in possible worlds considered as counterfactual, which the concept expresses, the *secondary intension* of the concept. In the case of the concept *water*, its referent in the given possible world considered as counterfactual needs to meet the condition of being H_2O.[118]

We can, of course also distinguish two dimensions of the meaning of propositions. It is natural to think of the meaning of a proposition in terms of the proposition's intension where the intension is, roughly, a set of requirements expressed by the proposition which are such that if the given world satisfies these, the proposition is true. Once again, we will get different results if we consider the given possible world as actual and if we consider it as counterfactual. We can see that the proposition “Water is XYZ” is true in the XYZ world if the world is considered as actual, i.e. the XYZ world verifies the proposition. This is just another way of saying that the primary intension of this proposition is true in the XYZ world. The proposition “Water is XYZ” is, on the contrary, false in the XYZ world if the world is considered as counterfactual and water is H_2O in the actual world, i.e. the XYZ world fails to satisfy the proposition. This is just another way of saying that the secondary intension of this proposition is false in the XYZ world, given that water is actually H_2O.

With this terminology on the table, we can say that if there is a possible world which verifies “Water is XYZ”, then the statement expresses a scenario which is primarily possible. If, moreover, there is a possible world which satisfies “Water is XYZ”, then the statement expresses a scenario which is also secondarily possible. At this point we can see that no world satisfies “Water is XYZ”

[117]Chalmers (2010, p. 146).
[118]See Chalmers (2010, pp. 544–545).

and, as a result, the statement expresses a scenario which is not secondarily possible. There seem to be, however, possible worlds which verify "Water is XYZ" so the statement expresses a scenario which looks primarily possible. Which worlds will these be? Intuitively, they will be worlds in which the watery stuff in rivers and lakes is XYZ.

Given these considerations, we can see that the scenario expressed by "Water is XYZ" is primarily possible but not secondarily possible. The objection which appeals to the Kripkean cases then challenges a link from primary conceivability to secondary possibility. We have already seen that these cases are compatible with a link from secondary conceivability to secondary possibility. At this point we can see that the Kripkean cases are also compatible with a link from primary conceivability to primary possibility. The Kripkean challenge then leaves some C-P links intact, namely a link from primary conceivability to primary possibility as well as the link from secondary conceivability to secondary possibility.

It seems reasonable to conclude then that there is at least a *prima facie* case for a C-P link, if perhaps a highly attenuated one. What will this link be? It is easy to see at this point that the relevant link will be one from ideal primary conceivability to primary metaphysical possibility. Does this link suffice to establish a case for a step from the epistemic gap to the metaphysical gap? Let me now explain why I think there are reasons to believe that it does. It seems clear that if *P&~Q* is ideally primarily conceivable, then there is the epistemic gap. Indeed, it is natural to view the notion of an epistemic gap as relying on ideal rather than merely *prima facie* conceivability. Moreover, it seems clear that the notion should rely on primary conceivability since, as has been mentioned, the actual reference of phenomenal concepts is in dispute.

A less straightforward question is whether the primary metaphysical possibility of *P&~Q* implies an ontological gap. Here one could object, firstly, that perhaps we need to have established the physical possibility of *P&~Q* in order to get to the ontological gap. If that is true, metaphysical possibility would not be strong enough to lead us to the ontological gap.

It is, however, unclear what would be the motivation for holding that physical possibility of *P&~Q* leads us to the ontological gap while the metaphysical possibility of *P&~Q* does not. Recall that the metaphysical possibility of *P&~Q* means that there is a possible world which is physically just like ours although with different laws of nature and yet there is no consciousness in this world. I think, intuitively, this speaks quite strongly for the ontological gap, i.e. the view that consciousness is non-physical. Of course, the physicalist can appeal here to the view that our world simply has some laws of nature which explain why certain configurations of physical entities give rise to consciousness. We can call this subset of the actual laws of nature the psycho-physical laws.

The physicalists who hold that only the physical possibility of a zombie world would disprove their view, then appeal to psycho-physical laws which are true about the actual world in order to show that consciousness is physical. It is easy to see, however, that this appeal cannot help the physicalists here. Recall, after all, that if a physical replica of our world without consciousness is metaphysically possible, this replica will already include all the physical properties although it will not, of course, include psycho-physical laws. The extra properties or entities whose production is supposed to be explained by the appeal to the psychophysical laws will therefore need to be non-physical. It seems to me therefore that the appeal to the psycho-physical laws cannot save the physicalist here. As I see it, the view that $P\&\sim Q$ is not physically possible although it is metaphysically possible is in fact a non-reductive view of consciousness which is best described as a form of emergentism, a view I shall further discuss in chapter 5.

Another objection to the claim that primary metaphysical possibility of $P\&\sim Q$ implies a metaphysical gap tells us that only secondary metaphysical possibility of $P\&\sim Q$ in fact implies a metaphysical gap.

To answer this objection it will help to specify what the primary metaphysical possibility of $P\&\sim Q$ amounts to. Given what we said above, it amounts to the fact that there is a world, call it W, which verifies $P\&\sim Q$ although this world may not satisfy $P\&\sim Q$. How could this be the case? Here the thought seems to be that physical and/or phenomenal properties may have a hidden structure which the concepts pick out in the actual world but which would – in the case of physical properties be missing in W, or which would – in the case of phenomenal properties – be present in W. Given that, W would not satisfy $P\&\sim Q$ although it would verify this statement. What, however, does it mean that W verifies $P\&\sim Q$ (although it may not satisfy $P\&\sim Q$)? It means that $P\&\sim Q$ is true in W when evaluated in terms of its primary intension (although $P\&\sim Q$ may be false in W when evaluated in terms of its secondary intension). What, however, is the primary intension of phenomenal concepts? The concept of phenomenal red, for example, plausibly refers to a property in the actual world iff this property features a specific qualitative feel (whatever its alleged hidden nature is). Similarly, it is arguable that a physical concept refers to a particular property in the actual world iff this property plays a particular role in a given physical theory and in the world (whatever the hidden nature of the property is).[119] The primary metaphysical possibility of $P\&\sim Q$ then means, very roughly, that there is a possible world which is just like our world when it comes to the micro-physical roles instantiated but lacks the phenomenal feels instantiated in our world. Should this possibility motivate us to give up physicalism and posit the ontological gap? I think so because this possible world is arguably, physically, just like our world but includes no consciousness, which directly

[119]See chapter 6 for much more on this.

implies the ontological gap.

The physicalists could object against this conclusion that perhaps this possible world only lacks the appearance of consciousness in our world, i.e. it lacks what phenomenal concepts refer to via their primary intensions, and does not lack consciousness itself. It seems, however, that this very fact implies a sort of ontological gap, a gap between the physical world and the appearance of consciousness. This gap, however, will be, I believe, all the anti-physicalists need. Moreover, one can argue – with Chalmers – that phenomenal concepts have identical primary and secondary intensions.[120] It is, after all, highly plausible that what consciousness is precisely the appearance, which means that consciousness does not have, unlike, for example, water, any hidden structure which would lead to the difference between the primary and secondary intension of the relevant concept.

The physicalist could further object to this that there is an important difference between a replica of our world with respect to the microphysical roles instantiated and a replica of our world with respect to all physical properties instantiated (i.e. with respect to the given role-properties as well as the supposed hidden natures realising these roles). In order for this objection to have any force, the physicalist would, however, need to specify how the additional properties which are not role-properties could render *P&~Q* inconceivable, given that they allow that *P&~Q* is conceivable if P only involves instantiations of role-properties. As far as I can tell, the most natural way to do that would be to go the way of Russellian monism, according to which micro-physical role-properties are realised by protophenomenal or phenomenal properties, often called quiddities. While I find this view fascinating and will have much to say about it in chapters 6 and 7, it has little to do with physicalism as it is normally understood by the a posteriori physicalists and I shall therefore ignore it here.

I hope these, somewhat cursory, considerations have shown that there is at least a strong *prima facie* case for the view that the epistemic gap leads to the ontological gap. As already mentioned, however, the a posteriori physicalists still challenge this step. Let me now explore their reasons for thinking that this step should be resisted.

2. Phenomenal Concept Strategy

Perhaps the most promising version of a posteriori physicalism has been advanced by Brian Loar, David Papineau, Janet Levin, Peter Carruthers, Esa Diaz-Leon and others and has been called the

[120]See Chalmers (2010, s. 150).

"phenomenal concept strategy" by Daniel Stoljar.[121] The proponents of the phenomenal concept strategy suggest that the existence of the epistemic gap results from the nature of our phenomenal concepts which, nevertheless, can be, they insist, explained in a manner compatible with physicalism.[122] As a result, we should, according to these thinkers, resist the step from the epistemic gap to the metaphysical gap. According to Loar, for example, our phenomenal concepts are recognitional concepts of a special kind and have the same referents as the physical concepts by means of which we refer to some of our brain-processes. Since, Loar argues, truths involving recognitional concepts are, in general, not a priori entailed by physical truths, it is hardly surprising that truths involving our phenomenal concepts are also not a priori entailed by physical truths and that we end up with an epistemic gap. Given that the existence of the epistemic gap is thus explainable in a way compatible with physicalism – recognitional concepts seem to pose no special problems for physicalism – it would seem strange, according to Loar, to think that the epistemic gap indicates the existence of the metaphysical gap.

According to David Papineau, another vocal proponent of the phenomenal concept strategy, the epistemic gap between physical and phenomenal truths can be explained in term of the "use-mention feature" of our phenomenal concepts.[123] When a phenomenal quality is *mentioned* in our thinking (i.e. when we think about a certain phenomenal quality), we, Papineau suggests, thereby *use* it, i.e. roughly, the quality itself is activated when we employ the phenomenal concept by means of which we conceive of this quality. Given this feature of our phenomenal concepts, it shouldn't surprise us, Papineau thinks, that truths containing physical or functional concepts – concepts which do not in this way *activate* their referents – do not a priori entail truths about phenomenal qualities.

The proponents of the phenomenal concept strategy then hope to block the anti-materialists' move from the epistemic gap to the ontological gap by appealing to a sort of second-order reflection. Namely, they suggest that in order to block this step and the anti-physicalist arguments which support it, it will not be enough to think about consciousness; instead, we will need to step one level up and *think about the way we think about consciousness*. Importantly, the phenomenal concept strategists do not think that by doing this, we can arrive at an explanation of the existence of consciousness in the physical world. Instead, they suggest, we will be able to understand why we are not able to explain consciousness in physical terms and understand that our failure to do so does not indicate that consciousness is non-physical. The fact that the epistemic gap is explicable with reference to our psychological condition which amounts to a kind of conceptual dualism is supposed to show us, the phenomenal concepts strategists argue, that we do not need to accept the

[121]See e.g. Loar (1999, 2002), Papineau (2002, 2007), Stoljar (2005).
[122]For a different variety of a posteriori physicalism see e.g. Block – Stalnaker (1999).
[123]Papineau (2007).

metaphysical conclusions of the anti-materialist arguments. Whether the phenomenal concept strategy is successful at blocking the central anti-materialist arguments and in this way block the step from the epistemic to the ontological gap, is controversial and currently extensively debated. In this chapter I would like to contribute to this debate by evaluating the prospects of the view of phenomenal concepts held by Loar, who is one of the founders of the phenomenal concept strategy, in the light of a critique pressed by Chalmers.[124]

3. Loar on Phenomenal Concepts

According to Loar, we can accept the epistemic gap and yet sensibly deny that there is the ontological gap.[125] The key, in Loar's view, is to notice that our phenomenal concepts, unlike our physical concepts, are recognitional concepts of a kind. Once we understand that, Loar thinks, we will not expect phenomenal truths to be entailed by physical truths since, as he argues, recognitional truths never are.[126] He introduces the following example of the process in the course of which a recognitional concept of a particular kind of succulent is formed:

> *Suppose you go into the California desert and spot a succulent never seen before. You become adept at recognizing instances, and gain a recognitional command of their kind, without a name for it; you are disposed to identify positive and negative instances and thereby pick out a kind.*[127]

According to Loar then, our disposition to recognise instances of a particular kind of succulents is grounded in a corresponding recognitional concept. Thanks to this recognitional concept we are able to conceive of a particular thing which we perceive as of "one of *that* kind", classifying it due to the way it affects our senses. Think, for example, of a particular shape of the succulent's leaves or trunk, the size and colour of the plant, etc. Seeing that the recognitional concepts which we thus acquire are concepts of particular kinds of things, Loar describes them as type-demonstratives.[128] Conceiving of a particular thing or organism as of "one of that kind" does not require possessing any theoretical physical or biological knowledge of organisms of that kind, nor does the recognitional concept imply any such knowledge; in the course of one's life one indeed often possesses a recognitional concept of a certain kind of entity long before one has any theoretical physical knowledge of the given kind of entity.[129] That, however, arguably does not prevent us from

[124]Chalmers (1999).
[125]See Loar (1999).
[126]By phenomenal, physical and recognitional truths I mean here truths involving, respectively, phenomenal, physical and recognitional concepts.
[127]Loar (2002, p. 298).
[128]Loar (2002, ibid.)
[129]Many pre-school children can, after all, recognise cats from dogs and rabbits and thus possess a recognitional concept

referring to physical kinds by means of our recognitional concepts. Often, of course, we possess physical concepts of the kinds of entities which we also grasp via our recognitional concepts. In these cases then the given physical concept and the corresponding recognitional concept will pick out the same kind of entity or, we can say, the two concepts will corefer.

These coreferring physical and recognitional concepts are, according to Loar, conceptually independent. Clearly, after all, truths about the given sensory appearance will not entail truths about the underlying physical and chemical structure or *vice versa*.[130] Given the conceptual independence of recognitional and physical concepts and the fact that they, nevertheless, corefer, we can see why, for Loar, the epistemic gap does not lead to a metaphysical gap or, in other words, conceivability of zombies is compatible with their metaphysical impossibility. If phenomenal truths are not a priori entailed by (complete) physical truth, then zombies are clearly conceivable. Given, however, that the two kinds of concepts corefer – both refer to physical properties of the brain – zombies are, according to the a posteriori physicalist, not metaphysically possible. Any possible world which is a physical replica or our world will, after all, in that case also include consciousness since the coreference thesis implies that conscious states are (identical) with certain physical processes in our brains.

The anti-physicalists could object against this that the claim that our phenomenal concepts are recognitional concepts is dubious as the two kinds of concepts are in important respects different. They could, for example, emphasise that when we conceive of a particular physical thing or organism via a recognitional concept as of one of *that* kind, we apply our concept to the thing because the thing or organism appears to us in a particular way. The way things of a certain kind appear to me, however, seems to be a contingent feature of things of that kind, as they could in principle appear to me quite differently (or not appear to me at all), depending on the kind of sensory apparatus I am equipped with, or the way sensory information is processed in my brain, etc. Given these considerations, it is natural to say that if one conceives of a particular physical kind via a recognitional concept, one conceives of it indirectly, via a contingent mode of presentation. One can, using Loar's distinction, say that the relevant concept *expresses* a particular sensory appearance but *refers to* a particular physical kind.[131]

Having established that, the anti-physicalist could go on to argue that phenomenal concepts, unlike these paradigm recognitional concepts, do not refer to our phenomenal states indirectly, i.e. via contingent modes of presentation. They could support this claim by drawing attention to the fact that when I, for example, conceive of pain as a kind of state which feels like *this*, then what my

of cats even though they know little or nothing about cats' physical, biological and functional constitution.
[130]Loar (2002, p. 298).
[131]Loar (1999, p. 466).

concept expresses is by no means contingent with respect to pain itself. This point has been famously emphasized by Kripke who writes, for example, "[p]ain […] is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality".[132] One way to express this point is to say that the way pain appears is (essential to) what pain is.[133] If Kripke is right, phenomenal concepts pick out pain not via contingent modes of presentation but rather directly or, one could say, via a necessary mode of presentation, i.e. by expressing a necessary property of pain. A phenomenal concept then, to use Loar's distinction, expresses the same property it refers to, i.e. a particular type of a phenomenal feel. The anti-physicalist could then object against Loar's proposal that given this difference, it is dubious that phenomenal concepts are, as Loar would have them, recognitional concepts of a kind.

In response to this kind of objection, Loar suggests that he can embrace direct reference of phenomenal concepts and yet hold that they are recognitional concepts of a kind.[134] Phenomenal concepts are, in Loar's view, unusual recognitional concepts in that they, like e.g. physical concepts, such as H_2O, pick out their referents directly. They are, however, still recognitional concepts in that they enable us to conceive of their referents as instances of particular (phenomenal) kinds, picking them out via the ways they appear to us – which coincide with the ways they are.

We can see now that there is much that Loar and the anti-physicalists agree on. Namely, both sides agree with Kripke that our phenomenal concepts pick out their referents directly, rather than via contingent modes of presentation. They also agree that phenomenal and physical concepts are conceptually independent. Controversy arises, however, over the ontological implications of these two claims which we can call the *direct reference* claim and the *conceptual independence* claim. Loar insists that the two claims are fully compatible with physicalism while the anti-physicalists deny this. The anti-physicalists' reasoning will presumably go as follows: plausibly, with the exception of phenomenal concepts, in all known cases of pairs of coreferring concepts which lack an a priori connection, i.e. are conceptually independent, at least one of the two coreferring concepts picks out its referent, indirectly, i.e. via a contingent mode of presentation.

This sort of reasoning can lead the anti-physicalist to suggest that for any pair of concepts which are conceptually independent of one another, but which, nevertheless, corefer, it is the case that at least one of them picks out its referent via a contingent mode of presentation. Loar calls such a thesis the "semantic premise", a term which I will adopt here too.[135] We can see that if the semantic premise,

[132]Kripke (2002, p. 332).
[133]Not everyone agrees here. David Rosenthal (see e.g. Rosenthal 1991), for example, explicitly denies this thesis. However, given that the thesis is, as I shall show, accepted by both the anti-physicalists and by many a posteriori physicalists, including Loar, I shall suppose here that it is sound.
[134]Loar (2002, p. 297).
[135]Loar (2002, ibid.). Loar himself phrases the semantic premise as follows: "A statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a

the conceptual independence claim and the direct reference claim are granted, the pairs of phenomenal-physical concepts cannot corefer. Loar, however, thinks there is no good reason to hold the semantic premise and indeed rejects it. We can see that to deny this premise means to allow that physical-phenomenal pairs of concepts can corefer even if we accept the conceptual independence and the direct reference claims. Without the semantic premise, therefore, the conceptual independence claim and the direct reference claim seem to be compatible with physicalism. If that is so, Loar has shown us how physicalists can live with the epistemic gap.

4. Chalmers on Strong Necessity

David Chalmers has in his writings offered multiple critiques of the phenomenal concept strategy as well as of the more general doctrine of a posteriori physicalism and my discussion will necessarily be selective. The key problem with a posteriori physicalism, according to Chalmers, is that its proponents, including Loar, are committed to the existence of *strong metaphysical necessity*.[136] One can characterise the view that there is *no* strong metaphysical necessity as the view that each logically possible world is also a metaphysically possible world, or in other words, that there is a single space of possible worlds. Here we can view logically possible worlds as the worlds which are ideally conceivable, i.e. whose description involves no logical contradiction. If, however, there is strong metaphysical necessity, then there are further, a posteriori, constraints on the space of metaphysically possible worlds (apart from the constraints imposed by logic) and, as a result, some logically possible worlds are not metaphysically possible. In other words, strong metaphysical necessity implies that there are two different spaces of possible worlds: that of logically possible worlds and that of metaphysically possible worlds. If there is strong metaphysical necessity then, for example, even though the zombie world is logically possible, it may not be metaphysically possible and therefore the logical possibility of the zombie world may be compatible with the truth of physicalism.

It is important, according to Chalmers, to distinguish between strong and weak metaphysical necessity.[137] While both are kinds of a posteriori necessity, the two kinds are different in crucial respects. Weak metaphysical necessity is the kind of a posteriori necessity discussed by Kripke.[138] Why, however, call the Kripkean necessity weak? To see this, recall the discussion of the Kripkean cases earlier in this chapter. Using the two-dimensional apparatus introduced there we can say that the Kripkean a posteriori necessities, such as "Water is H_2O", have secondary intensions which are

contingent property of that property."
[136]Chalmers (1996, pp. 136–138).
[137]Chalmers (1996, p. 137).
[138]Kripke (1980, p. 128).

true in all possible worlds, i.e. necessary, but also have primary intensions which are not true in all possible worlds; if, for example, the XYZ world is considered as actual, the proposition will be false there.

While many a posteriori physicalists have invoked Kripkean or weak metaphysical necessity to explain why zombies are logically but not metaphysically possible, Chalmers argues that weak metaphysical necessity has in fact nothing to do with the relevant link between logical and metaphysical possibility in question, i.e. with the link between logical and metaphysical possibility o f *worlds*. If Chalmers is right, then weak metaphysical necessity cannot help the a posteriori physicalists rule out the metaphysical possibility of a zombie world which, as they allow, is conceivable or logically possible.[139]

The problem with weak metaphysical necessity, Chalmers suggests, is that considerations about this type of necessity do not show us that any logically possible world is metaphysically impossible. They merely show us that certain descriptions of possible worlds, are in fact misdescriptions if these worlds are considered as counterfactual. The mistake which leads to these misdescriptions consists, according to Chalmers, in applying our concepts in virtue of their primary intensions rather than in virtue of their secondary intensions, as would be appropriate when we evaluate the reference of our concepts in possible worlds considered as counterfactual.

Another way to express this point is to say that given that the Kripkean identities express weak metaphysical necessity, they are compatible with being false in some worlds considered as actual. As we saw, after all, the primary intensions of the relevant statements are contingent. The Kripkean necessity "Water is H_2O" is, for example compatible with being false in the XYZ world if that is considered as actual and if thus the proposition is evaluated with respect to its primary intension. This, however, means that these necessities do not imply that some logically possible worlds (e.g. the XYZ world), are in fact metaphysically impossible.

Things are different when it comes to the supposed strong metaphysical necessity. If statements such as "Pain is C-fibres firing", express strong metaphysical necessity, then both the primary intension and the secondary intension of this statement must be necessary. There then cannot be a possible world considered as actual or as counterfactual in which the statement would be false. While such a world is conceivable, given what we know a priori, and it is therefore (primarily) logically possible, it is simply not (primarily or secondarily) metaphysically possible.

If then there is strong metaphysical necessity, some logically possible worlds are not metaphysically possible. This means that even though a particular possible world is conceivable for an ideal

[139]The physicalist who claims that a zombie world is not even logically possible or conceivable is, as we saw, not an a posteriori physicalist but rather an a priori physicalist.

reasoner, the world may still not be metaphysically possible, neither in the primary nor in the secondary sense. If there is strong metaphysical necessity, there is then, apart from the space of logically possible worlds, a separate space of metaphysically possible worlds where these correspond to some but not all logically possible worlds. The existence of strong necessity then means that apart from a priori constraints which apply to both the space of logically possible worlds and the space of metaphysically possible worlds, there are certain a posteriori constraints which apply merely to the space of metaphysically possible worlds. If there is strong metaphysical necessity, then the step from an epistemic gap to the ontological gap, or – in other words – the step from the logical possibility of a zombie world to its metaphysical possibility is, at least in some cases for reasons unavailable to even ideal a priori reflection, blocked and the conceivability argument is unsound.

According to Chalmers, the doctrine of strong necessity is unjustified and there are good reasons to reject it.[140] He observes, for example, that the existence of strong metaphysical necessity implies that there is yet another kind of modality apart from *logical* possibility and necessity and *natural* or *physical* possibility and necessity. Namely, there must then also be *metaphysical* possibility and necessity. The zombie world, for example, would then be, for an a posteriori physicalist, logically possible although physically and metaphysically impossible. Chalmers thinks that there is no reason to believe that such a separate metaphysical modality exists.[141] His view is that any logically possible world is also metaphysically possible, while it may not be physically possible, given the laws of nature which exist in the actual world. Using a theological metaphor, he argues that any logically possible world could have been created by God and is therefore metaphysically possible.[142]

What's more, claiming that the additional modality of strong metaphysical necessity is viable is, Chalmers thinks, clearly *ad hoc* since, as he puts it, "the only motivation for this view would seem to be to save physicalism at all costs".[143] There is, he emphasises, no other phenomenon in the universe which would motivate and support such "proliferation of modalities", since nothing else seems to provide a reason to believe in strong metaphysical necessity. Moreover, Chalmers argues, the believer in strong metaphysical necessity could never *know* that some worlds which are logically possible are metaphysically impossible since neither a priori considerations nor a posteriori considerations could tell us that.[144] Our a priori considerations, of course, tell us about the *logical* possibility and impossibility of worlds while our a posteriori findings and considerations only tell us about the state of the actual world (and about the laws of nature which place constraints

[140]Chalmers (1996, pp. 136-138).
[141]Chalmers (1996, p. 137).
[142]Chalmers (1996, p. 138).
[143]Chalmers (1996, p. 138).
[144]Chalmers (1996, p. 137).

on the space of physically possible worlds). Neither kind of considerations could, therefore, place constraints on the space of metaphysically possible worlds.

Why, however, should one think that Loar is committed to the existence of strong metaphysical necessity? The issue is fairly straightforward: he agrees with the anti-physicalists that zombies are conceivable or logically possible but – being a physicalist – he cannot accept that a zombie world is metaphysically possible (since that would, as shown above, imply that consciousness is a metaphysical extra added to the material world). That means, however, that he must hold that some logically possible worlds are not metaphysically possible which is, as we saw, precisely what the doctrine of strong metaphysical necessity amounts to.

The problem with Loar's view then is, according to Chalmers, that Loar simply assumes the correctness of the controversial doctrine of strong necessity without offering any justification for it.[145] While Loar manages to account for the logical possibility of zombies in terms of the conceptual independence between phenomenal and physical concepts, he fails, according to Chalmers, to explain and justify that despite its apparent conceivability, the zombie world is, due to strong metaphysical necessity, metaphysically impossible.

It may seem as if Loar is trying to provide us with just such an explanation. As we saw, after all, Loar thinks that phenomenal and physical concepts corefer. If they do, then the two members of each physical-phenomenal conceptual pair (e.g. *firing of C-fibres* and *pain*) refer to the same kind of state and, given that the physical concept refers to a physical state, then so does the phenomenal concept. The coreference claim then implies that the phenomenal state and the physical state are actually identical. That, however, means that in no metaphysically possible world could the physical state exist without being phenomenal. We can see this if we notice that were the coreference claim true, a zombie world would clearly be secondarily impossible since as we saw, both concepts designate rigidly. The coreference claim, moreover, implies also a primary impossibility of the zombie world. We can see this if we realise that there could not be a possible world which would include a property that would have the appearance of consciousness but had a non-physical nature. If such a world was considered as actual, then, someone could suggest, consciousness would be non-physical in it and it would thus be primarily possible (although not secondarily possible) that consciousness is non-physical. The existence of such a world is, however, incompatible with the coreference claim given that, as we saw, a property which appears like consciousness in all respects, simply *is* consciousness. If we then grant Loar the coreference claim, primary conceivability of the zombie world is compatible with its both primary and secondary impossibility, which means that there must be strong metaphysical necessity.

[145]Chalmers (1999, p. 489).

The trouble with this kind of explanation is, of course, that Loar cannot simply assume the coreference claim as this claim is disputed and requires justification. How could Loar argue for the coreference claim? He could appeal to his view that phenomenal concepts are recognitional concepts of a kind and that recognitional and physical concepts typically corefer. The kinds which our recognitional concepts enable us to pick out can, as we saw, also be picked out via physical concepts so there will be many pairs of physical and recognitional concepts which corefer. Given that, should one not expect that phenomenal concepts (being, on Loar's view, themselves recognitional) and physical concepts corefer as well?

As I see it, this reply presupposes that property dualism is false. If, after all, property dualism were true, then our recognitional phenomenal concepts would pick out different properties than our theoretical physical concepts. On such a view then we would have no theoretical concepts of non-physical phenomenal properties, only, given that Loar is right, recognitional concepts of those. Neither would we have any special recognitional concepts of our physical brain properties (those concepts which Loar identifies with phenomenal concepts), only physical theoretical concepts of these properties. Given that all that Loar says is compatible with dualism on which physical and phenomenal concepts do not corefer, it is clear that he does not really justify the coreference claim but rather, at best, explains how physical-phenomenal coreference could work.

Even this more modest project of explaining how physical and phenomenal concepts corefer is, however, as I see it, ultimately unsuccessful. This is well expressed by Chalmers who suggests that Loar's explanation of coreference is undermined by his direct reference claim, i.e. his claim that phenomenal concepts pick out their referents directly. Given this direct reference claim, phenomenal concepts have a peculiar and exceptional status among recognitional concepts most of which, as we saw, refer via contingent modes of presentation.[146] As a result, it may seem that the explanatory model involving the paradigm recognitional concepts is ill-suited for the task of explaining the coreference of physical-phenomenal conceptual pairs.

The problem with this direct reference claim is, Chalmers thinks, not merely that it renders phenomenal concepts an exception among recognitional concepts. A deeper difficulty lies, Chalmers suggests, in the fact that the very feature which explains coreference of conceptual pairs of standard, or paradigm recognitional concepts (i.e. those which refer via contingent modes of presentation) and physical concepts is missing in the cases of conceptual pairs of phenomenal and physical concepts. This crucial feature is, according to Chalmers, the indirect manner of reference of paradigm recognitional concepts. It is precisely this referential indirectness which explains how physical and standard recognitional concepts can corefer despite their cognitive distinctness.

[146]Chalmers (1999, p. 488).

Chalmers writes:

> *[...] this coreference is explained by the two-dimensional nature of such recognitional concepts: they typically conceive of their referent as "the cause of such-and-such experience", or under some similar contingent mode of presentation. If we remove this feature of recognitional concepts [...], we no longer have any reason to believe that recognitional concepts and distinct theoretical concepts should corefer.*[147]

This passage neatly captures what I take to be the core of Chalmers's worry about Loar's view, namely that we can explain how recognitional concepts can (a) pick out the same referents as some physical concepts and yet (b) be cognitively quite distinct from these physical concepts only if the recognitional concepts involved refer via contingent modes of presentation. The contingent mode of presentation of a recognitional concept is, after all, cognitively different from the purely physical content of the coreferring physical concept (as for (b)), yet via this non-physical mode of presentation the concept picks out a purely physical referent – a physical kind (as for (a)).[148]

We can see that the situation is quite different in the case of pairs of directly referring concepts which supposedly corefer, such as, if Loar is right, conceptual pairs consisting of a phenomenal and a corresponding physical concept. In these cases we have, Chalmers suggests, no way of explaining why the two cognitively different concepts corefer. Consider, for example, the phenomenal concept "pain" and the physical concept "the firing of C-fibres". The physicalist and his opponent agree that both concepts refer directly which means that they do not pick out their referents via contingent modes of presentation. If, however, the concept of pain lacks a contingent mode of presentation, it *ipso facto* lacks the two-dimensional semantic structure which, as we saw, explains how a paradigm recognitional concept can have a different cognitive content from the corresponding physical concept and yet have the same referent as that concept. With both the physical and the phenomenal concept referring directly, both concepts refer to what they express, and the properties which they express certainly seem different enough. Chalmers concludes that Loar's view provides us with no resources to explain or justify how the two concepts could corefer. Loar can of course, Chalmers notices, simply insist that they corefer but that may seem dogmatic and question-begging. If, however, the coreference claim is unjustified and unexplained, then so is the claim that a zombie world is metaphysically impossible while being logically possible. As we saw, after all, the latter

[147]Chalmers (1999, p. 488).

[148]Another way to express the same point is to say that while pairs of coreferring paradigm recognitional and physical concepts have identical secondary intensions (a particular physical kind) their primary intensions are quite different – the primary intension of the recognitional concept picks out, plausibly, the cause of such and such sensory experience (which happens to be a particular physical kind in the actual world), while the primary intension of the coreferring physical concept picks out, just like its secondary intension, such-and-such physical kind. The difference in primary intensions explains why the corresponding identity statements are not a priori while the identity of secondary intensions explains why the two concepts corefer.

was meant to be justified and explained by the former. Loar therefore, if Chalmers is right, offers us no reason to think that there exists strong metaphysical necessity.[149] In that case, however, Loar's a posteriori physicalism seems to be under serious threat.

5. Loar and the Conceivability-Possibility Link

I shall now suggest a possible line of defense of Loar's view against Chalmers's objection. Let me first say that I agree with Chalmers's claim that Loar does not justify the coreference claim which is clearly an integral part of his position. Clearly, this would be a problem for Loar if it were his strategy to offer a direct argument for physicalism. One could, however, also view Loar's suggestion in a different light. Namely, one can read Loar as arguing for materialism indirectly by showing that the common challenges to it, such as the conceivability argument or the knowledge argument needn't worry the physicalists. Both of these arguments, as we saw, draw ontological, anti-physicalist conclusions from their epistemic premises. As I see it, Loar tries to show that if we adopt a certain view of the conceptual furniture of our minds, a view which seems *prima facie* plausible, then we do not need to draw these ontological conclusions. If this view is correct, after all, then the conceivability of zombies does not imply their metaphysical possibility and the conceivability claim is thus fully compatible with a materialist picture of the world. While Chalmers argues then – and I think correctly – that Loar does not justify his denial of the metaphysical possibility of zombies, what Loar does instead is try to cast doubt on the supposed implication between the conceivability of zombies and their metaphysical possibility. He does that by sketching out a view of the universe in which zombies are conceivable but not metaphysically possible. The epistemic gap is then, on this view, merely an interesting symptom of the way we think about consciousness, rather than a reliable guide to metaphysics. It is true that Loar doesn't really argue for strong metaphysical necessity, but one could say in his defence that he does not need to do so in order to achieve his goal which is, as I see it, casting doubt on the link from logical possibility to metaphysical possibility and thus on the plausibility of the anti-physicalist arguments and the step from the epistemic gap to the ontological gap. This then is a possible line of defence for Loar against Chalmers's criticism.

Does Loar succeed in this negative task? To see why not, recall Loar's reason for casting doubt on the link from the epistemic gap to the ontological gap or, in the key of conceivability, from the (ideal) conceivability or logical possibility of zombies to their metaphysical possibility. As we saw, for Loar the reason is the fact that if his sort of a posteriori physicalism is true, then zombies are logically possible but not metaphysically possible. This reasoning, however, bites back. If, after all,

[149]Chalmers (1999, p. 489).

zombies are logically possible but not metaphysically possible, as Loar would have it, then there must be the dubious strong metaphysical necessity, critiqued by Chalmers. Given, however, that the doctrine of strong metaphysical necessity looks, as we saw, dubious, then, by *modus tollens*, so is Loar's view. Of course, as I tried to show, Loar tries to justify or at least explain why there is strong necessity in the consciousness case, but if the above-introduced reasoning is correct his explanation ultimately fails.

Is there any other way in which the a posteriori physicalists could try to justify their reliance on strong necessity? One strategy here would be to attempt to offer examples of strong necessities outside the mind-body case. While this is not a strategy Loar pursues, this sort of defence of a posteriori physicalism is still worth mentioning here. Here the materialists could appeal, for example, to the views of philosophers, such as Alexander Bird or Sydney Shoemaker, who reject the contingency of the laws of nature and argue that these laws hold with metaphysical necessity.[150] The physicalists could appeal to this conception and argue that if this view is correct, then, arguably, while one can conceive of worlds which are nomologically different from the actual one, and as a result, such worlds are logically possible, these worlds are in fact not metaphysically possible.

As I see it, however, the appeal to the supposed metaphysical necessity of the laws of nature can hardly help the physicalist here. We can see this if we consider what Bird and others mean by claiming that laws of nature hold with metaphysical necessity. Laws of nature for them amount to dispositions of things to behave in a certain way in the presence of a particular stimulus.[151] It is natural, however, to think that dispositions of a thing are rooted in what the thing is, which suggests that a thing has its dispositions necessarily. Bird offers us an example of salt.[152] He argues that in order for salt to exist, Coulomb's law needs to exist as this law governs the electrostatic attraction which is a condition of the existence of salt. The same law, however, ensures that salt dissolves in water. Supposing that Bird is right about this, then, clearly, the lawful regularity, which amounts, roughly, to the fact that salt dissolves in water, is not a contingent but rather a metaphysically necessary fact about salt.

It is not difficult to see that Bird's dispositionalist conception of the laws of nature will be of little help to the physicalist. Even if we, after all, suppose that this conception is correct, it is clear that it does not justify the claim that some logically possible worlds are not metaphysically possible, i.e. it does not justify strong metaphysical necessity. It shows us at most that things have their dispositions essentially. That however, is compatible with the claim that there are other possible worlds in which things have different dispositions – only we could not, given that Bird's considerations are correct,

[150]See Bird (2005), Shoemaker (1980). See e.g. Armstrong (1983) for the contingency conception of the laws of nature.
[151]Bird (2005, p. 354–355).
[152]Bird (2005, p. 364).

describe these possible worlds as worlds which include the same things as our world, only with different dispositions. Imagine, for example a possible world in which the stuff which otherwise (more or less) corresponds to salt in our world (it has the typical taste, colour, etc.) does not dissolve in water. While nothing Bird says shows us that such a world could not exist, if Bird's conception is correct, then it would be wrong to call this salt-like stuff "salt" and it would be equally wrong to say that salt is governed by different laws of nature in that world. That, however, by no means offers support to the existence of strong necessities. It does not, after all, rule out the existence of any metaphysically possible world which is logically possible, all it shows us is that some descriptions of possible worlds are in fact misdescriptions. We can conclude then, that the dispositionalist account of the laws of nature, which conceives of these laws as metaphysically necessary, will be of no help to the a posteriori physicalists.

6. Conclusion

In this chapter I have started my discussion of a posteriori physicalism which will continue in the following chapter. A posteriori physicalism is currently perhaps the most popular variety of physicalism with respect to consciousness as it tries to show how a physicalist can rationally live with the epistemic gap. My discussion focused, almost exclusively, on the variety of a posteriori physicalism called the phenomenal concept strategy. While this approach promises to provide us with reasons to reject the step from the epistemic gap to the ontological gap, as I tried to demonstrate, using the example of the view of Brian Loar, one of the founders of the phenomenal concept strategy, there are reasons to think that this promise is not quite fulfilled. Appealing to the critique of Loar's view which has been advanced by Chalmers, I argued that while Loar provides us with an interesting explanation of why zombies are conceivable, he fails to properly justify his claim that zombies are not metaphysically possible, which is of course crucial for the defence of a posteriori physicalism. Another way to express this point is to say that Loar relies on the existence of strong metaphysical necessity which he fails to properly justify. Namely, I argued that Loar's claim that phenomenal concepts are cognitively distinct from physical concepts, together with the Kripkean view, accepted by Loar, that phenomenal concepts pick out their referents directly, i.e. not via contingent modes of presentation, render the coreference of physical and phenomenal concepts, on which Loar's claim that zombies are metaphysically impossible rests, quite unjustified.

I tried to show how Loar's view could be defended against Chalmers's critique by suggesting a more charitable way of understanding Loar's project. On this understanding, Loar does not try to justify the metaphysical impossibility of zombies, but rather question the step from their conceivability to

their metaphysical possibility. One could, after all, argue that if a posteriori physicalism is true, this step needs to be rejected. While this way of reading Loar may seem to render his position plausible, I argue that even on this reading Loar's project ultimately fails as this reply to the zombie argument relies on the unexplained and unjustified existence of strong metaphysical necessity. I conclude the chapter by considering another way in which the a posteriori physicalists could try to reply to the above-mentioned objection that their view relies on strong metaphysical necessity. This reply appeals to the notion that the laws of nature are metaphysically necessary, even if nomologically different worlds are conceivable. I suggested that this objection ultimately fails as the metaphysical necessity in question fails to establish that some logically possible worlds are not metaphysically possible. It seems reasonable to conclude that it is at least a serious challenge for the a posteriori physicalist to account for and justify the highly dubious existence of strong metaphysical necessity.

4. What Do Our Phenomenal Concepts Reveal to Us?

1. A Posteriori Physicalism and Opacity of Phenomenal Concepts

In the previous chapter I argued that the doctrine of a posteriori physicalism, which is currently perhaps the most popular variety of physicalism, brings along the uncomfortable metaphysical baggage in the form of a commitment to the existence of strong metaphysical necessity which, I attempted to show, is difficult, if not impossible to justify or account for. In this chapter I shall focus on another critique of a posteriori physicalism. This critique, versions of which have been advanced by thinkers such as Joseph Levine, Martine Nida-Rümelin, David Chalmers and Philip Goff, is based on the thought that some of our phenomenal concepts, namely those by means of which we conceive of our conscious states in terms of what it is like to be in them, provide us with rich and substantial knowledge of the nature of their referents.[153] The existence and nature of this phenomenal knowledge is, according to these critics, incompatible with a posteriori physicalism.[154]

The critique of a posteriori physicalism which I shall focus on here has not received as much attention in the literature as the knowledge argument or the conceivability argument, which I introduced in the previous chapters, but is still, as I shall argue, a serious challenge for physicalism. In the literature, this phenomenal knowledge, supposedly revealed to us by our phenomenal concepts, has been characterised in various ways. Nida-Rümelin, for example, writes:

> *If you have a phenomenal concept of a phenomenal property, then you know what it is to have an experience with that subjective feel. You thereby know what it is to have that property: you grasp the phenomenal property via your phenomenal concept.*[155]

Joseph Levine, another proponent of this kind of critique, writes in a passage concerning phenomenal thought:

> *The first-person access we have to the properties of experience seems quite rich; we are afforded a very substantive and determinate conception of a reddish experience merely by having it.*[156]

Despite certain differences, both of these passages express the idea that our phenomenal concepts

[153]See e.g. Nida-Rümelin (2007b), Levine (2007), Goff (2011).
[154]In what follows I shall thus concentrate on what Chalmers calls *pure phenomenal concepts*, i.e. those phenomenal concepts by means of which we conceive of our phenomenal states with respect to what it is like to be in these states (Chalmers [2003]). These concepts then enable us to conceive, for example, of pain in terms of what it is like for a subject to be in pain. In particular, I shall be interested in those pure phenomenal concepts which we are able to form solely by attending to a currently experienced phenomenal state. Chalmers calls phenomenal concepts of this sort direct phenomenal concepts (ibid.).
[155]Nida-Rümelin (2007b, p. 307).
[156]Levine (2007, p. 163).

provide us with a way of understanding their referents – phenomenal states. How far this understanding reaches and whether this understanding poses a challenge for a posteriori physicalism are the questions I shall tackle in this chapter.

I shall first focus here on how this type of critique of a posteriori physicalism has been developed by Philip Goff, who argues for the thesis of translucency of phenomenal concepts. Thereafter, I shall discuss the view of Martine Nida-Rümelin who argues that we grasp our phenomenal properties via our phenomenal concepts. Finally, I shall try to use some of the lessons learnt in these two discussions in developing an argument against a posteriori physicalism.

The translucency thesis defended by Goff tells us that our phenomenal concepts a priori reveal to us some knowledge of the nature of their referents, our phenomenal states.[157] This thesis, however, is incompatible, Goff argues, with a posteriori physicalism. How are we to understand the claim that phenomenal concepts a priori reveal to us knowledge of their referents? In order to understand this claim, it will help to start with Goff's distinction between *opaque*, *translucent* and *transparent* concepts. This distinction concerns concepts in general, not merely phenomenal concepts. According to Goff, a concept is opaque iff it does not reveal to us a priori anything non-trivial about the nature of its referent. Applying this general thought to the concepts of properties, Goff arrives at the following definition:

> *[...] a concept C of a property F is opaque iff C reveals nothing of what it is (or what it would be) for an object to have F.*[158]

Consider, for example, the concept *David's favourite shape*[159] and suppose that David is fond of sphericity. It is arguable that the concept *David's favourite shape* does not reveal anything essential about its referent, the spherical shape. I may then have mastered an opaque concept without knowing a priori anything non-trivial about what the instantiation of its referent consists in, i.e. without, in this case, knowing that it consists in being spherical.

If, on the contrary, we have a concept which is not opaque, the concept provides us a priori, merely in virtue of the fact that we have it, with at least some non-trivial knowledge of its referent, i.e. knowledge about what the instantiation of its referent consists in. According to Goff, non-opaque concepts can be divided into two distinct categories – they are either transparent, or they are merely translucent. Transparent concepts are understood by Goff as those concepts which (often implicitly) provide us with knowledge of the complete essential nature of their referents, i.e. they provide us a

[157]Goff (2011).
[158]Goff (2011, p. 192).
[159]In this chapter I shall italicise individual words and short expressions which refer to specific concepts.

priori with complete knowledge of what it is for their referents to be instantiated.[160] Translucent concepts, on the other hand, provide us a priori with knowledge of merely a part of the essential nature of their referents, or – we can say – they tell us a priori merely something (not everything) about what it is for the referents of these concepts to be instantiated.

Among the examples of transparent concepts provided by Goff are concepts, such as *sphericity in Euclidean geometry, friend* or *party*. Anyone who, for example, has mastered the concept *sphericity in Euclidean geometry* will know a priori that a thing is spherical in Euclidean geometry iff all points of its surface are equidistant from its center. An example of a concept which is plausibly translucent, although not transparent, is the compound concept *being of the same height as John* which has a transparent part – *being of the same height as –,* and an opaque part – John.

It is important to be clear about what Goff means when he writes that a given transparent or translucent concept reveals to us some knowledge a priori. What I think Goff has in mind here is that once I have acquired the concept, for example the concept *sphericity*, I shall not need any more empirical knowledge in order to know what a given shape must be like to count as an instance of sphericity. This is, of course, compatible with the view that I may sometimes need empirical knowledge to acquire the given transparent or translucent concept in the first place. Presumably, for example, one needs certain kinds of experiences to acquire concepts such as *party* or *friend.*

The notion of an opaque concept is of key importance for Goff's argument against a posteriori physicalism. This argument can be sketched as follows:

(1) Careful armchair reflection on the semantic intuitions I have about my phenomenal concepts tells me that phenomenal concepts are not opaque.

(2) A posteriori physicalists are committed to the view that phenomenal concepts are opaque.

———

(3) A posteriori physicalists are committed to a counterintuitive view of phenomenal concepts.

This argument looks valid, although it is not clear whether it is also sound. Let me now discuss its premises one by one.

Premise (1) amounts to the rejection of the opaque view of phenomenal concepts. Why think then that phenomenal concepts are non-opaque? Here Goff notices that we can refer to the property of feeling pain using the concept *being in pain* but also, for example using the concept *the property Kevin is thinking about*, given that Kevin is thinking about the property of feeling pain. The concept

[160]Goff (2011, p. 194).

the property Kevin is thinking about is, as Goff emphasises, opaque and so it does not reveal to us a priori anything non-trivial about its referent, i.e. the property of feeling pain. Things seem significantly different when it comes to the phenomenal concept *feeling pain.* If someone tells me, for example, that Pete is feeling pain, I shall, it seems, know a priori quite a lot about the unpleasant state to which the concept *feeling pain* refers and I shall perhaps, supposing that I care about how Pete is feeling, offer him a pain killer, etc. This, Goff argues, indicates that the phenomenal concept *feeling pain* is not opaque as it seems to provide us a priori with non-trivial knowledge of its referent, i.e. of the property of feeling pain. Why think that the relevant phenomenal knowledge is revealed to me a priori by the concept? Here the thought is that intuitively, everyone who has the concept of pain will not need any additional empirical knowledge in order to know what kind of conscious state Pete happens to be in.

Let me now mention another consideration which speaks against the opaque view of phenomenal concepts and, *ipso facto,* supports premise (1). This consideration is based on what we can call the *conceivability test* and should lead us to the thought that claims such as "pain is unpleasant", "pain is an unpleasant feeling" or "a mental state which feels like pain is pain" are a priori. Arguably, nobody who has the concept *pain* (and relevant other concepts) will need, as the conceivability test indicates, any additional knowledge in order to know these truths. The claim that these statements are a priori is supported by the fact that their negations will fail in the conceivability test, i.e. they are inconceivable. Here the thought is that, for any statement *S*, if ~*S* is inconceivable, then *S* is a priori.[161]

Is it, for example, conceivable that someone is in pain but the pain is not unpleasant? I do not think such a scenario is conceivable and so it seems that the knowledge that pain is unpleasant is revealed to us a priori by the concept of pain. Here, of course, someone could raise the objection that, for example, for body builders doing weight-lifting training, the muscle pain which follows will not be unpleasant as it will mean for them that the training went as it should have.[162] A critic can also appeal here to a monk who in the name of penitence would put on a hairshirt or perhaps even a masochist who gets sexual pleasure from pain. It seems to me, however, that one can reply to this objection that the mentioned cases to not really indicate that there could be pain which is not unpleasant. All it seems to show is that our experience is complex and unpleasant states can, under certain circumstances bring us pleasure or satisfaction. As a result, the body builder can enjoy just the unpleasant feeling in the muscles, the monk would behave in this way precisely because pain is unpleasant and masochists gets satisfaction precisely from the fact that they are suffering, i.e. experiencing unpleasant states. It seems to me, therefore, that this objection does not threaten the

[161]More precisely, I have here in mind primary ideal conceivability, a notion I introduced in the previous chapter.
[162]I would like to thank James Hill for pressing this point.

claim that the statement “pain is unpleasant” is a priori.

Compare this with the statement “this state is not unpleasant”. Such a statement may, of course, be false but we will, arguably, not know this a priori merely by analysing, however carefully, the demonstrative concept *this state*. Similarly, let us run the conceivability test on the negation of the statement that a mental state which feels like pain really is pain. Even here it seems that the negation, i.e. the statement that there are mental states which feel like pain but which are not really pain, will not pass the conceivability test. I think we would take the fact that a speaker utters such a claim to indicate that the speaker does not understand the concept *pain*. It seems reasonable, therefore, to view the original item of knowledge as a priori revealed to us by the concept. Similar considerations apply to other phenomenal concepts. Consider, for example, the statement that the experience of bright red is more similar to the experience of orange than to the experience of blue. This is another claim whose negation (the statement that red experience is more similar to a blue experience than to an orange experience) fails the conceivability test and so the original statement seems to be a priori entailed by the given concept. Once again, we would tend to think that someone who expressed the negated statement does not understand at least one of the concepts involved.

Finally, one could argue for the view that phenomenal concepts provide us a priori with knowledge of their referents by appealing to the hypothetical case of Jackson's Mary. Imagine the situation that Mary, upon her release, sees red for the first time. What is it that happens to Mary, according to the a posteriori physicalists? Most of them will presumably say that she acquires a new concept, namely the phenomenal concept of phenomenal red, by means of which she newly conceives of the physical state which is identical with her phenomenal state and which is already known by her under the physical concept. Intuitively, the new phenomenal concept will provide her with much, or at least some, knowledge of its referent, the phenomenally red state. Supposing that, however, the opaque view of phenomenal concepts looks quite unattractive.

I hope that these considerations – Goff's attempt to contrast phenomenal concepts with concepts which are clearly opaque, my consideration working with the conceivability test and the appeal to Mary's newly acquired concept – sufficiently justify premise (1) of Goff's argument.

Premise (2) of Goff's argument tells us that a posteriori physicalists are committed to opacity of phenomenal concepts. This premise is supported by the consideration that if phenomenal concepts were not opaque and, at the same time, a posteriori physicalism was true, then phenomenal concepts would need to reveal to us, partly or perhaps even completely, the physical or functional nature of their referents.[163] The concept *feeling pain* would then, argues Goff, need to provide us with at least

[163]Goff (2011, p. 196).

a partial understanding of the physical (or functional) nature of its referent, roughly, the property of C-fibres firing in one's brain (or perhaps of the particular functional role which is, according to the functionalists, the essence of pain).[164] That, however, is, according to Goff, precisely what the a posteriori physicalists deny – if after all, phenomenal concepts provided us a priori with knowledge of the physical (or functional) nature of their referents, then phenomenal truths would arguably be entailed by physical truths and a posteriori physicalism would collapse into a priori physicalism as there would be no epistemic gap. A posteriori physicalists are therefore, according to Goff, committed to the view that phenomenal concepts are opaque, which is in conflict with our semantic intuitions about phenomenal concepts.

If we thus accept premise (1) which tells us, in effect, that according to our epistemic intuitions, our phenomenal concepts are non-opaque, as well as premise (2) which states that a posteriori physicalism entails opacity of phenomenal concepts, we get the conclusion (3) that a posteriori physicalism brings about a counterintuitive view of phenomenal concepts.

Is there a way for the a posteriori physicalists to reply to Goff's argument? Here one option is for them to bite the bullet and embrace the view that perhaps we should just accept that some of our semantic intuitions about our phenomenal concepts, are misleading. Such a reply, as I see it, amounts to simply accepting the conclusion of the argument. This kind of view has been suggested by David Papineau who explicitly endorses the opacity of our phenomenal concepts.[165] Papineau allows that we have certain intuitions, according to which our phenomenal concepts provide us with an insight into the true non-physical nature of their referents – phenomenal states. There are, however, good reasons, he thinks, not to place too much weight on these intuitions.[166] According to Papineau, our phenomenal concepts do not in fact a priori reveal to us anything about their referents, which he think is due to the fact that these concepts are primitive or atomic, i.e. they are not composed of other, more primitive concepts. Papineau thus allows that some concepts really do a priori reveal to us information about their referents, but this is only a matter of composed concepts.[167] The notion of transparency or translucency of our phenomenal concepts is thus, in effect, if Papineau's view is correct, a mere illusion which needs to be rejected.[168]

Still, as far as I know, Papineau does not really justify his claim that phenomenal concepts are opaque and it certainly does not seem obvious that atomic concepts could not reveal anything about their referents to their possessors. Moreover, to accept the thesis of the opacity of phenomenal

[164]The neuronal correlate of pain is in fact more complex, but here I shall for the sake of simplicity stick to the traditional philosophical usage.
[165]Papineau (2006, p. 106).
[166]Papineau (2006, p. 102).
[167]Papineau (2006, ibid.).
[168]Papineau also offers us an explanation of our intuitive view that phenomenal concepts are transparent. This view, he suggests, results from the “use-mention” feature of our phenomenal concepts. See Papineau (2007, pp. 123-124).

concepts held by Papineau means to give up our deeply held semantic intuitions about phenomenal concepts, which I tried to espouse above. Such a step, however, arguably means undermining the motivation behind a posteriori physicalism. If a posteriori physicalism fails to accommodate our intuitions about phenomenal concepts, why not embrace a priori physicalism instead? The conclusion of Goff's argument thus, as I see it, should in the end worry the a posteriori physicalists. Let me therefore ask now whether they can reject one of the premises of the argument.

Some a posteriori physicalists have attempted to do just that, arguing, in effect, against premise (2) which expresses the commitment on the side of a posteriori physicalists to the opacity of phenomenal concepts. Namely, some a posteriori physicalists have offered theories of phenomenal concepts which deny their opacity and thereby, arguably, accommodate the intuitive view that our concepts of phenomenal states reveal to us at least something about the nature of their referents. According to Janet Levin, for example, we conceive of our phenomenal states via *hybrid concepts* which have both a type-demonstrative component and a descriptive component where the descriptive component provides us with a priori knowledge of these concepts' referents, phenomenal states.[169] Levine thus, we can say, adopts the type-demonstrative account embraced by Brian Loar, discussed in the last chapter, but complements it with the view that phenomenal concepts also have a descriptive part which is available to us a priori. According to Levin, we thus conceive, for example, of the phenomenally red experience demonstratively as "this type of inner state" and at the same time descriptively – roughly, as "the type of inner state which is phenomenally very similar to an orange experience, much less similar to a blue experience, etc.".[170] The descriptive aspect of the concept thus provides us with information about the unique place of the state in a sensory space of qualities, as well as with information about what kinds of emotional reactions the state can trigger.

The given phenomenal concept has thus, according to Levin, two conditions when it comes to its reference – it refers, to put things simply, to the type of neural state which (1) triggers the application of this type of phenomenal concept and (2) holds an appropriate position in the system of functional states, inputs and outputs – it is the type of state which is capable of causally leading to relevant judgements about similarities and differences of phenomenal states and to the corresponding emotional reactions. Here (1) and (2) express the conditions which a brain state must satisfy in order to be the referent of a given phenomenal concept. We can see that while condition (2) is a priori available to the concept user, what type of state fulfills condition (1) is outside the sphere of the a priori and can only be discovered empirically.

[169]Levin (2002, p. 587).
[170]Levin (2002, p. 583).

Thanks to the demonstrative component, which Levin attributes to phenomenal concepts, these concepts are irreducible to purely functional concepts, so her view is compatible with the existence of the epistemic gap. Owing to their descriptive component, Levin's account is, on the other hand, compatible with the intuition espoused earlier that our phenomenal concepts provide us with some knowledge of their referents, i.e. they are not opaque. We can see that the knowledge which we are able to a priori arrive at is compatible with a posteriori physicalism: we learn, after all, about functional connectedness of the given state within a system of mental states and there is no problem in principle with these states being physically realisable. The variety of a posteriori physicalism held by Levin thus offers us a counterexample with respect to premise (2) of Goff's argument. Our phenomenal concepts are thus, according to Levin, by no means opaque, they in fact offer us relational knowledge of their referents.

Another attempt to render a posteriori physicalism compatible with non-opacity of phenomenal concepts, has been advanced by Robert Schroer.[171] Schroer, like Levin, proposes a hybrid view of phenomenal concepts but rejects Levin's claim that phenomenal concepts provide us with only relational knowledge of phenomenal states.[172] This claim, Schroer emphasises, is in conflict with the intuitive view that we have rich knowledge of the intrinsic nature of our phenomenal states, the claim taken, as we saw, very seriously by the anti-physicalists.[173] It is this intuition which Schroer tries to accommodate in his own view which is based on the understanding of phenomenal states as structurally complex entities.

Schroer views phenomenal states – and here his view contrasts with Papineau's view mentioned above – as composed out of many phenomenal elements organised in particular ratios. This can be illustrated by an appeal to the situation in which we get a particular shade of paint by mixing various amounts of paints of primary colours. Similarly, our experiences of colour, as well as other phenomenal states, are, according to Schroer, composed from a number of primitive components in various amounts. The ultimate simple elements of colour experiences are, however, according to Schroer, not phenomenal colours themselves but rather their elements such as hue, brightness and saturation, or perhaps even simpler elements, such as what he calls “strength” or “warmth”.[174] It is this inner, compositional complexity of phenomenal states which, according to Schroer, our phenomenal concepts capture. Phenomenal concepts then, according to Schroer, have the form of descriptions such as: “the quality with such-and-such level of *this element* and such-and-such level of *that element*, etc.”[175] The expressions which are printed in italics in this description represent the

[171]Schroer (2010).
[172]Schroer (2010, p. 512).
[173]See e.g. Chalmers (1996, p. 235).
[174]Schroer (2010, p. 515).
[175]Schroer (2010, p. 517).

type-demonstrative components of phenomenal concepts while the instances of the phrase “such-and-such” stand for particular amounts of the primitive elements, which the concept expresses, for example, as percentages.

According to Schroer's view, once again, the demonstrative components of the given concept account for the fact that we encounter an epistemic gap (since, as Loar taught us, demonstrative truths are not entailed by physical truths). The referents of these (type-)demonstratives, i.e. the primitive components of our phenomenal states are, after all, not a priori specified any closer, they are merely pointed to by the demonstrative elements of phenomenal concepts. The content of phenomenal concepts which is available to a priori reflection is, nevertheless, according to Schroer, rich as it includes information about the amounts or levels of the primitive phenomenal elements present in the resulting phenomenal state.

This a priori available knowledge, moreover, does not concern – unlike in Levin's view – only relations between various phenomenal states – but rather these phenomenal states themselves and thus, in a sense, the inner or intrinsic natures of these states. Schroer's view is also compatible with the ontological commitments of a priori physicalism as the demonstrative elements of the descriptions, which he thinks our phenomenal concepts reveal to us, refer, according to his view, to brain states. Given that Schroer's view also explicitly rejects the opacity of phenomenal concepts, it can be used as another counterexample with respect to premise (2) of Goff's argument.

The examples of Levin's and Schroer's views give us thus a good reason to reject premise (2) of Goff's argument. These views, after all, explicitly reject the opacity of phenomenal concepts and thus allow the a posteriori physicalist to embrace the intuitive view that phenomenal concepts provide us with at least some knowledge of their referents.

It seems then that Goff's argument, as described above, is ultimately not quite successful. Still, I think, the argument points the anti-physicalist in the right direction. Namely, once it is allowed that phenomenal concepts reveal to us some knowledge of their referents, it is natural to ask *what kind of knowledge* phenomenal concepts provide us with. Here, given that it is highly plausible that phenomenal concepts do not refer to their referents via contingent modes of presentation, it is highly plausible that these concepts provide us with knowledge of the necessary properties of phenomenal states and perhaps, as some have argued, even with knowledge of the complete essences or natures of these states. To see what I mean here, consider the concept *water* for example. It is highly plausible that this concept provides us with some a priori knowledge of its referent, but, arguably, this knowledge will not concern the chemical essence of water but rather water's contingent properties, such as that it is liquid, that it exists in lakes and rivers, that it is

drinkable, etc. In the case of phenomenal concepts, it is plausible, however, that our phenomenal concepts, or at least the pure phenomenal concepts which we form solely by attention to a particular phenomenal state, do not reveal to us knowledge of contingent properties of phenomenal states. There does not, after all, seem to be any contingent appearance of pain, analogous to the contingent appearance of water. If that is the case, however, it seems that we must conclude that our phenomenal concepts provide us with knowledge of necessary or essential properties of their referents, phenomenal states.

The thought that our phenomenal concepts provide us with essential knowledge of their referents has recently resonated among the critics of physicalism as it seems to cast doubt on the physicalist doctrine. Here the thought is that if, thanks to our phenomenal concepts, we understand the nature of our phenomenal states as phenomenal, it is not clear that these phenomenal states could have, in reality, a physical nature, as the physicalists hold. A thorough philosophical explication of this line of reasoning has recently been offered by Martine Nida-Rümelin. Let me now try to sketch and evaluate her argument.

2. A Posteriori Physicalism and the Grasping Thesis

Nida-Rümelin's argument is based on the *phenomenal essentialism* thesis which states that solely in virtue of having a phenomenal concept, we grasp the corresponding phenomenal property.[176] Here to grasp a property via its concept amounts, according to Nida-Rümelin, to more than simply having a concept of that property. It amounts to knowing the nature of the property which the concept refers to. To grasp the nature of a given property often requires more than having a concept of that property, often, for example, additional empirical knowledge is needed. As, however, Nida-Rümelin argues, phenomenal properties (and presumably some other properties) are such that we are able to grasp them merely in virtue of having their phenomenal concepts. The phenomenal essentialism claim then tells us that solely in virtue of having a particular phenomenal concept we know what it is to have the corresponding phenomenal property or, in other words, what the nature of this phenomenal property consists in.[177]

What, however, does grasping a property amount to? According to Nida-Rümelin, it amounts to knowing what all things that have this property necessarily share.[178] If we wish to explore this question, it will not suffice, Nida-Rümelin emphasizes, to look for what all things which have this property in fact share. Perhaps, for example, in the actual world all organisms which have

[176]Nida-Rümelin (2007b, pp. 307–308).
[177]Nida-Rümelin (2007b, p. 308).
[178]Nida-Rümelin (2007b, p. 311).

consciousness also have eyes but having eyes is, nevertheless, clearly contingent with respect to the property of having consciousness. In order to distinguish the contingent features of a thing from its essential features, it is necessary to ask what attributes *need to be* instantiated by the entities which have that property. Nida-Rümelin suggests therefore that to grasp a property *P* means to know the counterfactual extension of some concept *C* of this property. Here the counterfactual extension of *C* is to be understood as what *P* refers to in possible worlds understood as counterfactual.[179]

As already mentioned, we often do *not* grasp a property, which we have a concept of, merely in virtue of having the concept. It is, for example, arguable that we do not grasp the property "being water" merely in virtue of having the concept of water. If, after all, Putnam and Kripke are right, then the counterfactual extension of the concept *water* is the chemical substance whose composition is expressed in the H_2O formula. In order to know this, however, we need, apart from the concept *water,* also the relevant piece of empirical knowledge about the actual world, namely the knowledge that water is H_2O in it. Things are different, according to Nida-Rümelin, when it comes to our phenomenal concepts: as the phenomenal essentialism thesis tells us, we are able to grasp the referents of phenomenal concepts merely in virtue of having these concepts, no additional empirical information is needed.[180]

Nida-Rümelin suggests that phenomenal essentialism can be justified in two steps. The first step amounts to the observation that our concepts in general provide us with a priori knowledge of their referents in the form of what Nida-Rümelin calls *essentiality conditionals*.[181] The second step amounts to the observation that those who have a phenomenal concept, will *ipso facto* know which essentiality conditional has a true antecedent.

Here the first step is based on the thought that the counterfactual, or secondary extension of a concept depends on the concept's referent in the actual world. If, for example, the concept *water* refers to H_2O in the actual world, then even the counterfactual extension of *water* is H_2O (and it is thus a Kripkean, or weak a posteriori necessity that water is H_2O). If, on the other hand, the actual world is – chemically speaking – such that *water* refers to XYZ in it, then the counterfactual extension of this concept is XYZ. These conditional claims are examples of what Nida-Rümelin calls essentiality conditionals – conditionals which express the dependence of counterfactual extension of a given concept on facts about the reference of the concept in the actual world. While the (implicit) knowledge of the essentiality conditionals is made a priori available to us by the given concept, this knowledge only implies facts about *possible* counterfactual extensions of the concept. We are, however, only able to grasp the property which the given concept refers to if we know

[179]Nida-Rümelin (2007b, p. 312).
[180]Nida-Rümelin (2007b, p. 323).
[181]Nida-Rümelin (2007b, p. 315).

which essentiality conditional has a true antecedent, which means that we know what the concept refers to in the actual world. This, of course, often requires empirical knowledge which goes beyond mere concept possession.

This leads us to the second step of the argument which amounts to Nida-Rümelin's appeal to the fact that in the case of phenomenal concepts, the very possession of the given concept already involves knowledge of which essentiality conditional has a true antecedent. It is, after all, the case that anyone who has the given phenomenal concept also knows its actual referent – the given phenomenal property. Phenomenal and similar concepts are thus often characterised as *actuality-independent* in the literature.[182] The idea here is that once we have the given concept, no additional information about the actual world is needed in order to know which essentiality conditional has a true antecedent. In this respect, of course, phenomenal concepts are importantly different from concepts such as *water*, *tiger* and similar everyday concepts. As we saw, after all, in order to find out, for example, what the actual referent of, for example, *water* is and which essentiality conditional associated with this concept has a true antecedent, we need additional empirical information – mere mastery of the concept is not sufficient here, which renders this and similar concepts *actuality-dependent*.[183]

We can now see more clearly the two steps of Nida-Rümelin's argument for phenomenal essentialism. The first step, as we saw, is the general semantic thesis that our concepts provide us a priori with knowledge of essentiality conditionals. The second step of the argument amounts to the claim that phenomenal concepts are actuality-independent, which means that we know, merely in virtue of having them, which essentiality conditional has a true antecedent. We can say that this claim in effect amounts to the claim that phenomenal concepts do not refer to any hidden structure but rather to the way our phenomenal states appear to us, i.e. how they are, phenomenally speaking.

The phenomenal essentialism thesis is merely one premise of Nida-Rümelin's anti-physicalist argument. Another key premise of her argument is the principle of cognitive transparency.[184] This principle is supposed to justify the claim that a given phenomenal property cannot be grasped via two different concepts, i.e. via a phenomenal and a physical concept. It is, of course, from this direction that one can expect most objections of the physicalists. They will presumably hold that our physical concepts reveal to us the nature of phenomenal properties (more on this later). If they are, at the same time, persuaded by the argument for phenomenal essentialism, they will likely hold that we grasp the nature of phenomenal properties via two distinct kinds of concepts, i.e. via phenomenal concepts and via physical concepts. That would mean that both the physical concept

[182]See e.g. Damnjanovic (2012, p. 80).
[183]See e.g. Damnjanovic (2012, p. 79).
[184]Nida-Rümelin (2007b, p. 327).

(together with the needed empirical knowledge) and the corresponding phenomenal concept provide us with knowledge of the nature of the given phenomenal property.

Here Nida-Rümelin says that it is indeed true in many cases that we grasp the same property via two distinct concepts, but we are in these cases able to rationally judge that the two concepts are necessarily coextensive, i.e. that they have identical counterfactual extensions. Nida-Rümelin calls this claim the principle of cognitive transparency.[185] This principle is based on what it means to grasp a property. Nida-Rümelin explains it as follows:

> *The idea of grasping a property implies fully understanding what having the property consists in. Therefore, every aspect of what it is to have the property should be in principle cognitively accessible to the subject. But then grasping the property in two conceptually different ways should necessarily go along with the capacity to realize that one and the same property has been cognitively penetrated.*[186]

This passage, as I see it, expresses the intuitive idea that one cannot truly understand or grasp a property in two distinct ways without being able to judge that the two conceptions are of the same property. Nida-Rümelin explicates this intuitive thought by means of the notion of necessary coextensiveness across possible worlds. If two concepts are necessarily coextensive, they will have the same extension in all possible worlds considered as counterfactual. In that case, however, the two concepts clearly refer to the same property. If then we know the counterfactual extensions of both concepts, we shall, according to Nida-Rümelin, be in a position to rationally judge whether the two concepts refer to the same property.[187]

Nida-Rümelin illustrates the plausibility of the cognitive transparency principle using the example of the concepts *water* and H_2O. Both of these concepts enable us, given the right cognitive background, to grasp the property of being water. Assuming that H_2O has no further hidden complexity, then, plausibly, once we have the H_2O concept, we shall not need any further knowledge in order to grasp the property of being water. One can also grasp the property of being water via the *water* concept although there one needs, apart from mastery of the concept, to know that *water* refers to H_2O in the actual world in order to know which of the essentiality conditionals associated with the concept has a true antecedent and thus to know what the counterfactual extension of the concept is.

We are then able to grasp the property of being water, at least given sufficient empirical knowledge, via two distinct concepts. At the same time, however, this "dual grasping" is not a counterexample

[185]Nida-Rümelin (2007b, ibid.).
[186]Nida-Rümelin (2007b, ibid.).
[187]Nida-Rümelin (2007b, p. 327).

with respect to the cognitive transparency principle. If, after all, we are in the epistemic situation of grasping the property of being water via these two concepts, we are *ipso facto* in the epistemic situation of being able to rationally judge, without further empirical information, that the two concepts are necessarily coextensive because we already know that water is H_2O.

The situation is quite different when it comes to the supposedly co-referring physical and phenomenal concepts. In the case of these pairs of concepts, argues Nida-Rümelin, we are never able to rationally judge that they are necessarily coextensive. Nida-Rümelin calls this claim the *cognitive independence* claim.[188] Here one can appeal to our intuitions regarding these concepts. Consider, on the one hand, the concept of pain, and, on the other hand, the concept of the firing of C-fibres, it seems that our understanding of the concepts involved together with with all relevant physical background knowledge gives us no reason to rationally judge that these concepts are necessarily coextensive. As Nida-Rümelin emphasises, if we, after all, were able to judge that, it would be hard to to explain the roots of our puzzlement about consciousness.[189]

The final thesis that needs to be established in order to complete Nida-Rümelin's anti-physicalist argument expresses something which may appear to be clear at first sight, namely that we grasp all physical properties via our physical concepts. Nida-Rümelin calls this claim the thesis of *cognitive accessibility of physical properties*.[190] This thesis is supported by the thought that we call physical properties physical precisely because they can be fully grasped via our physical concepts and fully expressed in physical terms. The thesis can, nevertheless, be questioned. According to some philosophers, after all, our fundamental physical concepts refer to particular causally defined micro-physical roles and these roles are realised by properties whose intrinsic nature is unknown to us. [191] These properties, which I shall say much more about in chapter 6, are sometimes called *quiddities*. The proponents of quiddities argue that the intrinsic nature of quiddities is in principle closed with respect to the methods of science. In spite of that, however, many thinkers view quiddities as physical properties.[192] In that case it is clear, however, that we do not grasp all physical properties by means of our physical concepts – physical entities thus have a hidden face which cannot be, even in principle, captured by means of physical concepts. I shall, however, ignore this difficulty here and shall instead presuppose, together with Nida-Rümelin and mainstream physicalists, that we grasp all physical properties via our physical concepts, or, more precisely, that we could in principle

[188]Nida-Rümelin (2007b, p. 327–328).
[189]Nida-Rümelin (2007b, p. 329).
[190]Nida-Rümelin (2007b, p. 326).
[191]See e.g. Lewis (2009).
[192]Using Chalmers's distinction which I shall discuss in detail in chapter 6 of this volume, quiddities are among broadly physical properties but they are not among narrowly physical properties (see Chalmers [2015]). Using Stoljar's distinction, quiddities are physical in the sense of the object-based conception of the physical but they are non-physical in the sense of the theory-based conception of the physical (see Stoljar [2001]). See also Strawson (2006).

grasp all physical properties via our physical concepts if we were in possession of completed physics.

We are now able to see the general structure of the anti-physicalist argument advanced by Nida-Rümelin. This argument is based on the claim that we are able to grasp our phenomenal properties via our phenomenal concepts (phenomenal essentialism). If we, at the same time, make the physicalistic assumption that our phenomenal properties are identical with particular physical properties, it means that we can also grasp them via physical concepts (cognitive accessibility of physical properties). If, however, we grasp the same property via two distinct concepts, the physical and the phenomenal, then we, according to Nida-Rümelin, need to be able to rationally judge, at least in principle, that the two concepts are coextensive (principle of cognitive transparency). We are, however, – even if we suppose that we have sufficient physical background knowledge – unable to rationally judge that the two concepts are necessarily coextensive (cognitive independence).[193] Therefore, concludes Nida-Rümelin, it is necessary to reject the supposition that phenomenal properties are identical with particular physical properties and, as a result, reject physicalism.

3. Grasping the Nature of a Property

Despite the fact that the premises of this complex argument look – at least at first sight – plausible, the argument has left many unpersuaded. One point of dispute has been the premise appealing to the cognitive transparency principle. Nic Damnjanovic and, informally, David Chalmers have both suggested that the a posteriori physicalists would have a good reason to reject this principle. [194] This principle tells us that if we know the counterfactual extensions of two distinct concepts and if the two concepts refer to the same property, we are able to rationally judge, without further empirical knowledge, i.e. a priori, that the two concepts are necessarily coextensive. This is, however, Damnjanovic argues, unacceptable for the a posteriori physicalists.

To see why, recall that according to the a posteriori physicalists, phenomenal concepts and physical concepts both refer to physical properties.[195] At the same time, however, a posteriori physicalists hold that phenomenal truths are not even in principle a priori entailed by micro-physical truths. Does it mean that a posteriori physicalism is incompatible with the principle of cognitive

[193]A more careful articulation of this point would require the note that we also have no reason to expect that any concepts of completed physics whose referents we would grasp would be such that we would be able to rationally judge that they are necessarily coextensive.
[194]Damnjanovic (2012, pp. 87–88), see also and informal commentary on Chalmers's blog: http://fragments.consc.net/djc/2006/12/nidarumelin_on_.html
[195]I am – like Nida-Rümelin – leaving out functionalism from this part of the discussion. It seems though that her argument could be in principle modified as to target functionalism.

transparency? Not on its own – this principle, after all, only applies to situations in which we *grasp* the same property via two distinct concepts while it does not apply to situations in which the two concepts merely refer to this property. However, the a posteriori physicalists could object that, according to their view, we indeed grasp, in the case of psycho-physical identities, the same property via two distinct concepts – at least if we accept the conception of grasping properties provided by Nida-Rümelin – and yet have a good reason not to expect that we should be able to rationally judge that the two concepts are necessarily coextensive.

We saw that, according to Nida-Rümelin, we grasp a given property iff we know the counterfactual extension of some concept of that property. As, however, Damnjanovic emphasizes, a posteriori physicalists hold precisely that we know the counterfactual extensions of both our phenomenal concepts and of our physical concepts.[196] To see why they hold this, recall that they embrace the intuitive view that our phenomenal concepts are actuality-independent, denying that these concepts refer via contingent modes of presentation and viewing them instead as directly referring rigid designators.[197] At the same time, however, they also hold that physical concepts refer directly and so are rigid designators. We, however, trivially know the counterfactual extension of rigid designators if we know their actual extension, since in every possible world considered as counterfactual they refer to what they refer to in the actual world. That, however, means that, according to a posteriori physicalists, these concepts meet Nida-Rümelin's criterion of enabling us to grasp the nature of their referents.

How, however, can the a posteriori physicalists hold that phenomenal concepts are actuality-independent? According to Loar, for example, our phenomenal concepts are type-demonstrative concepts which refer directly to particular types of sensory states. If so, however, they will refer to just that type of sensory state in every possible world considered as counterfactual. Even for the proponents of the hybrid account of phenomenal concepts, the reference of these concepts is in this way actuality-independent, given that these concepts have demonstrative, directly referring elements. According to Papineau, who rejects the demonstrative account, phenomenal concepts have the actuality-independence feature because their referents, phenomenal states are built into the concepts themselves. That, according to Papineau, clearly rules out any possibility that the given concept could, in counterfactual worlds, refer to anything else.[198]

Given that the a posteriori physicalists embrace the actuality-independence of phenomenal concepts, argues Damnjanovic, the principle of cognitive transparency formulated by Nida-Rümelin is in direct conflict with a posteriori physicalism. The proponents of this doctrine, after all, precisely

[196]Damnjanovic (2012, p. 88).
[197]Loar (2002, s. 298).
[198]Papineau (2007, p. 131).

deny that we are able to rationally judge, merely in virtue of grasping the referents of these concepts via these concepts, that phenomenal and physical concepts are actually coextensive, despite the fact that these concepts corefer. That, however, *a fortiori* means that they will also deny that we are able to judge that these concepts are necessarily coextensive. What is more, they provide us with detailed justification for this denial, appealing to the nature of our phenomenal concepts which renders these concepts logically independent of the theoretical notions of physics. We saw that Loar, for example, argues that phenomenal concepts must be logically independent given that they are recognitional concepts while Papineau appeals to the use-mention feature which supposedly renders these concepts non-deducible from physical truths. Nida-Rümelin thus cannot rely on the a posteriori physicalists accepting her principle of cognitive transparency. In the light of the objection expressed by Damnjanovic and Chalmers, her argument then seems to be ineffective against a posteriori physicalism.[199]

It seems to me that the present difficulty with Nida-Rümelin's argument roots from her view of the semantics of demonstrative concepts. We can see this if we recall that, according to Nida-Rümelin, our concepts a priori reveal to us associated essentiality conditionals which specify how the counterfactual extension of the given concept depends on the nature of its actual referent.[200] We saw, moreover, that Nida-Rümelin thinks that in the case of phenomenal concepts we know which essentiality conditional has a true antecedent merely in virtue of having the given concept and we thus need no additional empirical knowledge about the world in order to know the concept's counterfactual extension. This feature renders phenomenal concepts actuality-independent.[201] We also saw that, according to Damnjanovic, a posteriori physicalists endorse this view. This means that, in the case of Loar's view, phenomenal type-demonstratives must also be actuality-independent due to which we know their counterfactual extensions purely in virtue of having the relevant concept.

As I see it, however, Nida-Rümelin would disagree with the suggestion that demonstrative concepts are actuality-independent. Namely, I think she would argue that in order to know the counterfactual extension of a demonstrative concept, we need apart from mastery of the concept also additional empirical knowledge of its actual referent. The idea here is, roughly that I can think of a possible scenario in which my cognitive demonstrative act – conceiving of a state as of *that state* – would have referred to another state than the one it in fact referred to. If we now think of this possible scenario as actual, we can say that the counterfactual extension of my demonstrative concept *that state* would be different – the concept would refer to its actual referent in every possible world

[199]Nida-Rümelin's argument cannot be used against a priori physicalism either because the proponents of this doctrine would presumably reject her cognitive independence claim.
[200]Nida-Rümelin (2007b, p. 315).
[201]Nida-Rümelin (2007b, p. 319).

considered as counterfactual. If this consideration is correct, however, demonstrative concepts are then, after all, actuality-dependent.

This may seem like a minor point but it has significant implications for Nida-Rümelin's anti-physicalist argument. The idea here is that, according to Nida-Rümelin, in order to know the counterfactual extension of my demonstrative concept, I will need, just like in the case of the concept *water,* empirical knowledge of its actual referent which is not available to me merely in virtue of possessing that concept. If that is so, however, demonstrative concepts do not allow us, without additional empirical knowledge, to grasp their referents and so Loar's variety of a posteriori physicalism is not, after all, a counterexample to Nida-Rümelin's principle of cognitive transparency.

This defence of Nida-Rümelin's argument would be I think, effective if there were a strong argument for the view that demonstrative concepts are not actuality-independent. Nida-Rümelin, however, as far as I know, does not provide us with such an argument – the view merely follows from the general two-dimensional semantic framework which she recommends. While I think there is much to be said in favour of the two-dimensional framework, a sustained argument for it will need to be left for another occasion. Without such an argument, I suggest that we grant the a posteriori physicalists the claim that even on their view phenomenal concepts are actuality-independent.

4. Two Concepts of Essence

The last above-quoted passage above shows us rather vividly that Nida-Rümelin's principle of cognitive transparency has its roots in her notion of grasping properties. It would seem, after all, Nida-Rümelin tells us, that if we truly understand what the instantiation of a property consists in via two different concepts, we should also, upon sufficient reflection, be able to understand that we are dealing with two concepts of the same property. At the same time, however, we saw in the previous section that if a posteriori physicalism is true, then there are cases in which we know the counterfactual extensions of two distinct concepts of the same property without thereby being able to recognise, even in principle, that the two concepts necessarily corefer. This situation should, I think lead us to the suspicion that perhaps the explication of grasping a property via a concept in terms of knowledge of the counterfactual extension of the concept is inadequate. Certainly it seems odd to think that we truly understand the same property via two distinct concepts without being able to judge that the two concepts necessarily corefer. As I see it, this perceived discomfort should inspire us to look for a more appropriate way to articulate the notion of grasping or understanding a

property via a concept.

To see this point even more clearly, think once again of the demonstrative concept *that type of state*. Imagine that from time to time I encounter a particular strange sensation and I learn to recognise this sensation as *that type of inner state*. If, as a posteriori physicalists claim, the demonstrative concept is rigid, I will, in virtue of having acquired this concept also plausibly know its counterfactual extension – it will indeed in all possible worlds considered as counterfactual refer to the type of inner state it actually refers to. Despite knowing the counterfactual extension of the concept, however, there is a clear sense that the concept does not really cognitively penetrate its referent. Indeed, the concept is, arguably opaque, i.e. it does not provide me with any non-trivial knowledge of the nature of its referent. Why then think, however, that knowledge of the counterfactual extension implies true understanding, or grasping, of the nature of the property?

One way we could try to respond to this difficulty is to return to the intuitive conception of grasping a property or – which I think amounts to the same – to the intuitive conception of grasping the essence of a property.[202] Intuitively speaking, we grasp the essence of a given property if we know what it is for a thing to instantiate this property.[203] In view of this characterisation, we can see that the fact that we refer to a property by means of an actuality-independent concept, such as the demonstrative concept *this type of state* does not mean, at least not in this intuitive sense, that we know the essence of the property. We will, after all, in a real sense fail to know what the fact that a thing instantiates this property consists in. This conception of grasping the essence of nature of a property is thus stronger than the one which Nida-Rümelin appeals to as it seems that some concepts, whose counterfactual extension we know, do not provide us with knowledge of the essence of their referent in this stronger sense.

What does this stronger concept of an essence amount to? One interesting proposal along these lines has been made by Kit Fine, who rejects the modal understanding of essence and argues instead for a definitional view of essence, according to which an essential property of a thing is such a property which constitutes a real definition of the thing.[204] Here the thought is that while the properties which constitute a real definition of a thing are the properties which the thing has necessarily, it is not the case that every necessary property of a thing is a part of the thing's real definition. If we, to use Fine's example, consider two things whose natures are disconnected, the Eiffel Tower and Socrates, then “to be different from the Eiffel Tower” is a necessary property of Socrates, but it is not, intuitively, his essential property. A much better candidate for an essential property of Socrates is,

[202]Nida-Rümelin does not directly speak of grasping the essence but still, given that she speaks of “phenomenal essentialism”, I assume that this formulation would not be foreign to her thinking.
[203]See Goff (2011, p. 198).
[204]Fine (1995).

for example, “to be human”. Fine's conception is thus, as I see it, closer to the intuitive conception of essence which I described above than to the modal conception which prevails in current analytic philosophy.

In view of this stronger concept of essence we can ask whether we grasp or understand what is essential for a given phenomenal property or, which amounts to the same, whether our phenomenal concepts reveal to us the essential nature of our phenomenal states. One way to approach this question consists in appealing to the conceivability test introduced above. The conceivability test showed us that phenomenal concepts reveal to us some a priori knowledge of their referents. It remains to ask now whether phenomenal states reveal to us the essence of their referents.

Consider, for example, the phenomenal state of experiencing a particular shade of phenomenal red, such as when we, for example, look at a ripe strawberry (supposing that we are normal subjects). A natural suggestion here is that the way this shade of phenomenal red looks, the way it appears to us, is essential with respect to the given phenomenal state and, at the same time, this characteristic feature of the state (i.e. the way the shade looks) is revealed to us a priori, i.e. merely in having a phenomenal concept of the state. Let us first consider whether the way the red phenomenal quality looks is revealed to us a priori, merely in virtue of having the concept. That would mean that someone who has the concept of phenomenal red will know the way in which a phenomenal state needs to appear, the way it needs to look, in order for this phenomenal concept to apply to it. If that were so, our concept of phenomenal red would a priori reveal to us something along the lines “this state looks / appears so and so” where “so and so” stands for the particular phenomenal quality which the given phenomenally red state instantiates.

Let us now run the conceivability test and pose the question whether it may be that a phenomenally red state would not look or appear so and so (where “so and so”, once again, stands for the particular phenomenal quality). If that were possible, then our concept of a phenomenally red state would not a priori reveal to us the requirement on its referent (i.e. a satisfaction condition), according to which the referent needs to look so and so. As I see it, however, this is simply inconceivable. It seems to me therefore that a relevant phenomenal concept reveals to us a priori the way the given state must appear in order to count as phenomenally red, i.e. in order to be the referent of this concept. Things will be quite different if we consider the claim “this type of inner state looks so and so” in which the concept *this type of inner state* is a demonstrative concept as envisioned by Loar. The negation of this claim is, as I see it, conceivable as the concept “this inner state” is a blind pointer which does not reveal to us a priori anything non-trivial about its referent, or at least it does not reveal to us anything which would be incompatible with the claim that the referent “does not appear so and so” (where *so and so* represents the particular phenomenal quality).

Is the fact that a given phenomenal state looks or appears so and so an essential property of this phenomenal state? The alternative to an affirmative answer would be to say here that this property is contingent from the point of view of this state. That, however, would mean, I think, to misunderstand what the given phenomenal state is, i.e. a state whose appearance is its reality.[205] It seems therefore appropriate to say that our phenomenal concepts a priori reveal to us at least some, although perhaps not all, essential properties of their referents. We can call this claim – using Goff's notion of conceptual translucency introduced above – the *translucency claim.*

Could the physicalists make sense of the translucency claim? If the considerations introduced earlier in this chapter are correct, then the views of Loar or Papineau are incompatible with this claim. As we saw, after all, Papineau explicitly endorses the opacity of phenomenal concepts while Loar seems to be committed to this kind of opacity. Perhaps however the hybrid views of Levin and Schroer would have a better shot at making sense of this claim. It is arguable, after all, that the structural or relational properties which our phenomenal concepts reveal to us, according to these thinkers, are essential properties of our phenomenal states. Here the thought is that it may well be essential for phenomenal red that it is more similar to phenomenal orange than to phenomenal blue (as for Levin's view). Similarly, it may well be essential for phenomenal red that it is composed of many simpler phenomenal elements which come in particular amounts (as for Schroer's view). If we accept these considerations, then the proponents of hybrid views of phenomenal concepts are, after all, able to make sense of the translucency claim.

Does it mean that at least some form of the hybrid account of phenomenal concepts is plausible? As I see it, there is a good reason to reject the hybrid account. The reason is that the hybrid account implies that our phenomenal concepts reveal to us only structural features of their referents. This is particularly clear in the case of Levin, who holds that phenomenal concepts only a priori reveal to us the similarity relational of their referents. As we saw, Schroer thinks, on the contrary, that phenomenal concepts a priori reveal to us what he calls intrinsic properties of their referents. It is easy to see, however, that these supposed intrinsic properties really amount to structural properties. Our phenomenal concepts only reveal to us, after all, according to Schroer, how the supposed primitive phenomenal elements are combined and in what amounts they occur in the resulting phenomenal state, but we learn nothing about what these elements are like, qualitatively, on their own. It seems fair to say then that, according to the hybrid views, our phenomenal concepts reveal to us a priori only structural properties of their referents. We can call this thesis the *structural translucency claim*.

Let me now try to show why I think that this claim is implausible. The problem can be, once again

[205]See Searle (1997, p. 112).

demonstrated using the conceivability test. It is intuitively plausible that our phenomenal concepts reveal to us a priori that their referents feature what we can call, qualitative feels. Consider, for example the qualitative feel, or the quale of a slight migraine headache. It seems that the phenomenal concept which we use to conceive of this phenomenal state reveals to us a priori that the state features a highly specific kind of qualitative feel. We can express this as the thought "A migraine headache feels so and so". We can see why this thought is plausibly a priori if we try to consider its negation – the thought "A migraine headache does not feel so and so" looks inconceivable. Again, we can contrast this with the claim "This state does not feel so and so" where *this state* is understood as a demonstrative concept, which looks conceivable. It seems then that "A migraine headache feels so and so" is a priori, given that its negation is arguably inconceivable, at least if *migraine headache* is a phenomenal concept of the phenomenal state of having a migraine headache.

Can the proponents of the hybrid account make sense of the a priority of such claims? Hardly, I think, since, as I argued, phenomenal concepts as they are viewed by Levin and Schroer only a priori reveal to us structural properties of their referents (given that their demonstrative components are blind pointers). It is, however, highly plausible that, for example, the phenomenal feel of a migraine is not identical with or reducible to structural properties. On these accounts then our phenomenal concepts cannot reveal to us a priori the specific phenomenal feels of their referents. Then, however, claims like "A migraine headache feels so and so" cannot be a priori. Given, however, that our epistemic intuitions tell us that the claim is a priori (given that *migraine headache* is a phenomenal concept), we need to conclude that the hybrid accounts are incompatible with our epistemic intuitions concerning our phenomenal concepts. That, however, as we saw, is a serious challenge for a posteriori physicalism which has the ambition to meet our epistemic intuitions. Given this consideration, it seems reasonable to accept that our phenomenal concepts reveal to us a priori more knowledge than merely knowledge of structural features of their referents. We can call this version of the translucency claim the *non-structural translucency claim.*

Where has our discussion led us? Above I argued for the translucency claim, the view that our phenomenal concepts reveal to us some, although perhaps not all, essential features of their referents, i.e. phenomenal properties. We saw that the versions of a posteriori physicalism which work with the hybrid view of phenomenal concepts can make sense of a version of the translucency claim, according to which phenomenal concepts reveal to us a priori merely knowledge of broadly structural properties of their referents. We can call this version of the translucency claim the *structural translucency claim*. Subsequently, I argued for a stronger claim which I called the non-structural translucency claim, according to which our phenomenal concepts reveal to us a priori

more than merely structural features of their referents. Can the a posteriori physicalists make sense of this stronger claim? As I see it, it is far from clear that they can. Let me now explain why I think that the non-structural translucency claim is a serious challenge for a posteriori physicalism. To see this, consider first why the a posteriori physicalists have no problem with the structural translucency claim. Here I think the reason is that it is plausible that the phenomenal realm and the physical world are (partly) structurally isomorphic which means that they share some of their structural properties.[206] For the physicalists, of course, this perceived structural isomorphism is a result of the fact that the structure of the phenomenal realm is the very structure which is instantiated by the brain.

Why is the non-structural translucency claim a much harder problem for the a posteriori physicalists? The reason is that it is not clear how these non-structural phenomenal feels which our concepts a priori reveal to us, could exist in the physical world. Here, of course, the natural suggestion is that these phenomenal feels are simply identical with certain physical properties of the brain, or, alternatively, with some functional properties realised by the brain. As I see it, however, both of these options are ultimately implausible. Consider first the option that the non-structural or intrinsic properties knowledge of which our phenomenal concepts reveal to us are identical with some functional properties of the brain. The problem with this is obvious: while functional properties are essentially relational, the revealed knowledge informs us of phenomenal feel which arguably go beyond structure. It is far from clear therefore how the two kinds of properties could be identical.

How about the possibility that the phenomenal feels the knowledge of whose existence and nature our phenomenal concepts a priori reveal to us, are identical with some physical properties of the brain? Say, for example, that the phenomenal feel of seeing red is identical with a particular physical brain state. Can the physicalists make sense of this? I think it is far from clear that they can. To see why not, recall that if the considerations above are sound, then this phenomenal feel is an essential property of the particular phenomenal state. The phenomenal state is, however, according to the physicalists, a physical state and so all of its properties must be physical properties. That, however, means that a part of the essence of this phenomenal state is both physical and phenomenal. Namely, it is both a phenomenal feel and a purely physical state.

Can the physicalist make sense of this? The part of the essence of the phenomenal state which my phenomenal concept informs me about as of a phenomenal feel certainly seems to my understanding radically different than the corresponding physical property. Recall, moreover, that the physicalists cannot appeal here to one of these properties, which my concepts provide me with

[206]See e.g. Chalmers (1996, pp. 222-226).

knowledge of, being a mere contingent mode of presentation. Clearly, the physical concept does not reveal to me anything contingent, and if the present considerations are sound, the phenomenal feel the knowledge of whose existence and nature is revealed to me by the phenomenal concept, is also non-contingent but rather essential for the phenomenal state. Neither can the physicalists, given that the considerations in this chapter have been correct, appeal to the claim that the phenomenal concept reveals to me no substantial knowledge of its referent and, so to say, keeps me in the dark about its referent's nature. As a result we have two concepts which provide us with knowledge or understanding of a single property: a phenomenal concept which provides me with the knowledge that we are dealing with a phenomenal non-structural or intrinsic property and a physical concept which provides me with knowledge of the physical nature of this property. The a posteriori physicalist thus seems to be committed to the claim that we can understand a single property in two conceptually quite different ways. We can call this claim the *dual revelation thesis*.[207]

5. Dual Revelation

It is far from clear that the a posteriori physicalists can make sense of the dual revelation thesis, at least if this thesis is properly understood.[208] An interesting critique of this thesis has been offered by Philip Goff who draws our attention, once again, to the example of understanding what it is for something to be spherical in Euclidean geometry. He argues that we know the nature of sphericity if we know that a thing is spherical in Euclidean geometry if all its surface points are equidistant from its center. As he writes, it is hard to imagine that one could know this in a conceptually distinct way.

I think Goff emphasises here an important point but one needs to carefully consider why exactly the dual revelation thesis is implausible. We can see this if we consider a counterexample against this thesis which has been recently offered by John Henry Taylor. Taylor rejects the dual revelation thesis, arguing that a nature of a property can be expressed in two conceptually distinct ways. In his article he appeals to the ontological position called the *powerful qualities view*.[209] This view, held by C. B. Martin, John Heil, Galen Strawson and others, is one conception of the relation between dispositional properties (or powers) and qualitative (or categorical) properties of things. According to the proponents of the powerful qualities view, qualitative properties (e.g. size, shape, etc.) are at least in some cases identical with dispositional properties (e.g. fragility, inflammability). In such cases, they think there is only one property which is both dispositional and qualitative.

Taylor focuses on an example of sphericity. According to the powerful qualities view, we can

[207]This term has been introduced by Goff (2015b).
[208]Goff (2011) indeed calls this claim the *thesis of dubious intelligibility*.
[209]Taylor (2013).

understand this property in two conceptually distinct ways: we can understand it in qualitative terms as the property which a thing has if all its surface points are equidistant from its center, but we can also understand it in dispositional terms as the property which confers a particular set of dispositions on a thing.[210] The set includes, for example, the disposition to start rolling when placed on a slanted smooth hard surface or the disposition to leave an impression of a round shape when pushed into plasticine. According to the powerful qualities view then, the given dispositional and the given qualitative concept refer to one and the same property. If this view is correct, we can then know, Taylor suggests, what it is for the property of sphericity – and presumably also for other properties too – to be instantiated in two conceptually distinct ways which is, of course, exactly what Goff denies.

Consider, however, why the powerful qualities view is considered by its proponents to be reasonable in the first place. To see this, we need to ask why, even though in the case of sphericity we deal with two different conceptions, it is still viewed as reasonable by the proponents of the powerful qualities view to hold that we are dealing with two conceptions of a single property. Here I think the answer is that it makes sense to us that the particular property conceived of under the qualitative concept of sphericity is (identical with) the property which confers the given set of "spherical" dispositions on the object which instantiates it. We can see that Taylor's example is intelligible if we compare this suggestion with the suggestion that the property conceived of under the qualitative concept of sphericity is (identical to) the property which confers a set of dispositions associated with being cubical on the object which instantiates it. Such a suggestion seems, of course, utterly non-intelligible. Similarly, we lack intelligibility when it comes to the supposed coreference of phenomenal and physical concepts; we simply – and this point has been emphasized many times – do not understand how a phenomenal state could be a physical process.[211] Given, however, the considerations presented in this chapter, our phenomenal concepts provide us with the understanding of at least a part of the essence of our phenomenal states and our physical concepts provide us with understanding of the essence of physical states. If so, we have a good reason to conclude that our phenomenal states and physical states are not identical. Here one can, once again, appeal to the above-quoted passage from Nida-Rümelin, according to which, in effect, to truly understand a property in two ways means to understand that the same property has been cognitively penetrated.

As I see it, Taylor's counterexample shows that Goff's thesis that it is non-intelligible how we could understand a single property in two conceptually distinct ways is perhaps a bit too strong. Let us return at this point to the thesis of cognitive transparency, formulated by Nida-Rümelin and adapt

[210]Taylor (2013, p. 1290).
[211]See e.g. Nagel (1974).

this thesis for our purposes. The thesis tells us that *if we grasp what a property essentially is via two distinct concepts, we must be able to rationally judge that the concepts are coextensive.* We can see that Taylor's example is compatible with this thesis. Clearly, as we saw, the reason why some thinkers hold that the qualitative and the dispositional concept of sphericity provide us with understanding of the same property (which is both qualitative and dispositional) is that we are able to rationally judge, especially given sufficient background information, that the dispositional and the qualitative aspect of sphericity must "go together" and as a result, the two concepts are necessarily coextensive. How, after all, could the two aspects come apart? When it comes to, on the other hand to phenomenal and the corresponding physical concept, we clearly have, given the epistemic gap, no reason to think that the two concepts are necessarily coextensive.

The anti-physicalist should thus, as I see it, accept the non-structural translucency thesis, according to which, in effect, phenomenal concepts reveal to us the nature of phenomenal feels, in conjunction with the modified cognitive transparency claim, whose original version was formulated by Nida-Rümelin. Together with other theses defended in this chapter these two claims give us a reason to doubt the plausibility of a posteriori physicalism.

6. Strong Necessities and Translucency

It is interesting to consider the results which I have arrived at in this chapter, in the context of the previous chapter. We saw there that Loar does not quite succeed at explaining and justifying the existence of strong necessities which make it the case that some scenarios which are ideally conceivable, are, nevertheless, not metaphysically possible. As we saw, it is precisely this point which renders his a posteriori physicalism implausible. Some a posteriori physicalists have, however, reacted to this conclusion by producing examples of identity statements which they claim express strong metaphysical necessities and, at the same time, do not concern the mind-body case. One example of such a statement is "Cicero is Tully". Here a posteriori physicalists can argue that this statement is (ideally) conceivable but it is not metaphysically possible given that Cicero and Tully are the same person in the actual world and that proper names are, if Kripke is right, rigid designators. That, of course, on its own, does not suffice to show that "Cicero is Tully" describes a strong necessity. Chalmers, after all, can reply here that while the secondary intension of the statement is necessary, its primary intension is contingent given that we can conceive of a world in which the names "Cicero" and "Tully" will refer each to a different person.[212] If we considered such a world as actual, argues Chalmers, it would be true in it that Cicero is not Tully.[213]

[212]Chalmers (2010, p. 171).
[213]In other words, the possible world verifies "Cicero is not Tully".

It is at this point that some of Chalmers's critics object that Chalmers is supposing here that *Cicero* and *Tully* are not radically opaque concepts, i.e. they are not concepts which do not a priori reveal anything non-trivial about their referents.[214] According to Chalmers, these concepts a priori reveal to us at least some information about the contingent properties of their referents. It is precisely this supposed a priori available content of these concepts which enables us to evaluate their reference in possible worlds considered as actual. Some a posteriori physicalists, however, argue that these and similar concepts (e.g. *this* or *Prague*) are radically opaque. If that is the case though how could we know what the given concept refers to in a given possible world considered as actual? As we saw, in the case of some concepts, their reference in possible worlds considered as counterfactual is determined by a posteriori information about their reference in the actual world, but a posteriori information will not be helpful when it comes to their reference in possible worlds considered as actual. If, however, those concepts are opaque, we know a priori nothing non-trivial about their referents and, as a result, we shall not be able to evaluate their reference in possible worlds considered as actual. That means, Chalmers's critics argue, that the statement "Cicero is not Tully" which seems (primarily) conceivable does not describe any possibility, secondary (represented by a possible world considered as counterfactual which satisfies the statement), or primary (represented by a possible world considered as actual which verifies the statement). In that case, however, the original statement ("Cicero is Tully") expresses a strong necessity.

It is of course controversial whether *Cicero* and *Tully* are radically opaque concepts. Thinkers, such as Chalmers or Nida-Rümelin who, as we saw, accept two-dimensionalism, reject the existence of radically opaque concepts but many remain unpersuaded. If we accept that the concepts *Cicero* and *Tully* are radically opaque, physicalists can argue that the statements about phenomenal-physical identities are not the only ones which express strong necessities and it should therefore not be viewed as a weakness of a posteriori physicalism that they express strong necessities. Nobody, after all, doubts that "Cicero is Tully" expresses an identity although we are, at least according to a posteriori physicalists, dealing with strong necessity.

The material discussed in this chapter, nevertheless, provides the anti-physicalists with a reason to be sceptical about this objection. They can, after all, argue that the reason why "Cicero is Tully" expresses strong necessity is precisely the radical opacity of the concepts involved in the statement. If, however, the considerations in this chapter are plausible, our phenomenal concepts are translucent or perhaps even fully transparent. This conclusion provides the anti-physicalists with a reason to think that statements, such as "Cicero is Tully" cannot dispel doubt about the coreference of physical and phenomenal concepts. If the a posteriori physicalists are to dispel this doubt, it will

[214]Goff – Papineau (2014)

not suffice to appeal to statements outside the mind-body case which express strong necessities. In order to dispel doubt about the coreference of phenomenal and physical concepts, a posteriori physicalists would need to appeal to statements outside the mind-body case which express strong necessities and, at the same time, which are such that the concepts expressed by them are translucent or transparent, such as is the case with phenomenal and physical concepts.

7. Conclusion

In this chapter I tackled an objection against a posteriori physicalism, according to which our phenomenal concepts reveal to us certain knowledge of their referents which ultimately casts doubt on a posteriori physicalism. I have suggested that while the thesis of structural translucency is compatible with physicalism, the claim of non-structural translucency of phenomenal concepts can, together with other claims defended in this chapter, give us a good reason to be sceptical about the plausibility of a posteriori physicalism. I also tried to show that the results of this chapter can help the anti-physicalists reply to an influential objection which can be raised against the results of the previous chapter.

5. The Magic of Emergence

1. The Concept of Emergence

In the previous chapters I offered a number of considerations and arguments against the main forms of the physicalist reduction of consciousness. Supposing that those anti-materialist arguments and considerations are sound, it is natural to ask what options there are open for someone who views consciousness as a non-physical phenomenon but who, nevertheless, supposes that consciousness is an integral part of nature and that human and animal forms of consciousness are products of evolution. Is such a view of consciousness possible at all? The project of the remaining part of this book will be to search for just such a view which we can call *naturalism without materialism*.

One approach which has been popular among those looking for alternatives to the materialist reduction, is emergentism, i.e. the view that consciousness is an emergent property of some living organisms. Various versions of emergentism have been embraced or at least taken seriously by thinkers such as Tim Crane, Timothy O'Connor, Brian McLaughlin, Martine Nida-Rümelin, Hong Yu Wong, David Chalmers. The view has, however, at the same time been criticized from more than one direction. Apart from objections raised by proponents of physicalism, emergentism has also been criticized by some of those who lean towards other non-reductive views of consciousness, typically towards some form of panpsychism or neutral monism, views I shall focus on in the following chapters.

In this chapter I shall first introduce emergentism, moving on thereafter to describe and evaluate the objections which have been raised against it by Galen Strawson and Thomas Nagel. I will show that both arguments ultimately fail for the same reason – because they conflate epistemic or logical matters with physical matters. This problem has been expressed by James Van Cleve in his reply to Nagel's argument. I shall show that this objection also applies to Strawson's anti-emergentist argument. While for much of this chapter I shall be defending emergentism, I shall conclude the chapter in a critical key, arguing that emergentism faces challenges which make the view implausible and which should motivate us to look into alternative non-reductive views. Namely, I shall argue that emergentism faces serious threats concerning the alleged causal interactions between physical reality and emergent, non-physical phenomenal properties. Here I shall comment on the causal argument offered by David Papineau but will also offer my own critical comments with respect to emergentism.

Emergent properties are usually characterised, roughly, as properties of complex systems which

somehow arise in these systems but are, at the same time, irreducible to the properties of the systems' components even if the manner of organisation of these components is taken into consideration. That means, as it is sometimes stated in the literature, that emergent properties are something *over and above* the properties of the system's components, even including their mode of combination.[215]

Whether there are in fact any properties or entities which are emergent in this (roughly sketched) sense in nature, is a controversial matter; their existence is denied by the physicalists and, indeed, by many other thinkers. Emergentist ideas, nevertheless, have a long history, going back at least to John Stuart Mill and, according to some, to Galen.[216] While historically, various kinds of phenomena, such as certain chemical, biological or mental properties (e.g. the property of being alive) have been viewed as emergent, nowadays by far the most common form of emergentism is emergentism about consciousness. This is unsurprising since, as we saw, many philosophers nowadays view consciousness as not reducible to brain states and processes but at the same time as somehow arising from these. Indeed if consciousness is understood as a non-physical but yet natural phenomenon which arises from physical brain processes, some form of emergentism may seem like an attractive option when it comes to making sense of the presence of consciousness in nature.

Attractive it may be but even a cursory glance at the emergentist literature reveals that there are many different conceptions of emergence in play and that it is not easy to make sense of the very idea of emergence.[217] Instead of trying to map all these different conceptions, I shall focus here on a concept of emergence which is the most relevant from the point of view of this volume and which is also, arguably, the concept of emergence which Nagel and Strawson argue against. I suggest that this concept is one of ontological emergentism, i.e. emergentism, understood as an ontological thesis, a thesis about the general nature of reality.

Why, however, is ontological emergence the most relevant variety of emergence from the point of view of the project pursued in this volume? The reason is simply that once we give up the view that consciousness is physical and suppose, tentatively, that consciousness is an emergent phenomenon, we will need to accept that consciousness is ontologically emergent, i.e. that it is an ontologically new (i.e. non-physical) phenomenon, irreducible to the physical phenomena which underlie it.

It is arguably because the concept of ontological emergence is the most relevant concept of emergence in the current context, that it has become a target of critical attention of Nagel and Strawson who both prefer different non-reductive views. Nagel makes it quite clear that he has no

[215]See e.g. O'Connor – Wong (2012).
[216]Caston (1997).
[217]For a useful survey of the various concepts of emergence, see Van Gulick (2006).

objections against the existence of properties emergent in a weaker, epistemological sense. He writes, for example: “Emergence is an epistemological condition: it means that an observed feature of the system cannot be derived from the properties currently attributed to its constituents.” [218] Statements like these show us that Nagel allows for the existence of properties of complex systems which are not derivable from anything we currently know about the properties of the systems' parts and the way these are combined. I take it that Nagel, in this passage basically says that a property of a system is epistemologically emergent if our current knowledge concerning the properties of the system's parts and their organisation does not a priori entail, even given ideal reflection, the truths about the emergent feature. I thus understand Nagel's notion of *derivability* used in the quoted passage as a logical relation which is equivalent to the notion (a priori) deducibility. On this understanding, property *P* is, in Nagel's sense, derivable from property *R* iff once we know that R is instantiated and we have a concept of property P, we can in principle deduce that *P* is instantiated without need for any more empirical knowledge.

Crucially, Nagel thinks that the epistemological fact that certain properties of a system are currently non-derivable by us from the properties of the system's parts always results from the fact that the system's parts have properties which are currently unknown to us. If these properties of the parts were known to us, then, Nagel suggests, all the properties of the system, including the properties which we now view as epistemologically emergent, would be – at least in principle – derivable from the properties of the system's fundamental parts.

While accepting the existence of epistemologically emergent properties, Nagel argues against what he calls “truly emergent properties”. He writes:

> *The supposition that a diamond or an organism should have truly (not just epistemologically) emergent properties is that those properties appear at certain complex levels of organization but are not explainable in terms of any more fundamental properties, known or unknown, of the constituents of the system.* [219]

Here Nagel describes truly emergent properties as properties of a complex system which are not explainable, even in principle, in terms of any properties of the system's parts (including their organisation). True emergence, unlike epistemological emergence, is thus, according to Nagel, not a result of our limited knowledge of the properties of the system's parts. Truly emergent properties, after all, could not be explained in terms of the properties of the parts, even if we knew everything there is to know about the properties of the parts (including the laws that govern their physical behaviour) and if we were ideal reasoners. This suggests that there is, in the alleged cases of true

[218] Nagel (1979, p. 182).
[219] Nagel (1979, p. 186).

emergence, something about the way nature is owing to which some properties of complex systems are simply unexplainable in terms of the properties of the systems' parts. True emergence, as understood by Nagel, is then an ontological rather than epistemological condition, although Nagel defines it in terms of unexplainability which is, of course, a concept of epistemological origin.

It is natural to ask whether Nagel's way of characterising emergence would be acceptable for the proponents of emergentism or whether Nagel is, so to say, attacking a straw-man. His own characterisation certainly seems consistent with the famous remark of Samuel Alexander, a classic exponent of emergentism, who wrote that "[t]he existence of emergent qualities thus described is something to be noted, as some would say, under the compulsion of brute empirical fact […]" and "admits no explanation […]."[220] The categorical character of Alexander's claim suggests that he also understands emergence ontologically, not as a result of a mere contingent epistemological limitation.[221] Moreover, he clearly, just like Nagel, puts emphasis on the unexplainability of emergent properties whose existence, he thinks, cannot be explained in terms of any of the processes which belong to the "lower level" of nature, the level from which emergent properties are supposed to arise. It seems therefore that Alexander could accept Nagel's characterisation of true emergence.

C. D. Broad, another classic exponent of emergentism, claimed that some natural wholes have properties which are emergent in the sense of not being deducible from the properties of the parts forming these wholes even if their manner of combination is taken into account.[222] Broad, just like Alexander, doesn't have in mind here the contingent non-deducibility resulting from our current ignorance concerning the relevant properties of the wholes' parts or from our imperfect reasoning skills, but rather non-deducibility in principle or, as he puts it, non-deducibility in theory.[223] Once again, Broad plausibly uses the concept of non-deducibility in principle to define a concept which is ultimately ontological which is clear from the fact that he understands emergentism as an intermediate position between pure mechanism and what he calls Substantial Vitalism which surely are ontological views. Nagel's characterisation of emergent properties in terms of their non-explainability and non-derivability in principle can thus be viewed as in effect equivalent to Broad's influential account.

Nagel's characterisation should also be acceptable for David Chalmers, who takes emergentism seriously, although he does not quite subscribe to it.[224] He characterises what he calls a *strongly emergent phenomenon* as a high-level phenomenon which "arises (in some sense) from the low-

[220]Alexander (1920, pp. 46-47).
[221]This is the way O'Connor and Wong read Alexander as well. See O'Connor and Wong (2012).
[222]Broad (1925, p. 61).
[223]Broad (1925, ibid.).
[224]See e.g. Chalmers (2010, p. xviii–xix).

level domain, but truths concerning that phenomenon are not deducible even in principle from truths in the low-level domain."[225] Chalmers contrasts strong emergence with *weak emergence* where a weakly emergent phenomenon "arises from the low-level domain, but truths concerning that phenomenon are unexpected given the principles governing the low-level domain". [226] While the notion of weak emergence has been much more common in the scientific discussion of emergence, think of connectionist modelling or non-linear dynamics,[227] only strong emergence can, Chalmers argues, possibly be of help when it comes to accounting for the existence of consciousness in the physical world. Given that Chalmers's notion of non-deducibility in principle seems to be equivalent to the notions of unexplainability and non-derivability used by Nagel, it is arguable that Chalmers would also accept Nagel's characterisation of true emergence.

Nagel's understanding of emergence should then be acceptable for many emergentists as capturing the ontological notion of emergence. At the very least, it is plausible that unexplainability and non-derivability of emergent properties discussed by Nagel are necessary conditions, although perhaps not sufficient conditions, for any case of ontological emergence. If this is correct, Nagel's criticism applies to any case of ontological emergence.

Is the notion of emergence critiqued by Nagel also the notion criticised by Galen Strawson? Let me now show why I believe that we should answer this question affirmatively. Strawson characterizes emergentism about conscious experience as the view that:

> *[p]hysical stuff in itself, in its basic nature, is indeed a wholly non-conscious, non-experiential phenomenon. Nevertheless when parts of it combine in certain ways, experiential phenomena 'emerge'. Ultimates in themselves are wholly non-conscious, non-experiential phenomena. Nevertheless, when they combine in certain ways, experiential phenomena 'emerge'.*[228]

The existence of consciousness is then, according to emergentism as understood by Strawson, a result of a complex combination of physical ultimates (i.e. ultimate, or fundamental constituents of the physical universe which will be posited by completed physics) which are, considered in isolation, non-conscious. Could consciousness – or to use the word preferred by Strawson – *experience* be emergent in this sense? Strawson finds this notion of emergence hard to make sense of and claims that it is ultimately incoherent.

Just like Nagel, Strawson doesn't renounce emergence *en bloc*. He in fact describes a notion of

[225]Chalmers (2006, p. 244).
[226]Chalmers (2006, ibid.).
[227]See e.g. Bedau (1997).
[228]Strawson (2006, p. 12).

emergence which he has no objections against. For Strawson, an example of a phenomenon emergent in the permissible sense of the word is the familiar phenomenon of liquidity.[229] The example of liquidity is indeed sometimes put forward by the advocates of emergentism as a useful analogy which can help us understand the emergence of consciousness.[230] Liquidity is viewed here as a "novel" property possessed by some complex systems of H_2O molecules while not possessed by any of the individual H_2O molecules or by their (ultimate) constituents. In order for the property of liquidity to emerge, the molecules need to be combined or organised in an appropriate complex manner. Having thus identified an example of an emergent phenomenon, these emergentists invite us to view consciousness as another example of an emergent phenomenon, which, in this case emerges in complex biological systems.

Strawson, however, suggests that the liquidity analogy is unhelpful when it comes to making sense of how consciousness arises. He argues that while we have at least a rough understanding as to why and how certain complexes of H_2O molecules instantiate the property of liquidity (an understanding which, we believe, could be further fleshed out and perfected if we learned more about the physical and chemical properties of H_2O molecules), we utterly lack such understanding as to why and how certain complexes of neurons and synapses give rise to conscious experience. It seems then that it is to a considerable degree intelligible or explainable as to why liquid phenomena arise from non-liquid H_2O molecules while it is utterly unintelligible and inexplicable for us as to why consciousness could arise from non-conscious processes.

What sort of unintelligibility or unexplainability pertains to the supposed emergence of consciousness, according to Strawson? Is it mere contingent unintelligibility or unexplainability resulting from our unfortunate epistemic situation or is it some stronger notion of non-intelligibility or inexplicability?[231] Here Strawson is quite unequivocal. If consciousness did emerge from neural phenomena then it would not be even in principle – or, as Strawson puts it, even for God – intelligible or explainable as to why it emerges. This is clear from the fact that Strawson calls the notion of emergence which, he thinks, would have to pertain to the emergence of conscious experience from non-experiential processes, *brute emergence* while, as we saw, he views the emergence of liquidity from non-liquid phenomena as in principle intelligible.

We can see now that the notion of brute emergence criticized by Strawson is equivalent with the notion of true emergence, itself defined in terms of unexplainability in principle, which is criticised by Nagel. Both thinkers then target a strong, ontological notion of emergence, a notion which

[229]Strawson (2006, p. 13).
[230]See e.g. Searle (1998, p. 1940).
[231]Notice that if the notion in question were the former notion of non-intelligibility, then Strawson's notion of emergence would be equivalent to Nagel's notion of epistemological emergence discussed above.

involves unintelligibility or unexplainability in principle, while allowing for some weaker notions of emergence. At the same time, however, they insist that any weaker notions of emergence would be useless when it comes to accounting for the existence of conscious experience.

In what follows, I shall try to show why both Nagel and Strawson think that ontological emergentism is wrong and needs to be rejected, starting by introducing Nagel's argument which came chronologically first. Before I do it, let me remark that I have my doubts concerning defining ontological emergence by means of concepts which are epistemological in their origin, such as unexplainability, non-deducibility or unpredictability. While, after all, these definitions manage to distinguish ontological emergentism from a priori physicalism, whose proponents hold that there are no in-principle-unexplainable or non-deducible higher-level properties, it is far from clear that definitions of this sort manage to clearly distinguish emergentism from a posteriori physicalism. The advocates of this view, after all, precisely hold that there are such unexplainable or non-deducible higher-level properties. However, they, of course, deny that this unexplainability or non-deducibility results from the fact that consciousness is non-physical, holding instead that it is a result of having the kind of conceptual repertoire we have.

A natural way out of this difficulty is, as I see it, to define the concept of ontological emergence in ontological terms. Perhaps the simplest way to do this is to say that an emergent phenomenon is *ontologically new* with respect to the entities and properties which it emerges from, or to say that an emergent phenomenon is *something over and above* the entities and properties which give rise to it. Still, arguably, these expressions are, although perhaps helpful, ultimately metaphorical. To get a clearer idea, one can, for example, follow Timothy O'Connor and Hong Yu Wong in saying that emergent properties are basic properties had by composite individuals.[232] Here basic properties are such properties whose instantiation does not even in part consist in instantiation of other properties by the entity or its parts. Another interesting attempt to define ontological emergence in ontological terms has been recently suggested by Elizabeth Barnes who argues that emergent phenomena are to be understood as fundamental, yet not ontologically independent properties.[233] In preferring the ontological definitions of ontological emergence over the epistemological ones I do not, of course, deny that the above mentioned definitions which appeal to epistemic notions correctly describe emergent phenomena. They, nevertheless, fail to distinguish emergentism from a posteriori physicalism.

[232]O'Connor – Wong (2005, p. 664).
[233]Barnes (2012).

2. Nagel Against Emergence

That Nagel is opposed to emergentism is clear from multiple passages in his works. In his latest book, *Mind and Cosmos,* he remarks, for example:

> *That* [...] *purely physical elements, when combined in a certain way, should necessarily produce a state of the whole that is not constituted out of the properties and relations of the physical parts still seems like magic* […]*".*[234]

What, however, is Nagel's argument against emergentism? Nagel suggests that if mental properties, including consciousness, were truly, i.e. ontologically, emergent properties of the brain, they could not even be *caused* by the neural processes happening in the brain.[235] This surprising conclusion, apart from being hard to swallow for anybody conceiving of consciousness as of a natural phenomenon, is, as I shall now try to show, in conflict with the very doctrine of emergentism. As we saw, after all, emergentists normally hold that the brain happenings give rise to or determine the emergent property of consciousness. Using the metaphor of a vertical hierarchy, popular in the emergentism debate, we can say that emergentists are committed to some form of "upward determination", i.e. determination of the emergent properties by lower-level properties. That such a commitment is an integral part of emergentism is widely acknowledged in the literature: Chalmers, as we saw, talks about the emergent phenomena arising from the lower-level phenomena,[236] James Van Cleve talks about the emergent properties of the whole being dependent on or determined by the properties of the parts constituting the whole[237] and Jaegwon Kim talks about emergent properties being supervenient on the properties from which they emerge. By supervenience, which he understands as a necessary, although not sufficient condition of emergence, he means, roughly, that the facts about the emergent properties are completely determined by the facts about the properties from which these emerge and so there is no space for variation at the emergent level without variation at the base level.[238] These and similar appeals to upward necessitation or determination certainly make sense since, intuitively, without upward determination of some sort, the emergent properties would be merely free-floating and, plausibly, the resulting position would not be emergentist but rather a version of psycho-physical parallelism.

Given that some form of upward determination is an integral part of emergentism, does it mean that

[234]Nagel (2012, pp. 55-56).
[235]Nagel's true target in this article is emergentism about mental properties in general. However, since mental states crucially include conscious mental states (see Nagel [1974]), one can easily mount the argument against a more limited target, namely conscious experience, being agnostic about the existence and metaphysical status of non-conscious mental states. This is how I shall proceed here for the sake of convenience of expression.
[236]Chalmers (2006, p. 244).
[237]Van Cleve (1990, p. 221).
[238]See Kim (2006, p. 193) for Kim's definition of supervenience: "[t]o say that M supervenes on N_1,..., N_n is to say that any system that has the base properties N_1,..., N_n will necessarily have the supervenient property M".

its proponents are committed to the view that brain states cause conscious states – as Nagel suggests? It seems to me that the answer is plausibly affirmative if we accept that causality is the only kind of determination relation which can be found in nature. One could, of course, deny this and hold that apart from causality, there are also other determinative relations to be found in nature. Perhaps the strongest candidate for such a supposedly non-causal determination relation is the above-mentioned relation of supervenience. That supervenience is non-causal is held by O'Connor and Wong who distinguish between *supervenience emergentism*, according to which emergent properties are not caused by the lower-level phenomena but rather supervene on them, and *causal emergentism* according to which emergent properties are caused by the lower-level phenomena.[239] It is plausible, however, that the supervenience relation really amounts to a kind of synchronic causal relation when it is combined with the view that consciousness is non-physical and yet arises in complex physical systems. If this is so, Nagel's worry indeed applies to emergentism. Moreover, even if the supervenience relation is not understood as causal, and is perhaps understood as self-standing, it clearly should be viewed (indeed, it is in many cases defined) as a sort of upward natural necessitation or determination and Nagel's argument could then apply to emergentism in a slightly rephrased form.

How, however, does Nagel arrive at the above mentioned conclusion that ontologically emergent properties could not be caused by the brain's microstructure, a conclusion which, if the preceding considerations are correct, would be troubling for the emergentists? As we saw, truly emergent properties are for Nagel those properties of a complex system which are not even in principle explainable in terms of the properties of the system's parts and their combination. Most properties which we normally attribute to macro-objects are, according to Nagel, not like that – they are, at least in principle, explainable in terms of the known or unknown properties of the object's constituents and the way they are organised. As such, they are not truly, or ontologically, emergent. Nagel offers an example of a diamond on whose properties he comments as follows:

> *Some of them, like shape, size, weight, and crystal structure, are directly entailed by the physical properties and relations of its constituents and their effects on each other when they are so combined. Others, like color, glitter, and hardness, involve interaction between the diamond and other things, and must be explained in terms of the effects of the diamond's constituents on those other things.*[240]

Nagel here suggests that some of the standard, non-emergent, properties of a diamond are directly entailed by its microphysical properties which, as we saw, makes them reductively explainable in

[239]O'Connor and Wong (2012)
[240]Nagel (1979, p. 186).

terms of these microphysical properties. Other properties, which arise as a result of interactions between the diamond and other objects, are reductively explainable if we take into account the diamond's microstructure and its interaction with the microstructures of the other objects.

We saw that emergentism is the view that some objects, apart from these properties which we are able to explain given what we know about the objects' microstructure (and perhaps microstructures of other objects which interact with it), also possess emergent properties which are not explainable, even in principle, although they are regularly and uniformly correlated with the properties of the objects' constituents and their combination. Here, the most promising candidates are, of course, phenomenal properties. These uniform correlations between physical and emergent mental properties can even, Nagel notices, be captured in the form of laws such as "If an organism is in physical state *P*, it is also in mental state *M*". This, however, raises the crucial question as to why these lawful correlations obtain. A natural thing to say here is, of course, that they obtain owing to the fact that the supposed emergent properties are *caused* by the appropriately organised micro-physical entities of the brain's parts.[241] Nagel, however, insists that the emergentist is not entitled to an answer of this kind. The problem, argues Nagel, is that emergentists cannot hold that our brain states cause our mental states given that they hold that the brain only has physical, non-conscious micro-properties. No configuration of purely physical properties or entities can, after all, Nagel argues, necessitate the instantiation of phenomenal properties. Nagel thus holds that given that the physical microstructure of the brain does not necessitate the existence of consciousness, it cannot cause the existence of consciousness either.

The step from causation to necessitation in Nagel's argument rests on an assumption about the nature of causation which Nagel makes and which I take to be intuitively plausible, namely that causation requires necessitation. In Nagel's view, which, I would guess, most of us embrace when not philosophizing, for any pair of events *A* and *B*, it is true that if *A* causes *B* then *A* necessitates *B* or, in other words, makes the occurrence of *B* necessary.[242] Nagel thus rejects the view of causation as mere regularity or correlation which is often – Nagel thinks wrongly – attributed to David Hume. Nagel expresses his own, non-reductionist view of causality when he writes: "True causes *do* necessitate their effects: they make them happen or make them the case. Uniform correlations are at best evidence of such underlying necessities."[243]

If the non-reductive view of causation is true and brain micro-properties in fact cause phenomenal properties, then these must be somehow necessitated by brain events. There is, however, Nagel

[241]Whether one would be inclined to say that depends, among other things, on whether one thinks the relation of causation can be synchronic. Nagel, apparently, thinks so, since he views the supposed emergent properties as existing at the same time as underlying properties which are supposed to give rise to them (see Nagel [1979, p. 186]).
[242]Nagel (1979, p. 185–186).
[243]Nagel (1979, p. 186).

suggests, no necessitation whatsoever between the physical properties of the brain and the relevant organism's mental states. He writes: "[t]here is no sense in which my body's physical state *in itself* makes it the case that I am in mental state *M*."[244] In the same paragraph he adds: "There must be some kind of necessity here. What we cannot understand is how the heat, or the brain process, necessitates the sensation. So long as we remain at the level of a purely physical conception of what goes on in the brain, this will continue to appear impossible."[245]

In this interesting passage, Nagel, I think, refers to the fact that statements such as "physical state *P* makes it the case that I am experiencing conscious sensation *C*" leave us puzzled as we have no clue as how and why that could, or should be true. We certainly do not feel that the occurrence of *P* necessitates the occurrence of *C*. Nagel notices in this passage that we feel the same lack of necessitation whether we focus on the less immediate relation between heat, presumably understood as molecular motion in our environment, and our heat sensation, or whether we focus on the brain process which appears much further down in the relevant causal chain and, supposedly, directly causes the heat sensation. The symptom of this felt lack of necessitation is arguably the fact that most of us find philosophical zombies conceivable, even if perhaps metaphysically impossible. We would, plausibly, not be able to even conceive of zombies if we understood why and how e.g. a particular physical state necessitates, say, a perception of vermilion experienced by the organism which is in this physical state.

The apparent lack of necessitation, Nagel suggests, would allow for the existence of causality between physical properties of the brain and mental states, had the regularity view of causation been true, since there clearly exist relevant correlations – or constant conjunction – such as "whenever an organism is in physical state *P*, it is also in mental state *Q*". According to the regularity view, of course, such correlations are all that causation amounts to. If, however, causation requires necessitation, then, given that we accept that our physical brain states do not necessitate our mental states, we get the conclusion that our physical brain states do not cause our mental states. This conclusion leads, according to Nagel, to the following dilemma for the emergentist: either the brain constituents only have physical properties and our mental properties have no causal explanation at all, or there are other, non-physical properties of the brain constituents and these properties necessitate – and causally explain – our mental states.

Clearly, neither horn of the dilemma is acceptable for the emergentists. Above I explained that the first horn should be seen as inconsistent with emergentism due to emergentism's commitment to upward determination. The second horn should be, as I see it, equally unappealing to the

[244]Nagel (1979, p. 187).
[245]Nagel (ibid.)

emergentist since it amounts to attributing extra non-physical properties to the basic constituents of the brain (and the universe) in order to account for consciousness. This, however, is in sharp conflict with mainstream emergentism which postulates that complex systems have novel or irreducible properties while keeping an ontologically austere, i.e. materialistic, view of the systems' fundamental constituents.

To summarise, Nagel argues that consciousness could not be caused by the microstructure of the brain as long as this microstructure is viewed as purely physical. In doing this, he appeals to an intuitive claim that the relation of causation between two properties requires that one of them necessitates the other. However, given that, according to Nagel, there is no necessitation between microphysical properties of the brain and phenomenal properties, the former could not cause the latter. The emergentists therefore need to either accept that the emergent properties are not caused by the microstructure of the brain, or to allow that microphysical constituents of the brain have non-physical properties which necessitate the mental properties. However, as we saw, both of these options are in conflict with ontological emergentism. Clearly, if Nagel's argument works, ontological emergentists have a reason to worry. But does it work? Let me now investigate whether there is any way for the emergentists to block Nagel's argument.

3. Two Replies

Firstly, the emergentists could simply reject Nagel's assumption that the regularity view of causation is wrong. Without this assumption this argument cannot get off the ground since the crucial step from the thesis that the brain causes our mental states to the thesis that the brain necessitates our mental states is blocked. Since it is controversial whether causation in fact requires necessitation, a natural reaction to this objection is to say that Nagel's argument is sound under the condition that the regularity view of causation is false.

We can understand this objection better if we sketch Nagel's argument as follows:

1. Causation requires necessitation.

2. Physical properties of the brain constituents (including their organisation) fail to necessitate mental properties.

———

3. Either brain constituents do not cause mental properties, or they have some non-physical properties.

Here the objection motivates us to make the argument weaker by allowing that premise (1) is true only if the regularity view of causation is false. In view of this objection, we can modify the argument as follows:

1_m. If the regularity view of causation is false, causation requires necessitation.

2_m. Physical properties of the brain constituents (including their organisation) fail to necessitate mental properties.

———

3_m. If the regularity view of causation is false, then either brain constituents do not cause mental properties, or they have some non-physical properties.

The argument then, if indeed there are no other problems with it, shows that emergentism is incompatible with the irreducible view of causation which is, I think, still a remarkable result. That the argument should be taken seriously, even though it is merely conditionally sound, is, as I see it, further supported by the fact that it is the irreducible view of causation, according to which causation requires necessitation that most people pre-theoretically embrace and which should, as I see it, be viewed as the default position in the absence of a good reason for its rejection.

Secondly, one could, as I already mentioned, react to Nagel's argument by questioning the thesis that emergent properties of a system are *caused* by the properties and organisation of the system's parts. This objection then amounts to accepting the first horn of the dilemma expressed in (3). One could support the claim that emergentism does not require upward causation by arguing that the relation of causation is necessarily diachronic, i.e. that causes always chronologically precede their effects, while emergent properties are synchronically supervenient on the base properties and therefore emergence is not a causal relation. It is, however, far from clear that the relation of causation is necessarily diachronic, as the example of a weight hanging on a spring and causing it to hold stretched seems to show. One can surely doubt that what causes the spring to be stretched at any given moment is an event which happened at some previous moment. However it may be, Nagel himself, clearly, endorses synchronic causation and so it seems that simply rejecting the possibility of synchronic causation without a good reason is too dogmatic.

Another way to support the second objection and the claim that emergent properties need not be caused by their underlying physical properties is to deny that the upward determination relation integral to emergentism is causal. One way to pursue this strategy is to subscribe to supervenience emergentism and hold that supervenience is not a kind of synchronic causation. It seems, however, that even if the supervenience relation is understood as non-causal upward determination, Nagel's

argument could still apply to it in a modified form. Nagel could still plausibly hold that any kind of upward determination, and indeed any kind of natural determination in general, requires necessitation and given that, according to Nagel, necessitation is lacking in the brain-consciousness case, consciousness cannot be naturally upwardly determined which is, as we saw, in conflict with emergentism.

4. Van Cleve's Objection

Another reply to Nagel's critique of emergentism has been offered by James Van Cleve who argues that Nagel sets the bar unreasonably high when it comes to causal explainability of a property. [246] We have seen that an important step in Nagel's argument is his intuitive claim that in order for B-properties to be caused by A-properties and thus to be causally explainable with reference to A-properties, the former need to be necessitated by the latter. In other words, the existence of A-properties must in that case somehow necessarily lead to the existence of B-properties. Van Cleve, however, notices that the concept of necessity is ambiguous and poses the question as to what sort of necessity is on Nagel's mind here. He suggests that the necessity in question must be logical necessity.[247] However, he adds, the emergentists are free to appeal here to a weaker sort of necessity, namely what he calls nomological necessity.

Perhaps the best way to see why Nagel likely means logical necessity here is to consider his claim that physical properties of an organism do not necessitate its mental properties. This claim will surely be accepted by the emergentist if we understand it as being about logical necessity, i.e. as stating that truths about someone's bodily properties do not logically, or a priori, entail (all) the facts about his or her phenomenal properties.[248] This lack of logical necessity is, after all, articulated, for example, in the Mary and zombies thought experiments and, as we saw, it can easily be accepted by the emergentists.

According to Van Cleve, Nagel then, in effect, argues that in order for B-properties to be caused by A-properties, the former need to be *logically necessitated* by the latter. How plausible is this view though? Van Cleve suggests that embracing it commits Nagel to a version of *causal rationalism*, a view famously embraced by Spinoza.[249] According to causal rationalism, there is logical implication between causes and their effects, causes then imply, or logically necessitate, their effects. Once we

[246]Van Cleve (1990, p. 217).
[247]Van Cleve (1990, p. 217).
[248]Of course, logical necessity is, strictly speaking, a relation between statements, but both Nagel and Van Cleve understand it in a looser sense as a relation between properties. Here we can say that A-properties logically necessitate B-properties iff truths about A-properties logically necessitate, i.e. a priori entail truths about B-properties.
[249]See Bennett (1984, pp. 29-30).

assume that causal rationalism is true, Nagel's argument looks persuasive. Given that causes logically necessitate their effects, any case of property *A* causing property *B* is a case of *A* logically necessitating *B*. Given that physical properties of the brain do not logically necessitate consciousness, as the zombie and Mary thought experiments suggest, we should, according to Nagel, conclude that physical brain states do not cause phenomenal properties and, therefore, phenomenal properties are not causally explainable in terms of physical brain properties.

Some passages in Nagel's article provide reasons to think that he may actually lean towards some version of causal rationalism. He writes, for example, „[a]n electron is a particle with a certain charge and a certain mass. Those properties *imply* that it will interact in a definite way with fields and with other objects. [my italics]“[250] Given that implication is clearly a logical relation, it seems that Nagel might embrace a form of causal rationalism.

Is, however, Nagel committed to full-fledged causal rationalism? Strictly speaking, all Nagel needs for the purposes of his argument is rationalism about synchronic causation between parts of a system and the emergent properties of the whole system instantiated at the same time. Van Cleve calls this more limited, synchronic version of causal rationalism *mereological rationalism.*[251] According to mereological rationalism, all the macro-properties of complex systems which are synchronically caused by the properties of the systems' parts plus their manner of combination are also a priori deducible from, or (which, I take it, amounts to the same) logically necessitated by the properties of the systems' parts, plus their combination.

Van Cleve, however, suggests that there is little reason for the emergentists to accept the more limited thesis of mereological rationalism or, *a fortiori,* the general thesis of causal rationalism.[252] Instead, argues Van Cleve, they can grant Nagel the claim that causation requires necessitation, but insist that all that causation requires is weaker, nomological necessity. The emergentists could thus, Van Cleve thinks, claim that emergent phenomenal properties are caused by physical properties of the brain in virtue of being nomologically necessitated by these properties. This sort of necessitation, however, is quite compatible with a lack of logical necessitation of the effect by the cause.[253]

Van Cleve tells us that nomological necessity is “intermediate in strength between logical necessity and Humean regularity”.[254] Perhaps his concept of nomological necessity is based on the intuitive idea that events in nature produce or bring about one another without us being able to simply

[250]Nagel (1979, p. 186).
[251]Van Cleve (1990, p. 218).
[252]See Van Cleve (1990, p. 217).
[253]Van Cleve (1990, p. 218).
[254]Van Cleve (1990, p. 217).

deduce the effects from the causes. This sort of necessity with which a cause brings about an effect is, arguably not the absolute necessity of logic, as it seems at least conceivable that the laws of nature could be different – metaphorically, god could have created the world with slightly or vastly different laws of nature.

Consider, for example, Newton's third law of motion (i.e. the law of action and reaction). Intuitively, this law expresses a natural, or nomological necessity, however is not, arguably, logically necessary. There seems to be no logical contradiction in saying that when two objects collide, they experience forces which are not equal in magnitude or opposite in direction. Similarly, there seems to be no contradiction in thinking that it could be the case that the speed of sound is not 343 metres per second. The idea behind Van Cleve's objection is then that emergent consciousness is in this way brought about or produced by the brain without us being able to deduce that it will be so produced.

Van Cleve's argument is not quite uncontroversial. One could reply to it that once we truly know a thing, in the sense of having a complete knowledge of its nature, we will be, in principle, able to deduce all its possible effects given that we also have knowledge of its environment. The thought here is that in virtue of knowing everything about what a thing is, we would also know everything about its dispositions to act in a particular way given a particular stimulus. The laws of nature are on this conception reducible, roughly, to dispositions of things.[255] The fact that Van Cleve thinks that there can indeed be nomological necessity without logical necessity seems to reveal that he works with a conception of laws of nature, according to which these are contingent with respect to what the given thing essentially is, although these laws, at the same time nomologically necessitate the ways the thing behaves.[256] I will, however, leave this issue aside in what follows.

Van Cleve thus provides us with an interesting defense of emergentism against Nagel's critique. If his appeal to nomological necessity is plausible, the emergentists are not committed to the counterintuitive view that causes do not necessitate their effects. In the following section, I shall argue that Van Cleve's objection also applies to an argument against emergentism advanced by Galen Strawson.

5. Strawson Against Emergence

As we saw, for Galen Strawson, any philosophical position which holds that physical ultimates are non-conscious while certain configurations or systems of these ultimates are conscious, is a version

[255]See e.g. Bird (2005).
[256]See e.g. Armstrong (1983).

of emergentism. In this section I shall try to explain why Strawson believes that any such position faces grave problems and must be rejected. The trouble with emergentism is, Strawson argues, that it invokes a brute and inexplicable relation between the physical system which consciousness is supposed to emerge from, on the one hand, and conscious experience itself, on the other hand, a relation which, according to Strawson, amounts to a miracle.

Recall that Strawson sees an important disanalogy between the case of liquidity, sometimes offered as a clear case of emergence by the emergentists, and the case of consciousness. In the former case, we have, Strawson suggests, at least a rough understanding as to why and how the emergent phenomenon arises while in the latter case we lack even a rough grasp. Strawson goes on to emphasize that the mind-body case is in this respect analogous rather to other hypothetical cases in which we have no grasp as to how one kind of phenomena could emerge from another kind of phenomena, and in which a suggestion that one could emerge from the other would sound quite absurd to us. One such case is hypothetical emergence of extended phenomena from utterly non-extended phenomena, such as mathematical points.[257] Another such case would be hypothetical emergence of spatial phenomena from wholly non-spatial phenomena standing in wholly non-spatial relations.[258]

Given that these hypothetical cases look impossible or even absurd, shouldn't we also rule out the emergence of experiential phenomena out of utterly non-experiential phenomena? Strawson suggests here that we should since, as he puts it, the divide, between the non-experiential domain and the experiential domain is even more fundamental than the divide between the non-extended and the extended or between the non-spatial and the spatial.[259]

Strawson's appeal to analogies is, however, unlikely to make the emergentists change their minds about the matter. They would, I expect, after all, simply deny that the emergence of experience is impossible, while perhaps accepting that it is counterintuitive or even puzzling.[260] They could, moreover, emphasize that their adherence to emergentism is consistent with allowing that the two cases discussed by Strawson are indeed impossible.

As I see it, however, Strawson's appeals to absurd cases do not themselves establish his argument against emergentism and are instead meant to show the inadequacy of the liquidity analogy. What then is Strawson's real argument against emergentism? I think the argument is based on the thought that if consciousness were, *per impossibile*, emergent, the relevant relation of emergence would

[257]Strawson (2006, p. 15).
[258]Strawson (2006, p. 17).
[259]Strawson (2006, p. 17–18).
[260]See e.g. Nida-Rümelin (2007a, p. 281).

need to be brute or unintelligible and as such it would amount to a law-like miracle.[261] Reliance on miracles is, however, not merely in conflict with the broadly naturalistic spirit of emergentism, but it also arguably amounts to giving up scientific investigation in general.

While denying that the example of liquidity could help us make sense of the emergence of consciousness, Strawson thinks that this case, nevertheless teaches us an important lesson. Namely, it teaches us that emergence in general must be, at least in principle, fully perspicuous or intelligible. This means, reasons Strawson, that emergence cannot be *brute*, i.e. unintelligible even in principle, even to an ideal reasoner; in each case of emergence there must then, according to Strawson, be a genuine explanation as to how and why the emergent phenomenon emerges and why it has the nature it does.[262] What would such an explanation amount to? Arguably, it would identify the feature or features of the lower-level phenomena in virtue of which consciousness emerges, and show us how and why it emerges. Strawson writes:

> *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.*[263]

As long as we hold that the brain's constituents are wholly non-experiential or non-conscious, reasons Strawson, there is nothing about them in virtue of which consciousness could emerge from them. Given that, however, the emergence of consciousness is, according to Strawson, unintelligible, even in principle, even to an ideal reasoner, i.e. brute.

Why not, however, simply accept that emergence is brute in this sense? Strawson offers us two considerations against this option. Firstly, Strawson argues that the very concept of brute emergence is contradictory since the concept of emergence implies intelligibility. If it does, then surely "brute emergence" is a *contradictio in adjecto* akin to e.g. "married bachelor".[264] The emergentists could, however, reply to Strawson here that they are not committed to his use of the notion of emergence – indeed, as I showed earlier, many emergentists define emergence in a way that allows for the unintelligibility of the relation. This is clear, for example, from the above quoted remark of Alexander, that emergence of new qualities in nature admits of no explanation and should be accepted with natural piety. It seems therefore that Strawson's claim that the notion of brute emergence is contradictory will be susceptible to the objection that the meaning he attributes to the term is far from being universally accepted. A conceptual stipulation of this sort will therefore hardly help Strawson here and he will need to offer a different argument in order to cast doubt on

[261] Strawson (2006, p. 18).
[262] Strawson (2006, ibid.).
[263] Strawson (2006, p. 18).
[264] Strawson (2006, ibid.).

the possibility of brute emergence.

As already mentioned, Strawson's other consideration against brute emergence appeals to the claim that brute emergence would amount to a miracle. He writes, for example:

> *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*.[265]

According to Strawson, it thus follows from the very notion of brute emergence that it amounts to a miracle. A miracle can perhaps be seen as an event which cannot, even in principle, be explained in terms of any *natural* causes and therefore gets attributed to supernatural, perhaps divine, causes.

How are we to understand Strawson's argument here? As we saw, according to Strawson, the notion of brute emergence implies that there is nothing about the base-phenomenon in virtue of which the supposed emergent phenomenon arises and has the nature it does. If, however, that is the case, then, presumably, the emergent phenomenon cannot be explained in terms of any natural causes and therefore its existence amounts to a miracle.

Does Strawson's argument against brute emergence work? One problem with it is, as I see it, the objection which Van Cleve raises against Nagel's argument also applies to Strawson's argument. Let me explain why.

6. Strawson and Van Cleve's Objection

As I see it, Strawson's argument, just like the one offered by Nagel, presupposes mereological rationalism, a thesis which, as Van Cleve shows us, can easily be rejected by the emergentists. One way to see this is to realize that Strawson's notion of in-virtue-ness is ambiguous. When Strawson claims that the notion of brute emergence implies that there is nothing about the base-phenomenon in virtue of which the supposed emergent phenomenon could arise from it, he uses the notion of "in virtue of" in what can be called a *logical sense*. In this sense of "in virtue of", A arises in virtue of B as long as the instantiation of B is, given the background conditions, a reason for the instantiation of A, which, when stated, makes A's instantiation perspicuous or transparent. The felt transparency of the explanation is, I take it, a direct consequence of the fact that B without A is inconceivable which means that we cannot even imagine or conceive of the existence of B without the existence of A. This, however, means that if A arises in virtue of B, then B's instantiation logically necessitates A's instantiation.

[265]Strawson (2006, ibid.).

We can see that Strawson uses his notion of "in-virtue-ness" in the logical sense if we consider a possible reply the emergentists could offer in reply to Strawson's critique. They could reply that, *contra* Strawson, there actually is something in virtue of which consciousness arises, namely the specific complexity and organisation of the brain processes of the relevant conscious organism. Clearly, such a reply would not satisfy Strawson, since the complexity of the brain, combined with its physical properties, fails to make it perspicuous for us as to why the brain gives rise to consciousness. Strawson would surely insist that the appeal to complexity fails to make the emergence of consciousness any less brute.

The problem with offering complexity as an answer to the question as to why the brain gives rise to consciousness, is that the appeal to complexity brings about further questions, such as why it is that the complexity of a particular kind and degree gives rise to a particular phenomenal property, or any phenomenal property at all. Surely then, such an answer does not amount to a transparent explanation. This lack of transparency or perspicuity is further indicated by the fact that we are arguably able to conceive of zombies who share the complexity of our brain but are not conscious. If then the emergentist suggests that conscious experience emerges in virtue of the relevant kind of complexity of the brain phenomena, he is not using the notion of "in-virtue-ness" in the logical sense but rather in what we can call *nomological sense*. Given, however, that this kind of reply would clearly not be accepted by Strawson, he himself must be using the notion of "in-virtue-ness" in the logical rather than the nomological sense.

Strawson then holds something which can be expressed in the following conditional statement:

> S: If there is nothing about the base-phenomenon in virtue of which the supposed emergent phenomenon arises and has the nature it does, the emergent phenomenon cannot be explained in terms of any natural causes and therefore its existence amounts to a miracle.

Is, however, S plausible? We have seen that Strawson is entitled to the claim that in cases of brute emergence, there is nothing about the base-phenomenon in virtue of which the emergent phenomenon arises, as long as "in virtue of" relation is understood in the logical sense. The given implication then amounts to something like the following statement.

> S_L: If there is nothing about the base-phenomenon which would make it, at least to a super-scientist or an omniscient god, perspicuous and intelligible as to why and how the emergent phenomenon arises and has the nature it does, then the emergent phenomenon cannot be explained in terms of any natural causes and its existence therefore amounts to a miracle.

Conditional S_L would be plausible if a complete causal explanation of a property were necessarily fully intelligible or transparent. If however, a full causal explanation of a property can be provided

without the explanation being perspicuous, the implication does not hold.

What would a complete causal explanation of an instantiation of a property amount to? A natural reply here would be that such an explanation would list all the causes which together suffice to produce the instantiation of the property. Would such an explanation be necessarily also transparent (in that it would make the instantiation of the property intelligible)? I do not see why it would need to be. While the complete causal explanation would describe the micro-properties and micro-events which physically necessitate the instantiation of the macro-property, it is far from clear that truths about the micro-properties and micro-events would then also need to logically necessitate truths about the macro-property which is something which arguably renders an explanation transparent.

We should at this point be able to see a close parallel with Nagel's reasoning. As I emphasised earlier, after all, the key complaint which Nagel expresses about the emergentist picture is that according to this picture, the brain, as long as it is purely physical, includes nothing which could *necessitate* the emergence of consciousness. We, however, saw that Nagel's requirement of logical necessitation may well be too strong to be acceptable for the emergentist who can, as argued by Van Cleve, make do with physical or nomological necessitation. We should now be able to see that Strawson, just like Nagel, in effect requires that properties of complex systems are fully transparently explainable and thus logically necessitated by the properties of the systems' components and their combination. In other words, both thinkers presuppose mereological rationalism, a view which the emergentists have little reason to accept.

We can now see that both of the presented anti-emergentist arguments are affected by Van Cleve's objection in the sense that they both rely on the truth of mereological rationalism, a thesis which is controversial and indeed challenged by ontological emergentists.

While I think that emergentism escapes the two arguments, it faces, as I see it, at least two serious challenges which we are now in a better position to see. Both of these challenges have to do with the issue of causation. The first challenge, if successful, leads to the conclusion that ontological emergentism has difficulty accounting for the intuitive view that consciousness has its place in the causal chains in the natural world, i.e that consciousness causally contributes to nature. The other charge, if plausible, is even more serious as it threatens the general plausibility of ontological emergentism.

7. The Causal Argument

A number of thinkers have brought up the issue that emergentism has a serious problem with causation. Jaegwon Kim, for example, explicates the worry that emergent properties will be necessarily causally preempted by their underlying physical properties.[266] While I find his argument interesting, I shall focus here on a slightly different problem which has been expressed by David Papineau in the form of his causal argument for physicalism.[267] The causal argument is perhaps the strongest argument for physicalism and its primary target is dualism in all its forms. Given that ontological emergentism, as I have presented it here, amounts to a version of property dualism, it will clearly be affected by Papineau's argument. The argument in effect aims to establish that the proponents of ontological emergentism need to choose between the unattractive options of epiphenomenalism, systematic causal over-determination and interactionism – as Chalmers puts it, they face a sort of trilemma. I shall try to clarify these three notions and the reasons for their unattractiveness in the course of discussing the individual premises of the causal argument.

The argument has the following structure:

(1) Conscious mental states have physical effects.

(2) All physical effects are fully caused by physical processes.

(3) Physical effects of conscious mental causes are not always over-determined by distinct causes.

(4) Mental events are wholly grounded in physical events.

_____

(5) Physicalism is true.

The argument looks valid, although, of course, the ontological emergentist may try to avoid its conclusion by rejecting one or more of its premises. We shall see, however, that the denial of each premise brings along significant theoretical costs. Let me now consider the premises of the causal argument one by one.

Here premise (1) is very plausible as it denies that conscious mental states are utterly causally inefficacious with respect to physical processes in the universe. In other words, premise (1) amounts to the denial of *epiphenomenalism* about consciousness. If epiphenomenalism is true, phenomenal properties have no physical causal effects whatsoever. Using the famous comparison introduced by Thomas Huxley, we can say that if phenomenalism is true, then the casual significance of conscious

[266]Kim (1999).
[267]Papineau (2002, pp. 17–18).

events with respect to the physical world amounts to the causal significance of a steam-whistle of a steam engine with respect to the operation of this engine.

While epiphenomenalism is not a fatal problem for a theory of consciousness, incurs considerable theoretical costs. A psycho-physical theory which embraces epiphenomenalism will need to deal with many counter-intuitive consequences. It will, for example, need to explain how we can, as it seems, react to our phenomenal states – intuitively, after all, pain triggers retraction behaviour, an enjoyable taste of pistachio ice-cream makes us come back for more, etc. In all of these cases we seem to be dealing with physical or wholly physically grounded verbal or behavioural effects of what are, according to the ontological emergentists, non-physical phenomenal properties and the epiphenomenalists will need to provide us with an account of the occurrence of these effects which does not appeal to mental-physical causal chains. This is, I take it, a reasonably strong reason to accept premise (1).

Premise (2) articulates the principle of the causal closure of the physical. This principle has been rejected by some anti-physicalists and there is no knock-down argument for it. Still there are strong reasons to hold this principle which have to do with the principle of conservation of energy, one of the keystones of contemporary physics.[268] Denial of premise (2) amounts to an acceptance of *interactionism*, the thesis that there are causal interactions between non-physical consciousness and the physical world. Interactionism, once again, is a possible position, although one which brings about considerable theoretical costs.[269] Apart from denying causal closure of the physical, the proponents of interactionism will also need to explain how two domains which are – ontologically speaking – radically different can nevertheless causally affect one another. This, of course, is the old philosophical problem a form of which was once brought to Descartes's attention by Princess Elisabeth of Bohemia. As it is far from clear that there could exist a solution to this problem, there is, I take it, a strong case in favour of premise (2).

Premise (3) amounts to the rejection of over-determination of all the physical events which are caused by mental events. In cases of causal over-determination there are two or more causes which bring about a single effect such that each of these causes would have been sufficient, considered on its own, to bring about the effect. While causal over-determination is certainly a possibility – consider, for example, an unlucky man who has been shot and at the same time struck by lightning – denying premise (3), however, means accepting the strong, and I think highly implausible, thesis that any case in which a mental event causes a physical event is a case of causal over-determination. This claim – sometimes called systematic causal over-determination – certainly sounds very odd.

[268]See e.g. Papineau (2001).
[269]See e.g. Nida-Rümelin (2007a, p. 277) for a position which embraces interactionism and rejects the principle of the causal closure of the physical.

If all the complex physical effects of our mental life may just as well be caused by purely physical, non-mental processes, why then have we evolved to be beings with minds and not zombies instead? Systematic causal over-determination certainly feels in tension with the great sense of economy which nature normally exhibits as well as with the principle of Occam's razor (why postulate two distinct causes of a given event if one could produce the event on its own?). While there is then no knock-down argument against systematic causal over-determination, I take it that there is a fairly strong case for (3).

Premise (4) follows from premises (1) – (3). If, after all, as (1) tells us, conscious events sometimes have physical effects, the causal closure of the physical is true, as premise (2) tells us, and if we rule out systematic causal overdetermination, as (3) tells us, we get the result that mental events must be physical or wholly grounded in physical events, which is just what (4) tells us. Why, however, think that conclusion (5) follows from premise (4)? Here one can appeal to the standard definition of physicalism according to which the doctrine amounts to the claim that all things and events in the universe are wholly grounded in micro-physical entities.[270] With this definition of physicalism on the table we can see that premise (4) implies (5), i.e. the thesis of physicalism.

How can an ontological emergentist react to Papineau's causal argument? Here one example is Nida-Rümelin's dualist emergentism, a view which rejects the causal closure of the physical world, emphasizing that there is no decisive proof for the closure.[271] Another option is to accept epiphenomenalism, a strategy which has been explored in detail by Chalmers in *The Conscious Mind* but which he has since grown less sympathetic to.[272] While these and other replies are available to the ontological emergentist, each of them, nevertheless brings about serious theoretical costs which I have sketched above. I take it then that the causal argument should at least motivate the proponents of the non-reductive approach to consciousness to look for theoretical alternatives which would not bring about these serious drawbacks.

8. The Problem of Upward causation

It seems to me that there is another and perhaps more fundamental challenge which the emergentists need to face. To see this, consider what the result of our discussion of Van Cleve's objection was. There we saw that the emergentists can rely on mereological necessitation of the emergent phenomenon by the base-phenomena. This appeal, as I see it, in effect amounts to the claim that there is a psycho-physical law which states what the physical conditions of emergence are. Here the

[270]See chapter 2 for more on this.
[271]Nida-Rümelin (2007a, p. 280).
[272]Chalmers (2010, p. xiii).

conditions of emergence amount to the kind of physical state which the brain must be in if it is to give rise to the non-physical emergent phenomenal properties. Further, there will presumably be more specific conditions of emergence for more specific phenomenal states. These will specify what the neural correlates of the specific phenomenal states, such as having a pink after-image or feeling thirsty, are. Of course, given the emergentist commitment to the upward determination discussed above, these correlations have their roots in underlying nomological necessities, according to the emergentists. One day, the emergentist can dream, we will have a full understanding of which brain state gives rise to which emergent phenomenal state. Of course, these bridge laws will not be perspicuous and it will still be possible to ask just why this and not another set of psycho-physical laws holds, but that is, despite Nagel's and Strawson's reservations, the best we can do, the emergentists will presumably say.

How, however, are we to understand the claim that a physical state gives rise to a phenomenal state, which, according to the dualist version of emergentism espoused here, must be non-physical? I think we have here, once again, the commitment to interactionism on the side of the ontological emergentist. If, after all, the physical states are supposed to have any sort of influence on the emergent states, which I take to be the crucial claim of emergentism, then there must be upward causation or some sort of upward determination closely akin to causation. How could there be, however? Remember that we are dealing with two ontologically utterly distinct domains. The emergent properties do not have, according to the ontological emergentist, any hidden physical structure or aspect. Similarly, the physical brain, at least when it comes to its individual parts, has, according to the emergentists, no phenomenal or even protophenomenal properties or aspects. Given that, however, it is far from clear that the latter could causally produce the former. We certainly do not know of any uncontroversial cases which could help us clarify how it could.

We can see that this problem of interactionism is in one way even more pressing for the emergentists than the one raised by Papineau. There the problem was how non-physical consciousness could affect the physical processes in the world and the emergentist had there the escape route of epiphenomenalism. If, on the other hand, the emergentists are unable to account for how emergent properties can be given rise to at all, their position is doomed. In another way, the current problem of psycho-physical interactionism is, admittedly, easier for the emergentists than the one espoused by Papineau. The upward causal determination is, after all, compatible with the causal closure of the physical world which, we saw, is very plausible.

How could the emergentists reply to this objection? I think they would simply insist that despite our intuitions, it is not ruled out that certain complex systems produce properties which are ontologically new with respect to the properties of the systems' parts, even if their organisation is

taken into account. While this answer is available to the emergentist, it brings about the question as t o *how* these ontologically new properties are produced. Here I think the emergentist will presumably simply say that there is no explanation available. But such an answer will hardly satisfy the critic who will just return to his original claim that the causal production of phenomenal properties by purely physical properties seems utterly impossible.

Perhaps a more promising reply the emergentists can offer here is to appeal to the old-Humean view that causation amounts to mere regularity and there is then no question of how the physical can produce the non-physical. Still, I think the critics of emergentism can reply here with at least two things. Firstly, they can say that surely the intuitive view, which according to the proponents of the new-Humean reading was also held by Hume,[273] is that causation amounts to more than regularity and it is far from clear that there is a good reason to hold the opposite. Secondly, they could say that even if the old-Humean view is correct, the situation of the emergentist will be difficult. Namely, they will argue that in all known cases of causation we are dealing with causation between ontologically homogeneous, namely physical, events. The emergentist view will thus need to rely on an exception – only in the mind-body case we get the phenomenal out of the physical. Exceptions, however, are always suspect in science.

It seems to me that the problem of upward causation brings about yet another related worry for the emergentist. We can call this worry the problem of *ontological conservation*. Here the thought is simple. It seems that the emergentist is committed to the view that emergent properties, despite being produced by the physical brain, must somehow appear as something out of nothing. That, however, seems highly implausible and indeed utterly non-intelligible.

The thought behind the ontological conservation problem is that even if, for the sake of argument, we allowed that physical brain processes could cause phenomenal properties, the latter cannot be constituted by the former or even by some parts of the former – we are dealing, after all, with dualist emergentism where there are two radically distinct domains of properties. This, however, makes it mysterious how these phenomenal properties, which are viewed as non-physical, could occur or come into existence in the first place. The emergentist will probably reply here that they are indeed caused by the physical processes happening in the brain. I think, however, that this reply does not really sort out the issue as it appeals to a mysterious kind of causal chain – one which starts from a physical process and ends with the occurrence of a phenomenal property. As I see it, somewhere in this causal chain from the physical to the phenomenal, there would need to be a place where something radically new and non-physical comes into existence. Given that this new element will be non-physical, it will need to have nothing, ontologically speaking, to do with the preceding

[273]See e.g. Hill (2012).

physical process, except, of course that it is, *ex hypothesi* caused by the directly preceding event. In particular it will not – being non-physical – be, wholly or partially, constituted by the entities and properties involved in the preceding event. The claim that it is caused by the preceding physical process will, therefore, as I see it, fail to remove the feeling that the new element will be an ontological something out of nothing.

It is plausible, to conclude, that the anti-emergentist considerations which I have discussed in the second part of this chapter pose a considerable challenge for ontological emergentism. At the very least, I think, they should be properly addressed by the emergentists. Before the emergentists provide satisfactory replies to them, we would, as I see it, do well to explore alternative non-reductive views. This at least is what I shall try to do in the following chapter.

6. Panpsychic Universe

1. Russellian Monism

In the previous chapters I argued that there are good reasons to be sceptical of the prospects of the materialist reduction of phenomenal consciousness as well of strong emergentism about consciousness. The obstacles these accounts face lead us to the question whether there is a more plausible account of the existence of consciousness in nature. In this chapter I shall attempt to sketch and defend a view which, as I shall argue, can at least point us in the right direction. This approach, which has undergone a modest revival in the last decade or two, is inspired by the view introduced by Bertrand Russell in his 1927 book *The Analysis of Matter* and is referred to as Russellian monism in the literature.[274] Despite its name, it would be wrong to identify Russellian monism with the actual view held by Russell – in fact it is fair to say that some versions of Russellian monism would almost certainly be rejected by him.

Russellian monism can be characterised as the view meeting the following four conditions:

(1) Structural (or extrinsic) properties of microphysical entities are grounded in quiddities (or intrinsic properties).

(2) Quiddities lie, even in principle, outside the reach of physics.

(3) Quiddities play or occupy the roles defined by physics.

(4) Quiddities have a close relation to consciousness.

Let us now consider these four conditions one by one. Condition (1) uses the distinction between structural properties and quiddities. This distinction is based on the thesis – a version of which was held by Russell – that physics describes micro-physical entities, properties and relations in terms of their structural characteristics. These entities are thus characterised by physics in terms of the roles they play in the physical world. An electron, for example, is defined, roughly, as an entity which attracts protons, repels other electrons, neither attracts nor repels neutrons, etc. The physical conception, we can say, captures the role-properties of the electron, or – which amounts to the same – its structural properties.[275]

Russell and his current followers, who we can call neo-Russellians, insist that the nature of

[274]See e.g. Alter – Nagasawa (2015), Chalmers (2010, pp. 133–137).
[275]Why are the role-properties also called here structural properties? The reason is that these properties are possessed by the given entity entirely in virtue of occupying a certain position (and playing a certain role) in the causal or nomic structure described by physics.

fundamental micro-physical entities is not fully exhausted by their role-properties.[276] They hold that there must be some properties which realise the fundamental dispositions or roles described by the physical theory. The property which plays or occupies the theoretical role of the given fundamental entity can be called a quiddity. Quiddities are then the properties which play or realise the fundamental roles attributed to fundamental micro-physical entities by a fundamental physical theory.[277] Quiddities are then, for the neo-Russellians, distinct from the role-properties; while the former are, we can say, the *realisers*, the latter are the *realised properties*.

Russellian monism is often characterised in terms of the distinction between *intrinsic* and *extrinsic* properties instead of the distinction between structural properties and quiddities. Here extrinsic properties are, roughly, the properties which an entity possesses in virtue of being related in particular ways (causally, nomically, spatiotemporally, etc.) to other entities, while intrinsic properties are the properties which an entity possesses merely in virtue of being itself. While this intuitive distinction to an extent corresponds to the distinction between structural properties and quiddities, it is, I think, best for the neo-Russellian to work with the more technical distinction between structural properties and quiddities as the intrinsic/extrinsic distinction is notoriously difficult to make sense of philosophically.[278]

So much for condition (1) of my characterisation of Russellian monism. Conditions (2) and (3) have, hopefully, been partially clarified in what has already been said. According to condition (2), quiddities lie, even in principle, beyond the reach of physical methods. We have seen that the neo-Russellians hold that physics only informs us about the structural properties of microphysical entities while quiddities are by definition non-structural, which means that they must lie outside the reach of physics.[279] At this point, an opponent could object that even if we accepted that current physics tells us nothing about quiddities, it is not clear that some future physics could not provide us with knowledge of their nature. Physics, after all, has undergone multiple paradigm shifts so it is, the objector could say, only reasonable to expect that more are to come and perhaps one of these will help us uncover the nature of quiddities.

To this the proponent of Russellian monism can reply that the current state of physics gives us no good reason to expect such a development, in fact, the general tendency when it comes to progress

[276]See e.g. Strawson (2006, p. 10).
[277]Chalmers (2015, p. 254).
[278]See e.g. Langton – Lewis (1998).
[279]Although this insight is usually attributed to Russell, it is important to notice that Russell is by no means the only one who thinks that our physical knowledge is in this way necessarily limited. More recently, David Lewis has argued for the thesis of "Ramseyan humility" according to which we are ignorant of the intrinsic properties which ground the dispositional properties that physics tells us about (Lewis 2009). According to Rae Langton, it is, interestingly, this epistemological thesis which Kant had in mind when he talked about our irremediable ignorance of the *noumena* (Langton 1998).

of physics seems to be rather opposite. While in the past (roughly, before the 20th century) we had, for example, a conception of atoms as solid, extended, indivisible, etc. entities, our current knowledge of the micro-world consists more or less of 'pointer readings' – we seem to know little more than how the given micro-physical entity affects measuring devices and perhaps other micro-physical entities. Moreover, Russellian monists can appeal to the fact that they can make do with a weaker thesis than (2), call it (2*), which states that *some,* not all quiddities are in principle beyond the reach of physics. While the Russellian monists do not have a clear proof that some future physics will not uncover the nature of all quiddities, given the current state and methods of physics, as well as the described historical tendency, it is plausible to hold that at least some quiddities are necessarily beyond the reach of physics.

Condition (3) which states that quiddities need to play or occupy the roles defined by physical science, is a part of the above-provided definition of quiddities and I shall not therefore comment on it any further. Condition (4) which requires that there be a close or intimate relationship between quiddities and consciousness, on the other hand, deserves discussion. There are two basic options for the neo-Russellians when it comes to understanding the "close relation" between quiddities and consciousness. Either the neo-Russellians can hold that quiddities are micro-phenomenal properties, or they can hold that quiddities are micro-protophenomenal properties. We can say that those who hold that quiddities are micro-phenomenal properties, subscribe to a panpsychist version of Russellian monism, or simply to Russellian panpsychism. Those, on the other hand, who hold that quiddities are micro-protophenomenal properties, hold a panprotopsychist version of Russellian monism, or simply Russellian panprotopsychism. I take it that an entity instantiates a phenomenal property if there is something it is like for the entity to be itself.[280] I shall use the term "micro-phenomenal property" to refer to phenomenal properties instantiated by micro-physical entities and shall contrast micro-phenomenal properties with macro-phenomenal properties, which are instantiated by ordinary macro-entities, such as you, me or probably a rabbit, at least when we are awake or dreaming.

The notion of a protophenomenal property deserves some discussion. At first approximation we can say that protophenomenal properties are the properties which, although not themselves phenomenal, can collectively – when appropriately combined or organised – constitute, realise or give rise to phenomenal properties. This definition, however, is too permissive as it allows for mainstream physicalism to count as a version of Russellian monism. Physicalists, after all, typically hold that physical properties of the brain, although not themselves phenomenal, collectively constitute or realise phenomenal properties when appropriately combined or organised. If so, however, these

[280]See chapter 1 for further discussion.

physical properties will count as protophenomenal, which clearly goes against the spirit of Russellian monism and which means that the present definition of protophenomenal properties fails to establish Russellian monism as a self-standing non-reductive ontological position.

How can this unwelcome result be avoided? Given the above-introduced definition of Russellian monism, true protophenomenal properties need to be quiddities, which means that they cannot be structural properties. Such a constraint will prevent a priori physicalism from counting as a version of Russellian monism. Recall, after all, that according to a priori physicalism, a view discussed in chapter 2, phenomenal properties are grounded in/or constituted by structural (i.e. functional or behavioural) properties of physical systems. With this constraint in place, we get a new definition of protophenomenal properties as properties which are (a) not themselves phenomenal but which collectively constitute, realise or give rise to phenomenal properties and which are (b) non-structural.

While this definition helps us distinguish Russellian monism from a priori physicalism, things are much less clear when it comes to a posteriori physicalism, the view discussed in chapters 3 and 4. Clearly, a posteriori physicalism will not count as Russellian monism if the proponents of this physicalist approach hold that phenomenal properties are grounded in structural properties. Some a posteriori physicalists, however, deny that phenomenal properties are grounded in structural properties, or at least deny that phenomenal properties are wholly grounded in structural properties and hold instead that phenomenal properties are grounded (at least in part) in quiddities.[281] These versions of a posteriori physicalism would thus still count as versions of Russellian monism which, once again, is an unwelcome result for those who view Russellian monism as a self-standing ontological position.

This result can, as suggested by Chalmers, be avoided by imposing a further constraint on protophenomenal properties. Namely, he thinks that the "close relation" between quiddities and phenomenal properties, which is mentioned in the definition of Russellian monism, should be articulated in terms of the requirement that truths about phenomenal properties must be a priori entailed by truths about protophenomenal properties (perhaps together with truths about structural properties).[282] This means that it must be possible, at least for an ideal reasoner upon sufficient reflection, to a priori deduce the truths about phenomenal properties of the given organism from the complete set of truths about protophenomenal properties (perhaps in conjunction with truths about structural properties) of the given organism. We thus get a new definition of protophenomenal properties as properties which are (a) not themselves phenomenal but which can collectively

[281]See e.g. Lewis (1995, 2009), Papineau (2002).
[282]Chalmers (2015, p. 260)

constitute, realise or give rise to phenomenal properties, which are (b) non-structural, and which are (c) such that truths about them (perhaps in conjunction with some structural truths) a priori entail truths about phenomenal properties. Constraint (c) in this definition will prevent a posteriori physicalism from counting as a version of Russellian monism as, according to a posteriori physicalists, truths about phenomenal properties are precisely *not* a priori entailed by truths about physical properties. It will also, given constraint (b), prevent a priori physicalism from counting as a version of Russellian monism.

One could worry whether constraint (c) is not too restrictive; perhaps there could be a strong emergentist version of Russellian monism according to which phenomenal properties emerge from appropriately organised quiddities which are protophenomenal. Strong emergentism, however, as we saw in the last chapter, tells us that truths about the emergent phenomenon are not even in principle a priori entailed by truths about the base phenomena. If so, then truths about the emergent phenomenal properties would not be a priori entailed by truths about quiddities which means that the relevant quiddities could not, after all, count as protophenomenal. Isn't then this latest constraint (expressed by condition [c]) a bit too restrictive? Here I think we should reply that this kind of view should count as a version of special, quiddistic emergentism, rather than of Russellian monism as the relevant protophenomenal quiddities arguably do not have a sufficiently close relationship to phenomenal properties, which is why strong emergence is required in the first place.[283]

One could, moreover, question the motivation for such an emergentist view. As I tried to show, one reason for scepticism about emergentism is precisely the lack of intelligible connection between the emergence base and the emergent phenomena. It is indeed this lack of intelligibility brought about by ontological emergentism, which motivates some thinkers to explore Russellian monism. Quiddistic emergentism of the kind just mentioned would however, hardly be of any help here as according to this view the truths about the emergence base, which consists in instantiation of appropriately organised quiddities, would not a priori entail truths about phenomenal properties and, arguably, intelligibility requires a priori entailment. For these reasons I shall here, following Chalmers, integrate both of the above-mentioned constraints into the definition of protophenomenal properties.[284]

One question which can be raised in connection with this characterisation of Russellian monism is what is meant here by 'monism'. In the course of my discussion of condition (2) I suggested that Russellian monism, strictly speaking, only requires that some quiddities are to be necessarily

[283]While there could be panpsychist views which also deny the a priori entailment, there the "closeness" of quiddities to phenomenal properties is guaranteed by the fact that the quiddities are themselves phenomenal.
[284]I take it that my characterization of Russellian monism is very close to the one recently introduced by Alter and Nagasawa (Alter – Nagasawa [2015]).

beyond the reach of physics. That means, however, that it may be that only some quiddities are phenomenal or protophenomenal. These claims seem to imply that there may be multiple kinds of quiddities, which casts doubt on the claim that we are dealing with a truly monistic view. What, however, is meant here by "monism"? I take it that there are two basic varieties of this ontological position: thing-monism and stuff-monism.[285] According to thing-monism, a version of which was held by Spinoza, the whole universe is really a single thing, which objects, such as tables, mountains or planets, are only small parts of. It is quite clear that the neo-Russellians are not committed to thing-monism, although some have lately explored this alternative.[286] According to stuff-monism, everything in the universe is wholly grounded in one kind of stuff. A prime example of stuff monism is, of course, physicalism itself. While Russell clearly had stuff monism, rather than thing-monism in mind when he called his view "neutral monism", it is not clear how serious he was about the "monism" part. He, after all, allows that while some physical events may have the intrinsic nature of our percepts, others may not have this intrinsic nature.[287] Most neo-Russellians are equally permissive about the use of the term 'monism' and I shall be as well.

2. Russellian Panpsychism

In the course of discussing requirement (4) I introduced two versions of Russellian monism: Russellian panpsychism and Russellian panprotopsychism. In what follows I shall describe these two views in more detail and try to evaluate their respective merits and problems. In particular, I shall argue that the best option available to the Russellian monists is the view called constitutive Russellian panpsychism, according to which macro-phenomenal properties are constituted by complex systems of micro-physical entities exhibiting primitive micro-phenomenal properties. This conclusion is, on its own, not too original within the current panpsychist debate – it has been, among others, recommended by Chalmers. I shall, however, attempt to propose new reasons to reject the main competing views of constitutive Russellian panpsychism, namely, panqualityism and emergent Russellian panpsychism. In the following chapter I shall then attempt to suggest a way for the constitutive Russellian panpsychist to start solving the combination problem – which is commonly seen as the most important challenge for this interesting approach to consciousness.

According to Russellian panpsychism, quiddities are phenomenal properties. The view can be seen as one version of the more general doctrine of panpsychism, the doctrine which has recently regained some popularity in philosophy.[288] Panpsychism is usually understood as the view that

[285]See e.g. Strawson (forthcoming).
[286]See Shani (2015).
[287]See e.g. Russell (2015, p. 46).
[288]See e.g. Chalmers (1996, 2015, forthcoming), Rosenberg (2004), Strawson (2006), Skrbina (2009), Blamauer (2011),

fundamental physical entities instantiate micro-phenomenal properties, i.e. that there is something it is like for these entities to be themselves.[289] We can also describe panpsychism as the view that fundamental physical entities have primitive experiences or are endowed with primitive phenomenally conscious states.

One can make this definition of panpsychism weaker by holding instead that panpsychism is true iff all members of at least some fundamental physical kinds have micro-phenomenal properties.[290] Behind this weaker definition is the plausible thought that even if it turned out that there is a kind of fundamental physical entities whose members do not have phenomenal properties while all the members of the other kinds do, we would, intuitively, still allow that panpsychism is true in our universe. The weaker definition brings about questions as to just how sparse, or spatially constrained, the phenomenality-endowed fundamental entities can be in order for panpsychism to still be true. Given, however, that the definition talks about fundamental entities, one can presume that these will not be extremely sparse or spatially constrained. Still, I am happy to allow that there may be grey areas in which it will not be clear whether panpsychism, as just defined, will be true.

The term "panpsychism" literally means that everything has a mind, or rather a soul (*psyché*). From the point of view of the contemporary debate this is misleading in at least two respects. Firstly, panpsychism, as it is usually understood, does not entail the thesis that microphysical entities have minds, at least not in the sense in which we normally talk about mind as something which involves beliefs and desires, moods, emotions, intelligence, the ability to think, etc. Instead, panpsychists normally attribute merely primitive micro-phenomenal properties to fundamental micro-physical entities. Further, panpsychism, as it is usually understood, does not involve the thesis that everything has micro-phenomenal or phenomenal properties. Indeed, the proponents of panpsychism do not normally hold that macroscopic things, such as chairs, rocks, cities or planets, considered as wholes, possess phenomenal properties, although such a thesis is compatible with panpsychism, as defined above. Moreover, they do not normally hold that abstract objects, such as numbers or geometrical points, have phenomenal properties.

Admittedly, even with these disclaimers in place, panpsychism goes against many of our ordinary beliefs, perhaps even against "common sense" and its central claim that fundamental micro-physical entities have (primitive) experiences or feelings and that there is something it is like to be them for themselves may strike some as rather extreme and dubious. One may, in this respect, raise the question as to what kinds of experiences could photons, quarks or similar entities, which of course have no sense organs, possibly have. Here the panpsychist can answer that these micro-experiences

Müller – Watzka (2011), Goff (forthcoming 2).
[289]See e.g. Strawson (2006, p. 25).
[290]Chalmers (2015, p. 246).

are presumably much more primitive than even our simplest experiences – for example an experience of a single colour shade. Even such a monochrome experience can, after all, arguably be viewed as complex and consisting out of hue, brightness and saturation with these consisting of further, yet more primitive elements.[291] Our ignorance regarding the nature of micro-experiences and micro-phenomenal qualities should not on its own prevent us from attributing certain primitive experiences to fundamental micro-physical entities. We can see this if we consider the case of bats, famously discussed by Nagel. We do not have and perhaps could not have a good idea as to the nature of the experiences of these echolocating creatures, but it would be unfair to deny that they have experiences merely in virtue of our ignorance of the nature of these experiences. Such a denial would amount to a gross conflation of epistemology and metaphysics. It certainly seems that there could be experiences in the universe whose natures are quite alien to us who have the kind of experiences we have.

Another objection one may wish to raise against panpsychism is that the view is simply crazy or absurd. It is important to notice, however, that this on its own, is not a good objection against the view and may well be a result of our upbringing in the western intellectual climate of the 20th and 21st centuries. Things will presumably appear somewhat different to someone brought up, for example, in India or Southeast Asia.[292] Moreover, versions of panpsychism or closely related views have arguably been held by many significant figures of Western philosophy.[293] This, I believe, gives us some reason to keep an open mind as to whether panpsychism is true, especially given the peculiar kind of difficulty of the problem of consciousness. It is, of course, possible that we shall one day be able to rule out the truth of panpsychism on scientific or philosophical grounds, but one should not do so merely because the view conflicts with some of the pre-theoretical beliefs integral to our Western culture.

This brings us to the important point that panpsychism, although a rather speculative doctrine, is entirely compatible with the discoveries of contemporary science and its proponents happily accept the existence of the theoretical entities posited by correct scientific theories. At the same time, however, they, of course, need to deny that the repertoire of real natural properties is exhausted by the properties sciences inform us about. This repertoire will, according to the panpsychists, need to also include micro-phenomenal properties which should be attributed to the entities posted by the fundamental physical theory, whatever these turn out to be.

[291]See Schroer (2010) for a physicalist approach which views phenomenal qualities as complex. See also chapter 4 of the present volume.
[292]See Parkes (2009) for an argument that a version of panpsychism is widespread in traditional Japanese, Chinese and Korean thinking.
[293]See Skrbina (2005) for a historical overview. See also Tollar (2012) for an argument that the Czech phenomenologist Ladislav Hejdánek is a panpsychist.

Admittedly, allowing that panpsychism is logically coherent and not in conflict with the contemporary scientific worldview is still far away from having a good reason to endorse the view. I shall now therefore explore some arguments for panpsychism which have appeared in recent literature.

3. Argument to the Best Explanation

One influential line of reasoning in favour of panpsychism starts from noticing the difficulties associated with physicalism and strong emergentism and goes on to suggest that panpsychism is a more promising alternative. According to this consideration we thus need to attribute phenomenal properties to fundamental physical entities because it is the best way to account for the existence of consciousness in the physical universe. This kind of argument has been famously discussed by Galen Strawson.[294] One can, of course, view this whole volume as an extended attempt to vindicate this argument for panpsychism. The discussion which follows will therefore only draft the outline of this argument rather than go into detail about its individual premises.

We can see this kind of argument as exhibiting the following structure:

1. All concrete macro-properties are constituted, or given rise to by complex systems of fundamental micro-physical entities.

2. Macro-phenomenal properties are concrete macro-properties.

3. Macro-phenomenal properties are constituted, or given rise to by complex systems of fundamental micro-physical entities.

4. Macro-phenomenal properties could not be constituted out of or given rise to by fundamental micro-physical entities if these possessed no micro-phenomenal properties.

———————

5. Fundamental micro-physical entities have micro-phenomenal properties.

The argument looks valid, however, its premises require some discussion. Although most action in the argument happens in premise (4), it will help to discuss each of the argument's individual steps. Premise (1) tells us that all concrete macro-properties are constituted, or given rise to by complex systems of fundamental physical entities.[295] This is a disjunctive claim whose first part states that all

[294]See e.g. Strawson (2006, forthcoming).

[295]Those who insist that properties are never concrete but always abstract can feel free to substitute "property instances" for "properties" as presumably, they will allow that property instances are concrete. I take it that nothing of substance hangs on this in the discussion that follows.

concrete macro-properties (i.e. properties of macro-physical, or ordinary things) are constituted by complex systems of fundamental physical entities. This claim is very close to a common definition of physicalism, according to which all (concrete) entities, properties and relations are wholly grounded in fundamental micro-physical entities, properties and relations. As a result, most physicalists will also accept the constitution claim which entails premise (1). This premise should, moreover, be also acceptable for the emergentists, as it, being disjunctive, allows that some macro-properties need not be constituted by their underlying micro-entities but rather are given rise to by these – which is, I take it, just what the emergentists typically hold.

Some may wish to object at this point that premise (1) begs the question against (non-emergentist) substance dualism since presumably immaterial subjects, posited by that doctrine, feature concrete properties, such as presumably, phenomenal properties, yet these properties will clearly not be constituted out of or given rise to by fundamental micro-physical entities. Firstly, after all, it is difficult to think of immaterial souls as being constituted by – or arising from – anything more fundamental as it is far from clear what this more fundamental stuff is supposed to be. Secondly, even if we allowed that immaterial souls are constituted by or arise from some more fundamental entities, these entities would certainly not be physical. If, after all, they were physical, the soul as a whole could hardly be immaterial.

I think the best reply to this objection is to allow that the argument is only conditionally sound, i.e. it is sound insofar as substance dualism is false. Despite that, however, the argument is still interesting because substance dualism has few supporters in current philosophy as the view brings about some high theoretical costs. Firstly, the substance dualist will presumably need to deny that consciousness in any way results from evolutionary processes in nature. Intuitively, consciousness is highly beneficial to organisms who possess it so it is natural to think that it must have evolved. If, however, consciousness is a property of an immaterial soul or if it is indeed identical to an immaterial soul, it is hard to make sense of the claim that it has evolved. Accepting, however, the supposition that consciousness has not evolved, it is far from clear that the substance dualist can explain how and why immaterial souls, with all their complex structure, have come into existence without needing to appeal to divine intervention. Even though premise (1) is then only conditionally sound, given that there are fairly strong reasons to reject substance dualism, the argument is certainly still worth our attention.

Premise (2) states that phenomenal properties are concrete properties such as say the physical property of having mass or the biological property of being an ostrich. This premise will be rejected by phenomenal anti-realists, such as Dennett, who deny the reality of phenomenal properties. I argued against positions of this sort in chapter 2 so here I shall simply assume that they are

implausible. Premise (2) could also be challenged by those a priori physicalists who hold that phenomenal properties are functional properties and these are not concrete properties but rather abstract properties realised by concrete realisers – the neural processes in our brains. Here a natural reply is that the premise, as well as the whole argument, can easily be rephrased in terms of property realisations which the functionalists should have no problem accepting as concrete. Rephrased in this way, the functionalists should have also no problem accepting premise (2).[296]

While premise (3) is directly implied by (1) and (2) and is thus plausible, insofar as the first two premises are, premise (4) is controversial and its full defense would require much discussion. This premise tells us that macro-phenomenal properties could not be constituted or given rise to by complex systems of fundamental physical entities if these micro-entities had no micro-phenomenal properties. This is, of course the key insight behind contemporary panpsychism, a version of which was once well expressed by William James, who wrote that "[i]f evolution is to run smoothly, consciousness in some shape must have been present at the very origin of things".[297] James's remark, of course, concerns the diachronic evolution of consciousness rather than the synchronic mereological relation but still, I believe, well expresses the basic insight which can be simply expressed as the thought that you cannot make phenomenally conscious things by combining phenomenally non-conscious things.

However this may be, thesis (4) will surely be rejected by the emergentists who indeed hold just that. I argued against emergentism in detail in the previous chapter so for now I shall assume that this position is implausible, at least in its ontological variety which is relevant for our purposes. Premise (4) will also likely be rejected by non-eliminativist materialists who hold that systems of non-phenomenal fundamental physical entities constitute phenomenal properties. I argued against materialism at length earlier in the book so for now I shall simply assume that the view is implausible. Premise (4) will also be rejected by panprotopsychists who hold that macro-phenomenal properties are constituted by complex systems of fundamental micro-physical entities which feature micro-protophenomenal properties. While I think the panprotopsychist reply is much more promising than the materialist and the emergentist replies to (4), I shall later in this chapter argue that even this reply faces serious challenges.[298]

Premise (4) is then admittedly controversial. I am hoping, nevertheless, that the present volume as a whole demonstrates that there are strong reasons to accept (4) and, as a result, that the panpsychist

[296]For the sake of simplicity, I shall remain here with the original phrasing of the argument in terms of properties.
[297]James (1890, p. 149, original in italics).
[298]One could also object that (4) is, strictly speaking, ambiguous as it is not clear whether all fundamental micro-physical entities would need to have micro-phenomenal properties in order for macro-phenomenal properties to arise, or whether it would suffice that only some, not all, fundamental micro-physical entities would need to have micro-phenomenal properties. While I think that my arguments only support the latter option, I shall leave this issue aside here as my definition of panpsychism only requires the latter, weaker reading of premise (4).

conclusion (5) ought to be at least seriously considered. Still, I think it can be considered a weakness of this argument for panpsychism that it requires many additional arguments in order to establish premise (4). Of course, I attempt to provide these arguments in this volume but the very fact that they are needed makes this line of argument for panpsychism admittedly quite complex.

4. The Intrinsic Nature Argument

Another argument for panpsychism which deserves at least a brief mention here goes back to Russell's discussion of neutral monism in *The Analysis of Matter* and has been recently revived by William Seager.[299] This argument, called the "intrinsic nature argument" by Seager, is based on the considerations of the scope of our physical knowledge in combination with general metaphysical considerations concerning the nature of consciousness. We can sketch the argument as follows:

(1) Science reveals to us only the structural properties of physical entities.

(2) Structural properties require quiddities for their instantiation.

(3) There are no non-phenomenal quiddities.

_____

(4) Quiddities are phenomenal properties.

While I think that this argument is valid, each of its premises requires some discussion. Premise (1) is based on Russell's and others' considerations about the nature of our physical knowledge. I discussed the justification for this thesis earlier in this chapter and here I shall therefore consider it to be reasonably justified.[300] Premise (2), which was also briefly discussed above, tells us that role-properties or structural properties require quiddities for their instantiation. It will be worth adding here to what has already been said that perhaps the most important challenge for this premise is the view in the philosophy of science called "ontic structural realism" according to which the universe consists of pure structure which does not need quiddities for its instantiation.[301] While such a view is certainly counterintuitive, it is unclear that it is false or non-intelligible, which casts doubt on premise (2). The panpsychist relying on the intrinsic nature argument would therefore need to provide us with a good reason to deny ontic structural realism and it is not clear how such an argument would go. Premise (3) is, unfortunately for the panpsychist, also controversial as it relies on the falsity of Russellian panprotopsychism. Once again, the panpsychist would need to provide us with a reason to reject panprotopsychism. I shall discuss some of the reasons for scepticism

[299]See e.g. Seager (2006a, 2006b).
[300]See, however, Ney (2015) for a recent challenge to the thesis expressed by this premise.
[301]See e.g. Ladyman (1998) for considerations in favour of such a view.

about panprotopsychism later in this chapter.

The intrinsic nature argument for panpsychism is then interesting but, as even this brief discussion revealed, rather controversial and its proponents would need to provide us with reasons to reject ontic structural realism as well as panprotopsychism. Given these issues, I think it is a good idea for the panpsychist to look for an argument with lighter, less controversial premises.

5. The Hegelian Argument

Another argument in favour of panpsychism has recently been introduced by Chalmers.[302] It is called by Chalmers the "Hegelian argument" not because it would aspire to accurately represent Hegel's thinking but rather because it possesses the three-part dialectical structure of thesis, antithesis and synthesis, traditionally attributed to Hegel. An interesting point about the Hegelian argument is that it combines and synthesises different arguments which have previously appeared in the philosophical literature, namely the conceivability argument against physicalism and the causal argument against dualism.

How does the Hegelian argument go then? The thesis, and the starting point of the argument, is mainstream physicalism. Given, however, that there is, Chalmers suggests, a strong argument against physicalism, namely the conceivability argument, we need to take the step to its antithesis – dualism. Dualism, however, Chalmers thinks, faces another strong argument, namely the causal argument for physicalism, which I introduced towards the end of the last chapter. This argument, together with the conceivability argument, should thus lead us to the synthesis – Russellian panpsychism. Unfortunately, for the panpsychist, Chalmers's quasi-Hegelian considerations do not stop here and instead continue to a new antithesis – panprotopsychism. I shall, however, focus now merely on the first part of these considerations which constitutes Chalmers's Hegelian argument for Russellian panpsychism.

The first step of Chalmers's quasi-Hegelian dialectic leads us from physicalism (thesis) to dualism (antithesis). While Chalmers's motivation behind this step is the conceivability argument, one could in principle also be motivated by a different anti-physicalist argument, such as the knowledge argument or the argument appealing to revelation, which I suggested in the chapter 4. Although these arguments are admittedly somewhat controversial, if the considerations in the previous chapters are sound, they together, I take it, establish a fairly strong case against physicalism. I will therefore presume here that the first step of the Hegelian argument is well justified.

[302]Chalmers (2015).

The second step of Chalmers's quasi-Hegelian dialectic leads us from dualism (antithesis) to Russellian panpsychism (synthesis). The motivation behind this step is Papineau's causal argument for physicalism.[303] As we saw in the previous chapter, the causal argument in effect attempts to show that the proponents of dualism need to choose between the unattractive options of epiphenomenalism, systematic causal over-determination and interactionism.

Recall the general structure of the causal argument:

(1) Conscious mental states have physical effects.

(2) All physical effects are fully caused by physical processes.

(3) Physical effects of conscious mental causes are not always over-determined by distinct causes.

(4) Mental events are wholly grounded in physical events.

_____

(5) Physicalism is true.

The causal argument is usually considered to be the most persuasive argument for physicalism. Physicalism itself is, however, according to Chalmers, already discredited by the conceivability argument. Moreover, as we have seen, even if we reject the conceivability argument, there are other anti-physicalist arguments which may be able to do the job of the conceivability argument. This should then lead us to embrace Russellian panpsychism as the synthesis of physicalism and dualism and the view which is, according to Chalmers, unaffected by neither the conceivability argument, nor the causal argument. This, of course, gives Russellian panpsychism a considerable advantage over both physicalism and dualism.

Let us first consider the kind of reply the Russellian panpsychism can offer to the proponents of the conceivability argument. Recall that the conceivability argument infers the possibility of a world in which all physical truths about the actual world, call their complete set *P*, hold but in which some phenomenal truth, call it *Q,* does not hold, from the ideal conceivability of such a world. The first premise of the argument then is the claim that *P* & non-*Q* is conceivable. Here the Russellian panpsychist can argue that we need to distinguish two senses of the term "physical truth" used in the premise. In one sense, the term "physical truth" means structural truth, i.e. truth which concerns the instantiation of structural properties (or role-properties) only. In the other sense the term "physical truth" refers to truths which concern the instantiation of structural properties as well as quiddities. We can, following Chalmers, call the first kind of truths "narrowly physical" and the

[303]Papineau (2002, pp. 17–18).

second kind of truths "broadly physical".[304] Moreover, we can call the properties described by narrowly physical truths "narrowly physical properties" and call the properties described by broadly physical truths "broadly physical properties". Similarly, we can distinguish between narrowly physical processes, i.e. those which consist in instantiation of narrowly physical properties and are therefore described by narrowly physical truths, and broadly physical processes, i.e. those which consist in instantiation of broadly physical properties and are described by broadly physical truths.

Having made this distinction it is easy to see that the conceivability premise is ambiguous. In one sense the premise states that it is conceivable that all narrowly physical truths about our world hold without an arbitrary phenomenal truth about our world holding (such as the truth that Peter is having a red experience). Here Russellian panpsychists can agree but, luckily, they can also accept the conclusion of the conceivability argument thus understood – namely the claim that narrow physicalism is false. Narrow physicalism, after all, amounts to the doctrine that truths about consciousness are grounded in narrowly physical truths, which is rejected by the Russellian panpsychist who holds that truths about consciousness are grounded in broadly physical truths, i.e. the conjunction of truths about quiddities and structural truths.

This brings us to the other reading of the conceivability premise. In this other sense the premise states that it is conceivable that all broadly physical truths about our world hold without an arbitrary phenomenal truth about our world holding. There is, however, no obvious reason for the Russellian panpsychist to accept such a claim. Given that it is integral to Russellian panpsychism that there are quiddities which have a close relation to consciousness, Russellian panpsychists seem to have a clear reason to reject the conceivability premise and thus avoid the conclusion that broad physicalism is false.[305] Such a conclusion would be in conflict with their view as Russellian panpsychism amounts to a version of broad physicalism – it is, after all, the view that consciousness is wholly grounded in truths about quiddities in conjunction with some structural truths.

It seems fair to conclude then that Russellian panpsychism can offer a persuasive reply to the conceivability argument. How, however, does this doctrine cope when it comes to the causal argument? There are, I believe, strong reasons to think that even this argument is ineffective against Russellian panpsychism. Let me now try to explain why.

The above-introduced distinction between narrowly and widely physical processes enables us to distinguish between two versions of premise (2) of the causal argument. These can be expressed as follows:

(2a) All physical effects are fully caused by broadly physical processes.

[304]Chalmers (2015, p. 255–256).
[305]In the next chapter which focuses on the combination problem, we shall see that this issue is more complicated.

(2b) All physical effects are fully caused by narrowly physical processes.

We can see that the Russellian panpsychists can happily accept (2a) as it is natural for them to hold that all physical effects are caused by causes involving quiddities and structural properties. Moreover, the Russellian panpsychist can also accept premises (1) and (3) of the causal argument. Given that the argument now includes premise (2a) which concerns broadly physical processes, premise (4) of the argument will consequently need to be modified to (4a) which says that mental events are fully grounded in broadly physical events. (4a) leads to the conclusion that (5a) broad physicalism is true. To summarize, the argument now looks like this (where the term "physical" is not qualified, I mean broadly or narrowly physical):

(1) Conscious mental states have physical effects.

(2a) All physical effects are fully caused by broadly physical processes.

(3) Physical effects of conscious mental causes are not always over-determined by distinct causes.

(4a) Mental events are fully grounded in broadly physical events.

_____

(5a) Broad physicalism is true.

As we saw, conclusion (5a) is fully compatible with the truth of Russellian panpsychism, as the view is a version of broad physicalism.

The physicalist could at this point insist that the Russellian panpsychist needs to accept the stronger premise (2b) which says that all physical effects are fully caused by narrowly physical processes, i.e. processes involving instantiation of structural properties without instantiation of quiddities (assuming that this option is intelligible). With that premise in place the argument will lead to stronger versions of theses (4) and (5), namely:

(4b) Mental events are fully grounded in narrowly physical events.

(5b) Narrow physicalism is true.

These claims are, of course, incompatible with the truth of Russellian panpsychism as this view holds that there are quiddities and that quiddities, perhaps together with some structural properties, ground phenomenal properties. The Russellian panpsychists, however, have a good reason to reject (2b) and, as a result, are not committed to (4b) and (5b). Namely, they can emphasise that, according to Russellian panpsychism, structural properties are realised by quiddities and there is no reason to think that quiddities are causally inefficacious. It is, indeed, natural to suppose that the

properties which realise structural properties of an entity also realise its causal powers and are thus causally efficacious. Importantly, then the Russellian panpsychists' denial of (2b) is well justified and does not commit them to the acceptance of the dubious thesis of interactionism. If so, however, the Russellian panpsychists are well justified in rejecting (4b) and (5b), the theses which, as we saw, are incompatible with the truth of their view.

If these considerations are sound, Russellian panpsychists have a good response to the causal argument for physicalism. That is, of course, a considerable merit of their view since, as already mentioned, this argument is usually seen as the most persuasive argument for physicalism. Given that the view can offer good replies to both the conceivability argument and the causal argument, it can be seen, as argued by Chalmers, as the quasi-Hegelian synthesis of physicalism and dualism and therefore also as the conclusion of the Hegelian argument, or at least of its first part.

How persuasive is this argument? We have seen that Chalmers justifies his step from physicalism to dualism with appeal to the conceivability argument. This argument, however, is – like all the other anti-physicalist arguments – controversial and it has been challenged by many physicalists. The second step of the Hegelian argument, leading us from dualism to Russellian panpsychism, is justified by means of the causal argument which will be challenged by those dualists who subscribe to epiphenomenalism, interactionism or allow for systematic causal over-determination. It is fair to say then that also the Hegelian argument for Russellian panpsychism is controversial. Supposing, however, that the reasons to reject physicalism, discussed in the previous chapters, are sound and that epiphenomenalism, interactionism and systematic causal over-determination are implausible, the argument provides us with a good reason to accept Russellian panpsychism, or at least to take the view seriously.

6. Non-Russellian Panpsychism

We have seen that if the Hegelian argument is sound, Russellian panpsychism is more plausible than physicalism or dualism. The argument however also has other implications which are worth mentioning. Namely, it gives us a reason to prefer Russellian panpsychism to the non-Russellian versions of panpsychism and, what is more, it casts doubt on the emergentist variety of panpsychism. Let me now consider these implications.

According to Non-Russellian panpsychism, micro-physical entities have micro-phenomenal properties but these properties do not ground micro-physical dispositions which means that they are not quiddities. A non-Russellian panpsychist can hold either that there are no quiddities, or perhaps that there are quiddities but they are not micro-phenomenal (or micro-protophenomenal). Given that

for non-Russellian panpsychists micro-phenomenal properties are not quiddities, their view is threatened by the causal argument as they will only be able to reject premise (2) of the causal argument at the expense of accepting interactionism. If, after all, physical processes involve instantiation of physical properties only, which means that they do not involve instantiation of any non-physical properties, and if, as (2) states, physical effects are caused by physical processes only, micro-phenomenal properties will face a real threat of being causally inefficacious. Of course, non-Russellian physicalists can simply reject (2) but then they will need to embrace the dubious thesis of psycho-physical interactionism. Equally unattractive is, however, as we saw, the rejection of (1) as it amounts to embracing epiphenomenalism, or the rejection of (3) as it amounts to embracing systematic causal overdetermination.

Couldn't, however, the non-Russellian panpsychist simply insist that even though micro-phenomenal properties are not quiddities, they are, nevertheless micro-physical properties and therefore the non-Russellian panpsychist can accept premise (2) as well as conclusion (5)? This terminological postulation is of course possible but the physicalist will then simply reply to the non-Russellian panpsychist that there are good reasons to accept premise (2*) which states that all physical effects are fully caused by physical* processes, where physical* processes are those physical processes which do not involve instantiation of any micro-phenomenal properties. The non-Russellian panpsychists can reject this reply but thereby they will embrace interactionism between phenomenal properties, which they view as physical, and physical* processes. Such interactionism will, of course, be just as implausible as the original psycho-physical interactionist thesis.

If these considerations are sound, the non-Russellian panpsychists will need to make the difficult choice between three unattractive options: interactionism, epiphenomenalism and systematic causal overdetermination. It is clearly a significant theoretical advantage of Russellian panpsychism that it, as I tried to demonstrate, manages to escape this trilemma.

7. Four Varieties of Russellian Panpsychism

One objection against Russellian panpsychism states that although the view allows for causal efficacy of micro-phenomenal properties, it is far from clear that it also allows for causal efficacy of macro-phenomenal properties. Given, however, that our consciousness involves instantiation macro-phenomenal properties, it is unclear whether Russellian panpsychism allows for causal efficacy of our consciousness. So, the objection goes, epiphenomenalism still threatens Russellian panpsychism.

As I see it, the force of this objection depends on how Russellian panpsychists conceive of the relation between macro-phenomenal properties and micro-phenomenal properties. There are four basic theoretical possibilities open for the Russellian panpsychist: identity, constitution, emergence and autonomy.[306] Let me now explore which of these options, if any, allow for causal efficacy of macro-phenomenal properties. Those Russellian panpsychists who view the relation between micro-phenomenal properties and macro-phenomenal properties as that of identity hold a version of *identity Russellian panpsychism*. For a proponent of identity Russellian panpsychism, there is no problem of causal exclusion of macro-phenomenal properties because if micro-phenomenal properties are causally efficacious and macro-phenomenal properties are identical with certain micro-phenomenal properties, then macro-phenomenal properties are *ipso facto* causally efficacious too.

Those Russellian panpsychists who view the relation between micro-phenomenal properties and macro-phenomenal properties as that of constitution hold a version of *constitutive Russellian panpsychism*. While the relation of constitution may be understood as one which includes as its special case the relation of identity (we can say that a thing or a property constitutes itself), its paradigm examples will be rather the relation between a whole and its parts or between a system and its individual components. Cases such as individual Lego bricks together constituting a model house, or individual H_2O molecules constituting the watery content of a glass come to mind as examples. It is plausible that in cases of constitution the complex whole inherits the causal efficacy of its component parts.[307] The reason for this is that in the cases of constitution, the component parts plausibly realise the complex whole and realising entities plausibly also realise causal properties of the realised entity. It seems, after all, inconceivable, given what the realising relation amounts to, that thing A would be realised by thing B, B would be causally efficacious but A would not be causally efficacious. Constitutive Russellian panpsychism then, just like identity Russellian panpsychism, enables causal efficacy of macro-phenomenal properties.

Those Russellian panpsychists who view the relation between micro-phenomenal and macro-phenomenal properties as that of emergence, hold a version of emergent Russellian panpsychism. If macro-phenomenal properties emerge from (complex configurations of) micro-phenomenal properties, then macro-phenomenal properties are neither identical with micro-phenomenal properties nor are they constituted by micro-phenomenal properties. We have seen in the previous chapter that the relation of emergence, as usually understood, is instantiated iff entities or properties of a higher-level domain arise out of entities of a lower-level domain but if – at the same time – truths about the higher-level domain are not even in principle deducible from the complete truth

[306]See Chalmers (forthcoming, p. 20).
[307]Chalmers (2015, pp. 257–258), Kim (1999, p. 16).

about the lower-level domain.[308] That, however, means that, according to emergent Russellian panpsychism, the emergent macro-phenomenal properties are something over and above the mere set of micro-phenomenal properties which are quiddities, the former are ontologically new with respect to the latter.

Is there space for causal efficacy of macro-phenomenal properties if emergent Russellian panpsychism is true? Given that the emergent macro-phenomenal properties are ontologically new with respect to the base-level properties, there is, unlike in the case of constitution, no reason to think that they inherit the causal efficacy of the micro-phenomenal base properties which they emerge from. Does it mean that emergent Russellian panpsychists need to face the trilemma of interactionism, epiphenomenalism or over-determination? Such reasoning would, as I see it, be too hasty. The mentioned trilemma will clearly threaten emergent non-Russellian panpsychism but when it comes to the Russellian version of the view, things are more complicated. To see why, consider once again premise (2) of the causal argument. This premise states that all physical effects are fully caused by physical processes. As we saw, Russellian panpsychists have a good reason to reject (2b) which says that all physical effects are caused by narrowly physical processes because they hold that quiddities are causally efficacious. At this point the constitutive and identity Russellian panpsychist can insist that a rejection of (2b) in combination with a rejection of epiphenomenalism and overdetermination does, nevertheless, not commit them to the implausible thesis of interactionism. Instead they are free to accept the weaker premise (2a) which amounts to the claim that all physical effects are caused by broadly physical processes, i.e. processes which involve instantiation of structural properties and quiddities. This step is, of course, not available to emergent non-Russellian panpsychists as, according to their view, neither micro-phenomenal nor macro-phenomenal properties are quiddities.

According to the advocates of emergent Russellian panpsychism, on the other hand, micro-phenomenal properties are quiddities and their causal efficacy is thus fully compatible with the causal closure of the broadly physical. How will macro-phenomenal properties stand when it comes to causal efficacy if emergent Russellian panpsychism is true? Here the answer depends on whether macro-phenomenal properties are quiddities. If macro-phenomenal properties are not quiddities and they are yet to keep their causal efficacy, emergent Russellian panpsychists will need to choose between the implausible theses of interactionism and causal overdetermination as even the causal closure of the broadly physical will not be acceptable for them.

Presumably, however, emergent Russellian panpsychists will hold that macro-phenomenal

[308]As I argued in chapter 5, an ontological definition of strong emergence is preferable to the prevalent epistemological ones, such as the one mentioned here, because it enables us to easily distinguish strong emergentism from a posteriori physicalism.

properties are also quiddities. Above, we defined quiddities as the realisers of fundamental micro-physical roles but it is not clear that there are any fundamental micro-physical roles which the emergent macro-phenomenal properties could realise or play. It seems, after all, that all the fundamental micro-physical roles are already played by micro-phenomenal properties. Perhaps, however, the emergent Russellian panpsychists can define quiddities more broadly as simply realisers of fundamental physical roles, leaving out the condition of micro-physicality. Supposing that there are then fundamental macro-physical roles which require quiddities in order to be realised but which at the same time cannot be realised by micro-physical quiddities, then perhaps macro-phenomenal properties are the quiddities which play these roles. That would, however, mean that causal efficacy of macro-phenomenal properties is compatible with the causal closure of the broadly physical even if emergent panpsychism is true.[309]

Given these considerations, it would seem that emergent Russellian panpsychism, just like the constitutive and identity varieties of Russellian panpsychism, escapes the causal argument and is thus among the views which work as the synthesis of Chalmers's Hegelian argument. It seems to me, however, that such a conclusion would be too hasty and that emergent Russellian panpsychism faces some serious challenges. Let me now explain why.

The main problem with the claim that if emergent Russellian panpsychism is true, then macro-phenomenal properties are quiddities, is the fact that it is far from clear that there are any fundamental physical roles for the emergent macro-phenomenal properties to realise or play. Presumably, after all, all macro-physical roles, are already realised by complex systems of fundamental micro-physical entities. These entities are, however, according to the Russellian panpsychist, already realised by micro-phenomenal properties. That means that the emergent Russellian panpsychists would need to hold that there are certain physical roles which are not realised by systems of micro-phenomenal properties – we can call these supposed physical roles "fundamental macro-physical roles". It is, however, far from clear that there are any fundamental macro-physical roles for macro-phenomenal quiddities, to play.

The emergent Russellian panpsychist could reply here that there are indeed emergent fundamental macro-physical roles which cannot be played by systems of micro-physical entities but instead must be played by emergent macro-phenomenal properties. Such a view, however, implies strong emergentism about some macro-physical entities and processes. This kind of emergentism, while arguably favoured by the founders of emergentism, such as Samuel Alexander, C. D. Broad or C. Lloyd Morgan, finds little support in contemporary science and is even more controversial than emergentism about consciousness, discussed in the previous chapter. What would the fundamental

[309]Here I put aside the problem of causal preemption of emergent properties in general, formulated by Kim (1999).

macro-physical roles which are supposedly realised by macro-phenomenal properties, amount to? One possible candidate here would be the biological property of being alive, which, according to some, is not reducible to its underlying chemical and physical processes. Even if we, however, accepted such irreducibility, it is not clear that macro-phenomenal properties could realise this property. In dreamless sleep we, arguably, lack phenomenal consciousness, but yet are alive. At the very least, the emergent Russellian panpsychists would need to provide us with an argument for such a far-reaching form of strong emergentism according to which there are fundamental macro-physical (perhaps biological) roles and would need to suggest possible candidates for such properties.

Notice, moreover, that it is not clear that the emergent Russellian panpsychists could deny that macro-phenomenal properties are quiddities, even if they were willing to pay the price of interactionism, overdetermination or epiphenomenalism. It is, after all, a key claim of Russellian panpsychism that micro-phenomenal properties also have an "outer", relational side which means that they are realisers of micro-physical roles. How, however, could the Russellian panpsychist deny that macro-phenomenal properties have this "outer" aspect too? Such a denial would be highly suspect as macro-phenomenal properties are supposed to be, ontologically speaking, the same kind of properties as micro-phenomenal properties – this indeed is precisely what renders Russellian panpsychism explanatory with respect to instances of macro-phenomenal properties.

Considerations about causal efficacy as well as general theoretical considerations should thus lead the emergent Russellian panpsychist to embrace the thesis that macro-phenomenal properties are quiddities. Such a view, as I have just argued, requires that there of fundamental macro-physical roles. It is, however, far from clear that there are fundamental macro-physical roles. Their existence is not merely called into question given the prevalent reductionist spirit of contemporary science but also by the consideration that there do not seem to be any fundamental macro-physical roles which could be played by macro-phenomenal properties. At the very least it would be up to the emergent Russellian panpsychist to show that fundamental macro-physical roles exist and provide us with examples of such roles.

If these considerations are sound, they give us, as I see it, a good reason to doubt the coherence of emergent Russellian panpsychism, a version of which has recently been introduced by Goff.[310] Goff's emergent Russellian panpsychism, is, unlike the type I have discussed so far, intelligible which means that truths about base-level existents a priori entail, at least in principle, truths about the emergent-level existents. Still, despite this intelligibility, it is clear that the emergent macro-phenomenal properties are for Goff something over and above the base-level properties. He, after

[310]Goff (2015a).

all, takes both micro-phenomenal facts and macro-phenomenal facts (called "o-phenomenal facts" by him) to be fundamental with the latter being caused by the former.[311]

If this is the case, however, Goff's position seems to result in the dilemma for emergent panpsychism which I described above: either there must be emergent fundamental macro-physical roles which cannot be played by mere gatherings or fundamental quiddities – which is a highly controversial claim and a significant bet on the future of natural science – or, alternatively, Goff needs to deny that macro-phenomenal properties are quiddities, which however, seems to cast doubt on the explanatory aspirations of emergent Russellian panpsychism. If, after all, macro-phenomenal properties are not quiddities while micro-phenomenal properties are, then it is not clear how the former properties could be ontologically the same kind of properties as the latter. Without this identity however, it is not clear how micro-phenomenal properties could be theoretically useful when it comes to accounting for the existence of macro-phenomenal properties.

Finally, let me consider autonomous Russellian panpsychism, according to which the relation between micro-phenomenal properties and macro-phenomenal properties is one of utter autonomy. Although an interesting view, the proponents of autonomous Russellian panpsychism, will face an even more serious version of the problems facing emergent Russellian panpsychism. It seems that, given their explanatory aspirations, autonomous Russellian panpsychists will need to hold that the autonomous macro-phenomenal properties are quiddities. Such quiddities will, however, require the existence are fundamental macro-physical roles which will need to be autonomous of micro-physics. We saw that there is little reason to hold that there are emergent fundamental macro-physical roles and it is even less plausible that there are utterly autonomous fundamental macro-physical roles. These roles, when realised, would, presumably, be sort of free-floating with respect to the micro-physical realm. That, however, goes directly against the spirit of the contemporary scientific world view according to which the macro-level is determined, at least loosely, by the micro-level. The Russellian version of autonomous panpsychism holds therefore, as I see it, little promise. Autonomous panpsychists could, of course, give up their Russellianism, i.e. the claim that macro-phenomenal properties are quiddities, but then they would be threatened by the causal argument in the way which I described above.

If the considerations concerning causal efficacy of macro-phenomenal properties are sound, the Hegelian argument gives us reasons to prefer Russellian panpsychism to non-Russellian panpsychism and, at the same time, to prefer identity Russellian panpsychism and constitutive Russellian panpsychism to emergent and autonomous Russellian panpsychism.

[311]Goff (2015a, p. 395).

8. Russellian Panprotopsychism

One objection which could be raised against the described conclusions of the Hegelian argument concerns the other above-introduced variety of Russellian monism, namely Russellian panprotopsychism. This objection states that Russellian panprotopsychism, just like Russellian panpsychism, is immune to both the conceivability argument and the causal argument and so the view also should be included in the synthesis of the Hegelian argument. If so, however, the argument speaks for Russellian panpsychism only if we suppose that Russellian panprotopsychism is false. Chalmers, who introduced the Hegelian argument, is aware of this complication and views Russellian panprotopsychism as the new antithesis to Russellian panpsychism which is for him, as we saw, the original synthesis of physicalism and dualism.[312] In order to amend the argument, the proponents of Russellian panpsychism would need to provide us with a good reason to prefer their view to Russellian panprotopsychism. In this section I will explore the question whether such a reason can be found. First, however, let me say a bit more about Russellian panprotopsychism.

According to Russellian panprotopsychism, quiddities are protophenomenal properties. Russellian panprotopsychism is thus a variety of panprotopsychism, the more general view that at least some kinds of fundamental physical entities have protophenomenal properties, which are not necessarily quiddities. This view can, at first sight, look more attractive than panpsychism, as it does not involve the controversial claim that micro-physical entities have experiences which feature phenomenal properties. There is then, according to panprotopsychism, nothing it is like to be a quark, electron or whichever micro-physical entity turns out to be included in the fundamental physical inventory. Fundamental micro-physical entities, according to panprotopsychism, only possess protophenomenal properties, i.e. special properties which have a close relation to consciousness, namely such relation that when protophenomenal properties gather and are appropriately organised, they collectively constitute consciousness with its macro-phenomenal properties.

Why, however, think that Russellian panprotopsychism is, just like Russellian panpsychism, immune to both the conceivability argument and the causal argument and therefore casts doubt on the Hegelian argument for Russellian panpsychism? Consider the conceivability argument first. Here the thought is that given the especially close relation of protophenomenal properties to consciousness, $P \,\&\, \sim Q$ will not be conceivable as long as we take P to be a complete set of broadly physical truths which, of course, includes truths about quiddities. The Russellian panprotopsychists can, after all, appeal to the fact that their definition of protophenomenal properties (i.e. the definition I described at the beginning of this chapter) requires that truths about macro-

[312]See Chalmers (2015, pp. 259–261).

consciousness are in principle deducible from truths about protophenomenal properties (perhaps in conjunction with some structural truths). Russellian panprotopsychists can, on the other hand, allow for the conceivability of *P & ~Q* as long as *P* is understood as a complete set of narrowly physical truths. In that case, the conceivability of *P & ~Q* will, nevertheless, not threaten their view, according to which narrowly physical facts do not exhaust micro-physical reality. The Russellian panprotopsychists' reply to the conceivability argument thus closely mirrors the reply provided by the Russellian panpsychists, discussed in much detail above.

Russellian panprotopsychism copes equally well when it comes to the causal argument. Its proponents, after all, have a good reason to reject causal closure of the narrowly physical as it is natural to attribute causal powers to protophenomenal properties given that these are quiddities. Russellian panprotopsychists can, on the other hand, happily accept causal closure of the broadly physical as well as the conclusion that broad physicalism is true because their view is a version of broad physicalism. Their reply to the causal argument thus closely mirrors the reply to this argument provided by the Russellian panpsychist, which I discussed in detail above. We can see that such a reply will not be available to the non-Russellian panprotopsychists as they can deny causal closure of the physical only at the expense of a commitment to interactionism, epiphenomenalism, or overdetermination, given that according to their view protophenomenal properties are not quiddities (the problem is exactly parallel to the one faced by the non-Russellian panpsychists). As a result, the non-Russellian version of panprotopsychism will not be immune to the causal argument.

Given that Russellian panprotopsychism is immune to both the conceivability argument and the causal argument, and given the above-mentioned intuitive appeal of the view, should it not be preferred to the arguably less moderate Russellian panpsychism? While I think Russellian panprotopsychism is certainly an interesting view, it also faces some serious challenges. Perhaps the most significant challenge to Russellian panprotopsychism is that it is not clear at all what protophenomenal properties are supposed to be. While, as we saw, we have a pretty firm grasp of phenomenal properties, our conception of protophenomenal properties is mostly negative: we know what they are not. Namely, we know that they are not phenomenal properties or structural properties, but we have little to no clue as to what sort of properties they are, positively speaking. Of course, protophenomenal properties were above defined as such properties that truths about them (perhaps in conjunction with some structural truths) a priori entail truths about phenomenal properties. Still, that arguably makes the position of the panprotopsychist even more troublesome while adding little positive content to the concept of protophenomenal properties, as it is not clear at all that there are any properties that could satisfy the requirement of a priori entailment. Indeed, one could argue that no properties we know of (with the possible exception of phenomenal properties

themselves which will be discussed later) satisfy this requirement.

It seems to me then that panprotopsychists would need to provide us with a clearer idea as to what protophenomenal properties are supposed to be if their view were to be taken seriously, especially given that the panpsychist alternative relies on properties which we arguably have a good grasp of.

Still, one can find a view in the history of philosophy which can be understood as a version of panprotopsychism and which does provide us with a clearer idea as to the nature of protophenomenal properties. Recall here the view of Wilfrid Sellars, sketched in chapter 1, according to which sensory qualities with their feature of ultimate homogeneity are parts of the physical world. A similar kind of view was arguably held by Russell himself who suggested that percepts are parts of the physical world. While according to many thinkers, Russell's view amounts to a version of panpsychism, idealism or perhaps phenomenalism,[313] I think it is best viewed as distinct from these views. Indeed I think we should take a hint from the fact that Russell himself called his view "neutral monism" and arguably understood percepts, considered on their own, to be neutral, that is neither physical, nor mental.[314] This is, of course, a highly interesting claim for anyone in search for protophenomenal properties as these properties are supposed to be neither phenomenal, nor physical in any traditional sense. A similar view can also be found in the works of Herbert Feigl who suggests that the physical world includes what he calls "qualities", rejecting at the same time panpsychism.[315] Feigl calls this view, using the term of C. S. Pepper, "panqualityism" and since this term is somewhat established in the current debate, I shall also adopt it here.[316]

9. Panqualityism

Panqualityism arguably provides us with some of the much needed content for the otherwise rather vacuous concept of a protophenomenal property. What, however, is this content supposed to be? One possibility, proposed by Chalmers, is to think of protophenomenal properties as of what he calls Edenic qualities.[317] In order to understand this concept, it will help to remember how one thought about the world as a child – back then, the sky seemed to us really, inherently or intrinsically blue, ripe strawberries were really, inherently or intrinsically red, grass was really, inherently or intrinsically green, etc. It seems to be a part of this naïve, or Edenic view of colours that these qualities really inhere in things and remain in them even after dark, in different lighting or

[313]See e.g. Chalmers (1996, pp. 154–155), Strawson (2010, p. 97).
[314]I argue for this point in Mihálik (2013).
[315]Feigl (1971).
[316]See e.g. Chalmers (2015), Chalmers (forthcoming), Coleman (forthcoming).
[317]Chalmers (2015, p. 272).

when nobody is looking.[318] Chalmers calls these, "true" colours Edenic because upon getting involved with philosophy or science, most of us "fell from Eden", i.e. stopped believing that Edenic colours and other Edenic qualities exist out there in the world.

What sort of view would we get if Edenic qualities were protophenomenal properties? Perhaps the best way to think about it is as of a view according to which micro-physical roles or dispositions instantiated in the universe are realised by Edenic qualities. The idea here is, once again, that when these qualities are instantiated in objects around us, grounding the microstructure of these objects, we have thanks to science cognitive access merely to the structural properties which these qualities ground, not to these qualities themselves. The only Edenic qualities which we can cognitively access when it comes to their intrinsic nature, and which we can thus know as qualities, are some of the Edenic qualities which are instantiated in our brains.

Could protophenomenal properties be Edenic qualities? We have seen that protophenomenal properties are supposed to be non-phenomenal and, clearly, Edenic qualities meet this condition. Phenomenal properties are properties of conscious states (or perhaps of conscious subjects) in virtue of which there is something it is like for a subject to be in those conscious states (or perhaps to be itself). When we think of grass as really, Edenically green, we are not thereby saying that there is something it is like for grass to be itself, we are not thereby attributing consciousness to grass. Further, we saw that protophenomenal qualities are supposed to be non-structural. Here, once again, Edenic qualities fit the bill as it seems that they are paradigm examples of properties whose nature is not exhausted by their role within a given causal or nomological structure of states. There surely seems to be more to Edenic red than a particular causal role, red is more than merely what red does. It seems then that Edenic qualities do meet the first two conditions which the above-formulated definition of protophenomenal properties requires.

Things seem less promising when it comes to the third requirement which states that truths about phenomenal properties are supposed to be a priori entailed by truths about protophenomenal properties (perhaps in conjunction with some structural truths). Here, as has recently been argued by Chalmers, qualities do not seem to be the right candidates for protophenomenal properties.[319] Indeed, it seems intuitively obvious that one can conceive of qualities being instantiated without any conscious experience at all being instantiated. The panqualityist will hold that qualities are quiddities and thus are instantiated in our brains (among other places), but still, it seems that one can conceive of the brain being made of qualitative, i.e. colourful, odorous, loud etc. stuff without the relevant organism being conscious at all.

[318]See Chalmers (2010, chapter 12, pp. 381–454) for much more on the Edenic view of colours and other qualities.
[319]Chalmers (2015, pp. 273–274).

These considerations can lead one to the hypothesis of qualitative zombies, introduced by Chalmers, who are creatures whose brains instantiate rich Edenic qualities, as these are quiddities which realise the particular micro-physical roles in the given brain, without these creatures being conscious.[320] Qualitative zombies are thus our replicas when it comes to the qualities instantiated in our experience, as well as our micro-physical replicas without being conscious. It seems that it is not any harder to conceive of qualitative zombies than it is of Edenically green grass, or Edenically blue sky without consciousness. The existence of qualities, on their own, does not seem to imply the existence of phenomenal properties (*qualities* do not imply *qualia*, if you wish) and adding structural or organisational properties to the admixture seems to be of little help.

Panqualityism has recently been advocated by Sam Coleman who has questioned the conceivability of qualitative zombies.[321] Coleman characterises qualities as unexperienced qualia and argues that these can serve as the protophenomenal properties posited by panprotopsychism. He allows that we need more than qualities themselves to get consciousness with phenomenal properties. In order to be able to provide a full account of consciousness, the panqualityist also needs an account of the awareness relation which, according to Coleman can be accounted for in terms of mental representation.[322] Only those qualities, instantiated by an organism's brain, will, according to Coleman, occur in the conscious field of the given organism, which are suitably represented by a higher-order thought. Coleman then complements his panqualityism with a higher-order-thought (HOT) account of awareness, inspired by the views of David Rosenthal.[323] It is important to notice that since Coleman appeals to the HOT theory in order to account for awareness, his approach to awareness is reductive – we are aware of a given quality, roughly, if it is in an appropriate causal relation with respect to the higher-order representing state and iff at the same time the representing state is in an appropriate causal relation to other mental states and/or behavioural outputs.

We are now able to see why, according to views of the kind which Coleman defends, qualities can be viewed as protophenomenal properties. If phenomenal consciousness can be conceived of as awareness of qualities and the relation of awareness is viewed as explainable in functional terms, it could be suggested that truths about qualities in conjunction with structural truths concerning the functional organisation of the brain will imply truths about phenomenal properties instantiated in consciousness. If some approach along these lines is plausible then, qualities meet all three criteria of protophenomenal properties: they are neither structural nor phenomenal properties and truths about them, in conjunction with structural truths, imply truths about phenomenal properties. As a result, panqualityism can be viewed as a promising version of panprotopsychism.

[320]Chalmers (2015, p. 273).
[321]Coleman (forthcoming).
[322]Coleman (forthcoming, p. 30).
[323]See e.g. Rosenthal (1991).

10. Qualitative Zombies

How plausible, however, is Coleman's panqualityism? As already mentioned, all panqualityist accounts face the threat of qualitative zombies, so it will be crucial for Coleman to be able to defend his view against this specific zombie attack. According to Chalmers, the qualitative zombie threat is indeed fatal for Coleman's view.[324] The problem, as Chalmers views it, is that Coleman's account of awareness of qualities is too reductive and so this account does not prevent conceivability of creatures whose brains instantiate qualities, without any awareness of these qualities. We can, following Coleman, call this particular kind of qualitative zombies "awareness zombies". Clearly, if awareness zombies, our qualitative replicas without awareness, are conceivable, Coleman's panqualityism will be susceptible to a relevant sort of conceivability argument. This conceivability argument has the following structure:

(1) $QQ\&\sim Q$ is conceivable.
(2) If $QQ\&\sim Q$ is conceivable, it is metaphysically possible.
(3) If $QQ\&\sim Q$ is metaphysically possible, panqualityism is false.

———————

(4) Panqualityism is false.

In this argument, QQ stands for the set of all truths about quality instances in conjunction with structural truths, and Q stands for an arbitrary phenomenal truth. As I see it, the most promising strategy for the panqualityists wishing to escape this argument will be to question premise (1). The panqualityists are, after all, unlikely to question the step from conceivability to possibility (given their typical motivation for rejecting physicalism). If, however, our qualitative and structural replicas without awareness are conceivable and if, as Coleman holds, consciousness requires awareness of qualities, it seems that premise (1) is plausible.

Why, however, think that our qualitative and structural replicas without awareness are conceivable? Here Chalmers's thought seems to be that awareness involves phenomenology and we are able to conceive of any functional state in the absence of phenomenology.[325] Since, Chalmers suggests, Coleman attempts to account for awareness by means of his (functionalist) HOT theory, his panqualityist proposal is ultimately implausible.

Coleman has recently offered an interesting reply to Chalmers's critique, questioning Chalmers's claim that the relation of awareness involves phenomenology. According to Coleman, awareness involves no phenomenology and so there is, pace Chalmers, no reason to think that his HOT theory

[324]Chalmers (2015, p. 273), Chalmers (forthcoming, p. 26).
[325]Chalmers (forthcoming, p. 26).

cannot account for awareness.[326] Why think that the awareness relation involves no phenomenology? Here Coleman first emphasises the fact that there are no clear examples of phenomenology which would pertain to awareness itself and not to the experienced qualities. Moreover, he offers an interesting argument to the effect that there could not, even in principle, be phenomenology of awareness itself, call it awareness-phenomenology, in addition to, and separate from phenomenology of experienced qualities, call it quality-phenomenology.

Coleman's argument goes roughly as follows: the proponent of awareness-phenomenology would need to, according to Coleman, choose whether the phenomenology pertaining to awareness of a particular experienced quality also involves phenomenology of that particular quality, or, alternatively, is completely separate and distinct from this quality-phenomenology. According to Coleman, however, both alternatives are unacceptable. If the particular awareness-phenomenology involved no phenomenology of the quality which the relevant act of awareness relates to, then it would be quite mysterious as to why the given act of awareness is awareness of this quality and not another. If, on the other hand, the particular awareness-phenomenology did involve the particular quality-phenomenology so that the particular quality-phenomenology was somehow suffused within the awareness phenomenology, then, presumably, we would get the quality-phenomenology twice – once as part of the awareness-phenomenology and once as separate quality-phenomenology. *Ex hypothesi*, after all, the particular quality-phenomenology is supposed to be distinct from the particular awareness-phenomenology. Given that neither of these two options is plausible, we should, Coleman argues, give up the idea that there is distinct awareness-phenomenology.

It seems to me that Coleman's argument gives us a good reason to be sceptical of the idea that there is distinct awareness phenomenology and I think, therefore, that his reply to the challenge formulated by Chalmers is fairly persuasive. If, after all, awareness involves no phenomenology, then Coleman's HOT account cannot fail for the reasons suggested by Chalmers.

Still, I think that the presented considerations should lead us to the question whether, quite apart from the issue of phenomenology of awareness, panqualityism is able to account for any phenomenology whatsoever, even phenomenology of experienced qualities. What will Coleman have to say here? Presumably he will say that macro-phenomenology which we experience is the result of the structuring and organising activity performed by the brain mechanisms on the qualities which are the brain's fundamental constituents. He indeed sometimes illustrates this point when he, metaphorically, talks about the HOT mechanism as of a spotlight, which illuminates various qualities instantiated in the brain while keeping other qualities in the dark.

[326]Coleman (forthcoming, p. 40)

It seems to me, nevertheless, that Coleman's conception is still threatened by qualitative zombies. We, after all, seem to be able to conceive of a creature whose brain instantiates rich Edenic qualities, think colours, sounds, smells etc., without there being anything at all it is like for the creature to be itself. Moreover, our conception will hardly be threatened if we add, insofar as we are able to, the intricate functional organisation of the kind which is, according to Coleman's HOT theory, needed for awareness. If such a scenario without consciousness is really conceivable, panqualityism is not immune to the conceivability argument. As far as I'm concerned, I do not think it is any more difficult to conceive of a brain constituted by an organised system of Edenic qualities without the relevant organism being aware of these qualities than it is to conceive of, say, Edenically red rose petals without them being aware of their redness.

At this point, presumably, Coleman will object that by properly conceiving of all those structured qualities in the brain of such a creature, I am really conceiving of its phenomenal states, or consciousness and am then by no means conceiving of a zombie. He could, moreover, support this contention by pointing out that when it comes to sensory qualities, the supposed qualitative zombie is just the same as its actual conscious twin. There, is, after all, as I have allowed, no extra awareness-phenomenology in one case which is missing in the zombie case. That this is the way Coleman would probably reply to my critique is indicated by the following passage where he writes:

> […] *zombie arguments depend on there being a sensory quality 'toggle' between the actual world and putative zombie world. One conceives of the absence of the relevant 'zombified' property by conceiving of the absence of its associated sensory qualities.*[327]

Here Coleman argues that his panqualityism is immune to the threat of the conceivability argument because there is no sensory quality 'toggle' between the actual world and qualitative zombie world. The thought here is that in order to be able to raise a conceivability argument against a particular theory of consciousness, the picture which the theory provides us with must be missing certain (or all) sensory qualities when compared to the actual world. It is this sensory difference which allows us to switch in our imagination, as by means of a toggle, back and forth between the zombie scenario and the actual world. This sensory difference is, of course very clear in the case of, for example, the functionalist theory of consciousness as, arguably, our conception of the functional replica of our world may not include any sensory qualities at all. Given, on the other hand, that we have allowed that the panqualityist world is not missing any sensory qualities when compared to the actual world, we cannot, according to Coleman, really conceive of the qualitative zombie world. Conceivability of a particular zombie scenario then requires, according to Coleman, that the zombie

[327]Coleman (forthcoming, pp. 39, fn. 71).

scenario is sensorily different from the actual world.

How plausible, however, is Coleman's reply? Clearly, if we allow that we can launch conceivability arguments only if the conceived zombie scenario is sensorily different from our world, the qualitative zombie threat is averted. Why, however, hold that? Unfortunately, Coleman does not give us a reason. As I see it, conceivability arguments, reasonably understood, do not require that the relevant twin scenarios are sensorily different but rather merely that they are different with respect to the *explanandum*, i.e. the entity, property or relation which we are trying to explain while being identical with respect to the *explanans,* i.e. the entities, properties or relations which are supposed to reductively explain the *explanandum*. As long as there is this sort of difference, it seems that a conceivability argument can be constructed.

As I see it, the qualitative zombie world is at least not a priori ruled out. It seems, after all, that the thought that an organism has a brain which is – at the micro-physical level – realised by Edenic qualities, which at the same time instantiates the relevant HOT mechanisms but which lacks phenomenology is not incoherent. Here, however, Coleman could object that if we understand, as Chalmers allows, phenomenal consciousness as awareness of qualities and at the same time allow that the awareness relation can, as it features no extra phenomenology, be reductively explained in terms of functional mechanisms, then qualitative zombies are in fact not conceivable. Truths about the distribution of qualities, together with structural truths about functional organisation would then, at least in principle, imply the existence of consciousness.

How persuasive, however, is Coleman's objection? It brings to light, I think, the inadequacy of his functionalist conception of awareness. We can see this if we compare this conception with an intuitive conception of awareness. It seems clear that if a given organism instantiates the relevant sensory qualities and at the same time features awareness of these qualities, intuitively understood, it is not conceivable that the organism could be a zombie. Consider now, on the other hand, an organism which instantiates the relevant sensory qualities and at the same time features awareness of those qualities, when awareness is understood as a sort of functional mechanism. It seems that nothing in this scenario rules out the thought that the given organism is a zombie and thus that it experiences nothing, despite having all those qualities in its brain. As a result, one should, I think, reject the reduction of awareness to a type of functional mechanism. Without such a reduction, however, the present objection is hardly persuasive and it seems therefore reasonable to think that qualitative zombies are at least negatively conceivable.

Perhaps, however, Coleman could object here that the conceivability argument requires a stronger notion of conceivability which we can call *positive conceivability.* As I see it, there are reasons to

reject this requirement, but I shall here, for the sake of argument, accept the legitimacy of such a requirement which I think could be suggested by Coleman. Do we then have some kind of positive conception of a qualitative zombie? It is natural to understand positive conceivability in terms of imaginability. Here we can expect that Coleman will emphasize that given that we lack a sensory "toggle" between a conscious organism and its zombie-twin, our imaginative representation of both will be identical. It seems to me, however, that there are reasons to resist Coleman's proposal here. As I see it, after all, my imaginative representation of a qualitative zombie will involve, insofar as we can imagine it, absence of conscious experience – we can say that the qualitative zombie will be – phenomenally – in a situation in which we are in dreamless sleep, there will be nothing at all it is like for it to be itself. Given that its conscious twin will, on the other hand, of course, have rich experience, there will be some sort of "toggle" available to us, after all. The "toggle", however, will not be sensory but rather phenomenal. It seems to me that this phenomenal toggle suffices to launch a conceivability argument against panqualityism.

Here Coleman could object that I am forgetting that the relevant qualities are instantiated in the brain of the supposed zombie. It is not clear to me, however, that our imaginative representation of what is instantiated in the brain of an organism needs to be a part of our imaginative representation of its experience or phenomenology. The fact that certain qualities are instantiated in a brain of an organism does not mean that they are there, experientially, *for* the organism. I seem to be able to imagine, for example, that panqualityism is true and I am – microphysically – made out of qualities and yet when I am in the state of dreamless sleep, I do not have any experience. Since, however, the difference between waking and dreamless sleep will be, for the panqualityist, purely a difference in functional workings of the brain, there seems to be nothing that prevents me from imagining, insofar I can imagine that I am made out of qualities and yet am in dreamless sleep, that someone who behaves and functions as if they are awake and instantiates qualities, is nevertheless a zombie, i.e. is – experientially – in the state I am in when I am in dreamless sleep. It seems to me then that there is a real sense in which we can imagine a panexperientialist zombie.

I think, moreover, that there is a good reason why Coleman should resist the view that conceivability of a qualitative zombie requires that the imaginative representation of the zombie is sensorily different from its conscious twin. If it were so, after all, it would be unclear whether unexperienced qualities – the basic unit of Coleman's metaphysics – are conceivable. Our imaginative representation of unexperienced qualities is, after all, sensorily speaking, hardly different from our imaginative representation of an experienced quality, if, as Coleman holds, the awareness relation brings about no additional phenomenology.

These considerations should, as I see it, lead us to the conclusion that until the panqualityists

provide us with a clearer reason why their view is immune to the conceivability argument, the threat of qualitative zombies seems to be still looming with respect to their view. It seems to me therefore that we do not have as of yet a viable conception of panprotopsychism. I have tried to show that the version of panprotopsychism which perhaps holds the most promise is panqualityism but I have argued that this view, at least in the version defended by Coleman faces the serious threat of qualitative zombies.

11. Conclusion

Where has the discussion led us with respect to our overall argument? I have tried to show that constitutive and identity Russellian panpsychism are preferable to emergent Russellian panpsychism because they, unlike emergent Russellian panpsychism, are not threatened by the causal argument. In the last section I have argued that constitutive and identity Russellian panpsychism are also preferable to panqualityism, the most promising version of Russellian panprotopsychism, as they, unlike panqualityism, are arguably not threatened by conceivability considerations. As I see it then, constitutive and identity Russellian panpsychism have many advantages when compared to other versions of Russellian monism. In the next and final chapter I shall explore how these two views cope with respect to the so called “combination problem” for panpsychism.

7. The Combination Problem

1. How to Combine Experiences?

Perhaps the most serious challenge for Russellian panpsychism and – as we shall see – especially for constitutive Russellian panpsychism, the views which are argued for in the previous chapter, has proved to be one of explaining how micro-phenomenal properties of the micro-physical entities could constitute or give rise to macro-phenomenal properties or macro-consciousness of the kind we are familiar with. This difficulty is called "the combination problem" by William Seager who characterises it as "the problem of explaining how the myriad elements of 'atomic consciousness' can be combined into a new, complex and rich consciousness such as that we possess".[328] As I see it, the combination problem has its roots in the insight that while we have a sort of conception of how micro-physical entities can constitute macro-physical objects, we arguably utterly lack such a conception when it comes to combining experiences.[329]

Historically speaking, the combination problem is associated with William James who famously discussed it in the course of his critique of the mind-stuff theory which amounts to a form of panpsychism. James wrote, for example, the following:

> *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 may 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 feelings 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.*[330]

The basic thought behind this rich passage is that feelings or experiences are somehow necessarily metaphysically isolated from each other and as such they cannot be combined to create new feelings and experiences. If this is so, then panpsychism (including Russellian panpsychism) ultimately fails to account for the existence of our consciousness because it is not clear why a swarm or cloud of instances of micro-consciousness should constitute a macro-consciousness which, as it seems,

[328]Seager (1995, p. 278).
[329]See e.g. Goff (2006).
[330]James (1890, p. 160).

features a specific sort of unity, stability, etc, and how this swarm of instances of macro-consciousness could even in principle do such a job. Notice that we normally do not have a similar problem in accepting that a swarm or cloud of atoms can constitute a chair or a plant with their unity and stability.

While James focuses, roughly, on the problem of mixing or combining phenomenal properties, the combination problem also has other dimensions. We can see this if we consider that there, plausibly, cannot be an experience without it being experienced by an experiencing subject – the existence of experience, arguably, consists in being experienced and thus requires an experiencer.[331] The panpsychist is thus committed to the existence of primitive micro-subjects which experience the micro-phenomenal properties attributed by panpsychism to fundamental micro-physical entities. According to panpsychism these micro-subjects are supposed to somehow collectively constitute or give rise to a macro-subject. Intuitively, however, an ordinary macro-subject is not a complex system of myriads of micro-subjects. Insofar as we can trust our first-person perspective, an organism is one particular macro-subject and it is far from clear that a macro-subject can somehow arise from / be constituted by the vast number of micro-subjects whose existence panpsychism is committed to. We can call this dimension of the combination problem the *subject combination problem*. I think the subject combination problem is one of the most troubling aspects of the combination problem for panpsychism.[332] In what follows, I shall therefore focus mostly on this problem, although I shall also attempt to address other aspects of the combination problem.

The combination problem applies with most force to what is – at least initially – the most plausible variety of Russellian panpsychism – constitutive Russellian panpsychism. In the previous chapter, I argued that constitutive Russellian panpsychism has significant merits over emergent and autonomous Russellian panpsychism, over non-Russellian versions of panpsychism as well as over panprotopsychism. Why, however, think that constitutive Russellian panpsychism is more plausible than identity Russellian panpsychism, the view that macro-phenomenal properties are identical with some micro-phenomenal properties?[333] Here one could say that it certainly sounds more plausible to hold that the complex conscious macro-experience which we are familiar with is collectively constituted by a great number of micro-experiences than to think that it is identical with some particular micro-experience had by a micro-physical entity. Still, identity Russellian panpsychism deserves at least a brief mention if only because it is a variety of panpsychism which works as the synthesis of the Hegelian argument and at the same time does not face the combination problem.

[331]Strawson (2006, p. 26).
[332]We can see that the subject combination problem does not arise for the panprotopsychists who do not need to posit micro-subjects as protophenomenal properties do not require microsubjects for their existence.
[333]The content of all of these doctrines was clarified in the previous chapter.

One version of identity Russellian panpsychism amounts to the view that macro-experience is identical with micro-experience of a fundamental micro-physical entity, such as a photon or a quark. Chalmers calls this the "dominant monad" view as it resembles the view once introduced by Leibniz in the *Monadology*.[334] While certainly interesting, the dominant monad view faces serious problems. The fundamental entity whose micro-experience is, according to this view, identical with my macro-experience, could, after all, presumably easily leave my body at some point of my existence and it is not clear what would then happen with my consciousness – presumably, it would leave my body and travel throughout the universe. It is, however, extremely implausible to claim that absence of a single fundamental particle could have the far reaching effect that I would stop being conscious. Moreover, it is hard to see how this dominant fundamental micro-physical entity could feature the rich causal profile which my consciousness seems to have.[335] Finally, as Chalmers emphasises, if a single micro-physical entity has, phenomenally speaking, a fairly rich and complex inner life of the kind I seem to have and at the same time is physically just like myriads of other fundamental entities of the same fundamental kind, then, presumably, all the fundamental entities constituting my body should be phenomenally similar to me – there would then be in my body myriads of macro-subjects with complex experiences of the type I am now having – which sounds simply absurd.[336] Allowing, on the other hand, that physically identical fundamental entities could be radically phenomenally different raises the question as to why the "dominant monad" is special in having rich phenomenology which reliably tracks my environment, etc.

Perhaps more promising than the dominant monad view are the holistic varieties of identity Russellian panpsychism according to which macro-experience is identical to experience had by a fundamental entity which is not atomistic. Here one option is the view that ordinary macro-experience is identical with experience had by an entangled system which is (a part of) one's brain.[337] Although interesting and more intuitive than the dominant monad view, the quantum holistic account faces some serious worries and problems. Firstly, it is far from clear that there are entangled systems in the brain which exhibit the kind of stability which our conscious experience seems to exhibit. Even if they did, moreover, it is not clear that they could reliably map our environment in the way our phenomenal states normally seem to. Secondly, physicists are not in agreement as to whether there even is quantum entanglement. As a result, it seems fair to say that the Russellian panpsychist will certainly be better-off without needing to rely on quantum entanglement.

Another option for the identity Russellian panpsychist is to hold that the whole universe is a

[334]Chalmers (2015, p. 270).
[335]Chalmers (2015, ibid.).
[336]See Chalmers (forthcoming).
[337]See e.g. Seager (2010) for a view of this kind.

fundamental physical entity which is realised by a giant phenomenal quiddity, presumably experienced by a giant cosmic subject. This version of identity Russellian panpsychism amounts to a variety of *cosmopsychism*, the view according to which the whole universe is phenomenally conscious.[338] Proponents of cosmopsychism face the challenge of needing to explain how macro-subjects like us arise from the giant cosmic subject or perhaps how they are individuated within this subject. This problem, called the decomposition problem by Chalmers, is in essence a reversed combination problem.[339] It consists in the fact that the existence of a cosmic subject itself does not provide us with any explanation as to why you or I are not zombies or why we have the kinds of phenomenal states we have. According to Chalmers, the decomposition problem seems just as difficult as the combination problem. As I see it, this judgment may perhaps be somewhat premature as philosophers have only just started exploring cosmopsychism and the decomposition problem. Still, the problem seems to be very serious and it will certainly be a virtue of a theory of consciousness if it manages to avoid it.

Given the overall implausibility of the dominant monad view, the fact that the quantum holistic view involves a significant bet on the existence of quantum entanglement of a rather stable kind and the fact that the existence of a cosmic subject does not seem to explain the existence of ordinary subjects, it is, I believe, reasonable to think that constitutive Russellian panpsychism is more plausible than identity Russellian panpsychism. Still, one may object that despite these worries, identity Russellian panpsychism has one important, perhaps decisive advantage over constitutive Russellian panpsychism – it does not need to face the combination problem. Clearly, after all, identity Russellian panpsychism does not require any combination of micro-experiences in order for a macro-experience to arise. What is more, the dominant monad view and the quantum holistic variety of identity Russellian panpsychism do not even need to face the decomposition problem which arises for cosmopsychism.

Notice, however, that the fact that identity Russellian panpsychism does not face the combination problem can only be considered to be an advantage if one already accepts that the combination problem is insoluble. If it turned out that the combination problem can be solved, identity Russellian panpsychism would lose much of its appeal. The much debated question whether the combination problem can be solved will be my main topic in what follows. If it turned out that it can, it would mean that constitutive Russellian panpsychism is the most plausible version of panpsychism as well as the most plausible version or Russellian monism and – if the considerations in the previous chapters are sound – also the most promising account of the existence of consciousness in the physical world. The question whether the combination problem for constitutive

[338]See e.g. Shani (2015).
[339]Chalmers (forthcoming).

Russellian panpsychism is soluble is therefore of crucial importance.

In what follows I shall therefore tackle just this issue and explore the question whether there are any strong reasons to hold that phenomenal combination is impossible and that the combination problem is insoluble.[340] Before considering these, I shall start on a more positive note and suggest that we do in fact possess a notion of how phenomenal properties and perhaps also micro-subjects can combine.

2. The Relation of Co-Consciousness

Consider the state of my consciousness at this very moment. I am having a visual experience of seeing a white laptop screen, a tactile experience of touching the keyboard, an audible experience of cars outside passing my building and perhaps also the cognitive experience of thinking the thoughts which these sentences express. Altogether, these experiences form a sort of unified field, which we can, with Searle, call the conscious field.[341] This conscious field has boundaries which give it a sort of unity. Given these boundaries, there are experiences, sensory and others, which are not part of my conscious field, such as, for example, the experiences had by the driver of the car which is now passing my building.[342] We can say that my experiences then form a sort of unity which is due to the fact that they are all parts of my conscious field, and not yours or somebody else's. It is easy to recall at this point Leibniz's remarks, directed at perception, about multiplicity in unity as we can now see that rather diverse experiences (tactile, visual, perhaps cognitive) are, nevertheless, unified.[343] If we ask what is responsible for this sort of unity, what the mechanism behind the unity is (if any), or even simply what this unity consists in, our armchair reflection proves to be rather unhelpful and we find that we have no clue as to how to answer these questions. It, nevertheless, seems clear to us that there is unity of this kind.

Given the unity of the conscious field, we can observe that the experiences which are parts of this field are related in a specific way in virtue of being parts of the same conscious field and all of them being experienced by a single subject. This relation has been called *co-consciousness* by Barry Dainton.[344] We can say that two experiences or two phenomenal qualities are co-conscious iff they are jointly experienced by a single subject. It has been suggested by Dainton, Chalmers and others that the relation of co-consciousness can help us understand how micro-experiences, micro-

[340]The notion "phenomenal combination" is used here quite generally for combination of micro-subjects, combination of micro-experiences as well as for combination of (instances of) phenomenal properties.
[341]See e.g. Searle (1993).
[342]To deny this would amount to solipsism.
[343]Leibniz (1714/1898, § 14, p. 224).
[344]See e.g. Dainton (2011, p. 251).

phenomenal properties or micro-subjects can combine so as to constitute macro-experiences, macro-phenomenal qualities or macro-subjects.[345] If these suggestions point us in the right direction, the relation of co-consciousness is important for the panpsychists as it has potential to provide them if not with a solution to the combination problem then at least with an idea as to what a solution to the combination problem could look like. It seems worthwhile therefore to explore the relation of co-consciousness in a bit more detail.

One question we should ask is what kind of existents enter the relation of co-consciousness. Here it is perhaps best to understand this relation as a relation between phenomenal properties, or qualia. As the above-presented description of my experience brought to our attention, I am often simultaneously aware of multiple phenomenal properties. That, I take it, is just a way of saying that it is often the case that multiple phenomenal properties are co-conscious, i.e. simultaneously experienced by a single conscious subject.

Perhaps there could be an extreme position according to which one can only be aware of a single phenomenal property at a given moment. It is not clear, however, that there is any reason to hold a view of that kind. Indeed, if one thinks of say, listening to a Rolling Stones song and eating an apple at the same time, it is natural to view this case as one in which many phenomenal properties are co-conscious, collectively constituting a conscious field instantiating a particular phenomenal character. One would, I think, need a strong argument if one wished to reject this intuitive view.

I take it therefore that what I have so far said about co-consciousness is rather uncontroversial. In order, however, for the panpsychists to be able to usefully employ the relation of co-consciousness in their solution to the combination problem, they will need to hold that this relation sometimes holds between phenomenal properties experienced by different subjects. Recall, after all, that the theoretical role which co-consciousness is supposed to play in panpsychism is to explain how phenomenal qualities experienced each by "its own" micro-subject can at times be collectively experienced by a macro-subject. The panpsychists' explanation will likely go along these lines: given that it is the case that whenever two phenomenal properties are co-conscious, there is, by definition of co-consciousness, a subject collectively experiencing them and given that phenomenal properties experienced by distinct micro-subjects are, as panpsychists hold, sometimes co-conscious, there must as a result sometimes be a subject jointly experiencing phenomenal qualities which are, at the same time, individually experienced each by its own micro-subject. We can thus say that the panpsychists' reliance on the notion of co-consciousness commits them to the thesis which we can call *co-consciousness across micro-subjects*, which states that phenomenal properties experienced by distinct micro-subjects can, under the right conditions, be co-conscious.

[345]Dainton (2011), Chalmers (forthcoming).

The thesis of co-consciousness across micro-subjects is, of course, controversial as it implies that the same conscious content, the same phenomenal property can be simultaneously experienced by two different subjects. In particular, it implies that a given phenomenal property experienced by a micro-subject is also – typically as one of many phenomenal properties – experienced by a particular macro-subject. The constitutive panpsychists are then committed to what we can call the *sharing principle,* according to which numerically the same instance of a given phenomenal property can be shared, i.e. simultaneously experienced, by two or more distinct subjects.[346] I take it that this result may sound quite controversial to some. Still, however odd the sharing principle may sound at first, it is not clear why it could not be true. Why could, for example, the phenomenal property of red (or the phenomenal property of merely one phenomenal aspect of red) which I am currently experiencing, not be simultaneously experienced by a micro-subject? As far as I'm concerned, even careful reflection on the phenomenal properties which I experience does not reveal to me that they, or their various aspects, could not be simultaneously experienced by another subject, in this case a micro-subject. It is almost certainly conceivable that they are.

3. The Combination Problem as a Conceivability Argument

With these remarks on the table, we now need to understand why exactly phenomenal combination can be thought to be implausible or perhaps even impossible. Perhaps the most discussed source of worries are the considerations of the kind articulated in the above-quoted passage from James. There the basic line of thought is that instantiations of phenomenal properties do in no way necessitate other phenomenal properties to occur, nor do they explain why they occur. If this consideration is plausible, then the existence of a swarm of micro-physical entities exhibiting micro-phenomenal properties, which, according to Russellian panpsychists constitute my brain, will not necessitate any of my macro-phenomenal properties.

This line of thought has been philosophically explicated by Goff who introduced a sort of conceivability argument against panpsychism.[347] Moreover, arguments which lead us from conceivability to possibility are controversial. Still, many anti-physicalists, and, of course, many panpsychists among them, are led to the rejection of physicalism by conceivability considerations and, as a result, many of the countermoves which are arguably available to the physicalists will not be available to the panpsychists. That, of course, makes the presented conceivability argument against panpsychism all the more worrying. The argument therefore deserves careful attention on the side of the panpsychists.

[346]See Basile (2010, p. 108).
[347]Goff (2009).

Goff's argument, adopted by Chalmers to target specifically constitutive panpsychism has the following structure:

(1) *PP*&~*Q* is conceivable.

(2) If *PP*&~*Q* is conceivable, it is metaphysically possible.

(3) If *PP*&~*Q* is metaphysically possible, constitutive panpsychism is false.

———————

(4) Constitutive panpsychism is false.

Here *PP* is a conjunction of all microphysical truths and all micro-phenomenal truths (i.e. truths about instantiations of micro-phenomenal properties) and *Q* is some macro-phenomenal truth. Clearly, the crucial premise of this argument is premise (1) which tells us that our micro-physical and, at the same time, micro-phenomenal replicas without macro-consciousness, or ordinary consciousness, are conceivable. In support of this premise, one could simply appeal to the fact that nothing about our notion of a micro-phenomenal property suggests that if many micro-phenomenal properties are gathered and appropriately organised, one or multiple macro-phenomenal properties arise. A similar consideration, of course, applies to micro-subjects: nothing about our notion of a micro-subject suggests that if many micro-subjects are gathered and appropriately organised, one or multiple macro-subjects arise. As far as I can see, these considerations are very plausible and so premise (1) looks well justified. Moreover, it will be hard for panpsychists, who often justify their rejection of physicalism by conceivability considerations, to take issue with the other premises of the argument.

Is there a way then for the panpsychist to respond to this argument? One reply, which has been introduced by Goff himself, appeals to the relation of phenomenal bonding.[348] According to Goff, the micro-subjects posited by panpsychists enter, given that specific conditions are fulfilled, the relation of phenomenal bonding which is such that when two or more micro-subjects enter it, the resulting state of affairs necessitates that another, distinct subject comes into existence.

How, however, could the phenomenal bonding relation exist in the physical universe? Nothing physics tells us certainly suggests the existence of such a relation. Here Goff provides us with an interesting proposal. Recall that according to Russellian panpsychism, phenomenal properties are quiddities which realise fundamental micro-physical roles. Analogically we can, according to Goff's proposal, view the relation of phenomenal bonding as a sort of relational quiddity which realises some fundamental micro-physical relation.[349] The thought here is, once again, that – just like in the

[348]Goff (2009, forthcoming 1).
[349]Goff (forthcoming 1, p. 11).

case of micro-physical entities – physics provides us with merely structural knowledge of micro-physical relations, such as causal, nomic or spatio-temporal relations – all we learn from physics is the role of the relation in the given physical theory and – given that the theory is correct – in the world. Such knowledge, however, leaves it open as to what relation realises the given fundamental role and here one possible candidate is the relation of phenomenal bonding.

On this conception then, there is clear space for phenomenal bonding in nature as the relation is viewed as a relational quiddity realising some micro-physical relation. We can call this suggestion the *Russellian conception of phenomenal bonding*. Once again, however, the term "Russellian" is used here to designate that the view relies on the role-property / quiddity distinction rather than to suggest that this conception would actually be acceptable for Russell. Still, apart from the Russellian version, there seems to also be theoretical space for non-Russellian conceptions of phenomenal bonding according to which the bonding relation is not a relational quiddity and is thus instantiated independently of any particular fundamental micro-physical relation.

How could the relation of phenomenal bonding help us block the above-mentioned conceivability argument against constitutive panpsychism? If micro-subjects enter – given the right conditions – the phenomenal bonding relation, then the panpsychists have a good reason to reject premise (1) because micro-phenomenal truths will include truths about the phenomenal bonding relations. Given the relevant truths about the bonding relations in conjunction with other micro-phenomenal truths (and perhaps some structural truths), *PP&~Q* will not be conceivable. The proponent of constitutive Russellian panpsychism will then have a good reason to reject premise (1) as well as the whole conceivability argument against panpsychism.

Given the importance of the phenomenal bonding relation for panpsychism, it would certainly be of great help if we had a clear idea as to what this relation amounts to. The characteristics which I have provided so far, look, regrettably, rather abstract. Is there a way to make the notion more substantial? Goff himself holds a sort of mysterian view of phenomenal bonding which will be of little help here. The notion of mysterianism is, of course, associated with the positions of thinkers such as Colin McGinn or Noam Chomsky who hold, roughly, that the nature of our cognitive apparatus, which they understand as a product of evolution, renders it impossible for us to understand certain aspects of our mental lives. McGinn, for example, argues that minds of the type which we happen to possess are necessarily cognitively closed with respect to the property of the brain which would enable us to account for the fact that the brain gives rise to consciousness.[350]

Goff, who is a proponent of panpsychism, is, importantly, far from being a mysterian with respect to

[350]McGinn (1989, p. 350).

understanding how the brain gives rise to consciousness. His mysterianism is much more limited and specifically targeted than, for example, the mysterianism of McGinn – it concerns only the nature of the phenomenal bonding relation. Goff suggests that the nature of this relation is necessarily beyond our grasp, because we only have cognitive access to phenomenal states of a single consciousness, our own, and as a result, are unable to understand how two different instances of consciousness could bond. Of course, the Russellian panpsychist will insist that we also have, at least in principle, a sort of cognitive access to different instances of consciousness via our senses, insofar as we can observe the relevant parts of other people's brains. This sensory access, however, is necessarily merely structural and does not enable us to cognitively access the given person's consciousness *qua* consciousness. The relation of phenomenal bonding, however, cannot be merely a matter of structure, otherwise truths about micro-subjects in conjunction with structural truths would a priori entail truths about macro-subjects, which is clearly not the case. Given then that the phenomenal bonding relation is not a matter of structure – although it may have a structural aspect – we should not be surprised that our senses provide us with no knowledge of the nature of the relation.

One could worry that Goff's mysterian conception of phenomenal bonding appeals to a mystery right at the heart of the panpsychist proposal. Does this appeal perhaps not call into question the rest of the panpsychist theory? A sceptic could object here that given that we have no idea as to how and why micro-subjects bond, positing micro-subjects does not really help us explain how macro-subjects arise out of appropriately combined micro-subjects.[351] As I see it, however, such an objection is not quite fair with respect to Goff's theory as he provides us with an explanation as to why the nature of phenomenal bonding is necessarily mysterious for us. Panpsychism with phenomenal bonding then, Goff can reply to the sceptic, is the best theory of consciousness we are able to reach given our epistemological limitations, and the fact that the theory leaves us ignorant as to the nature of phenomenal bonding does not discredit or devalue the theory.

Although I think that the mysterian conception is a possible alternative for the proponent of phenomenal bonding, we certainly would be better off if we had a more positive and determinate grasp of phenomenal bonding. Here one suggestion is that perhaps the relation of phenomenal bonding is the above-discussed relation of co-consciousness.[352] Although, after all, our grasp of co-consciousness is imperfect and in no way transparent, the relation is not ordinarily viewed by us as mysterious and we have some sort of positive idea as to what co-consciousness amounts to.

The suggestion that the relation of co-consciousness is the phenomenal bonding relation may sound

[351]Coleman (forthcoming).
[352]Chalmers (forthcoming).

strange at first. We saw, after all, that the phenomenal bonding relation is defined by Goff as a relation between micro-subjects while the co-consciousness relation is, intuitively, a relation between phenomenal properties experienced by these micro-subjects. There seems, therefore, to be an important difference between the two relations. Still, this difference can be overcome if we conceive of phenomenal bonding as a relation such that it holds between two micro-subjects iff the phenomenal properties experienced by them, are co-conscious, i.e. are jointly experienced by a further subject. On such a conception then subjects are phenomenally bonded in virtue of their contents being co-conscious. Notice that such a step does not take us any closer towards mysterianism about phenomenal bonding as long as we insist that the fact that two subjects are phenomenally bonded amounts to nothing but the fact that the contents experienced by these subjects are co-conscious, i.e. if the relation of phenomenal bonding is fully reducible to the relation of co-consciousness.

According to the current proposal, which we can call *phenomenal bonding constitutive Russellian panpsychism* (*phenomenal bonding panpsychism*, for short)*,* physics describes the world in terms of particular micro-physical roles with some of these roles defining fundamental micro-physical entities while others defining fundamental micro-physical relations. These micro-physical roles, the proposal goes, require quiddities for their realisation. While the roles corresponding to micro-physical entities require monadic quiddities, the roles corresponding to micro-physical relation require relational quiddities for their realisation. Here, natural candidates for the monadic quiddities are phenomenal properties while a natural candidate for one sort of relational quiddity is the relation of co-consciousness. I believe that this proposal paints an interesting and noteworthy picture of the universe as it is a picture which integrates phenomenal consciousness into the world as it is described by physics. The proposal, as I see it, provides an account of macro-consciousness which is both non-reductive and naturalistic. It is non-reductive because it views the phenomenal features of consciousness as fundamental, not derivative from other, non-conscious existents, and it is naturalistic because it views consciousness as an integral part of nature which has a role in the causal chains in nature. The fact that phenomenal bonding panpsychism views phenomenal consciousness as irreducible gives us, as I have argued, a reason to prefer the view over mainstream physicalism. Moreover, the fact that phenomenal bonding panpsychism provides phenomenal consciousness with a space in the causal processes in nature, including the process of natural evolution, gives us, I think, a good reason to prefer the view over mainstream versions of idealism, emergentism and substance dualism. Despite these significant merits over the competing views, however, the account brings about some worries and questions, some of which I shall try to tackle in the remainder of this chapter.

4. Phenomenal Bonding: Transitive or Non-Transitive?

One worry inherent to phenomenal bonding panpsychism has recently been raised by Dainton who claims that while it is intuitive to view the relation of co-consciousness as transitive, such a conception ultimately leads to trouble for Russellian panpsychism.[353] We can say that the relation of co-consciousness is transitive if it is such that whenever phenomenal property Q_1 is co-conscious with phenomenal property Q_2 and at the same time Q_2 is co-conscious with phenomenal property Q_3, then Q_1, Q_2 and Q_3 are jointly experienced by a single subject. This sort of transitivity of co-consciousness looks intuitively plausible as it is natural to think that all of the phenomenal properties related by a relation of co-consciousness must be experienced by a single subject, must exist, so to say, for a single subject. If, however, transitivity of co-consciousness is combined with the Russellian claim that the relation of co-consciousness is a relational quiddity which realises some fundamental physical relation, we are pulled in the direction of a single giant, cosmic subject and the above-discussed cosmopsychism. Fundamental physical relations, such as spatio-temporal or causal relations, after all, presumably relate all microphysical entities in the universe with each other. If however, one of these fundamental micro-physical relations is realised by the relation of co-consciousness, we get the result that all phenomenal properties are co-conscious with each other. Given the transitivity of co-consciousness, however, all phenomenal properties in the universe should presumably be experienced by a single cosmic subject. We saw, however, that cosmopsychism faces the serious challenge of the decomposition problem to which we do not have a solution. In view if the decomposition problem, it seems fair to conclude, however, that the transitive conception of co-consciousness does not, on its own, provide us with what we are searching for – an account of how ordinary macro-subjects can result from combination of micro-subjects.

The threat of cosmopsychism can be, according to Dainton, averted if we deny that the relation of co-consciousness is transitive.[354] Such a denial amounts to the claim that if phenomenal property Q_1 is co-conscious with phenomenal property Q_2 and Q_2 is co-conscious with phenomenal property Q_3, then it is not the case that Q_1, Q_2 and Q_3 are jointly experienced by a single subject, instead we end up with two quasi-micro-subjects, one experiencing Q_1 & Q_2, the other experiencing Q_2 & Q_3.[355] While the threat of cosmopsychism is averted, it is far from clear that denying the transitivity of co-consciousness will significantly help the panpsychist here. Indeed, Chalmers has recently argued that if the relation of co-consciousness is understood as non-transitive, then it cannot help us explain

[353]Dainton (2011, p. 256).
[354]Dainton (2011, p. 257).
[355]I call these fragmentary subjects quasi-micro-subjects because I wish to reserve the term "micro-subject" for the subjects which realise fundamental physical entities.

how macro-subjects come into existence.[356] It seems, after all, that a macro-subject will need to consist of a large number of micro-subjects, otherwise we get the problems of the dominant monad view, discussed above, once again. If, however, the relation of co-consciousness always only relates two micro-subjects, we will, arguably, end up with a large number of fragmentary quasi-micro-subjects but no macro-subject.

The Russellian panpsychists working with co-consciousness then, Chalmers argues, face a dilemma: either they get a cosmic subject, or they get myriads of quasi-micro-subjects which are results of merely two micro-subjects being co-conscious.[357] In neither case is the panpsychist able to provide us with an account of how ordinary macro-subjects are generated by combining micro-subjects. This point, moreover, can be generalized to pose a challenge for all attempts at solving the combination problem by means of the phenomenal bonding relation, i.e. even for those, according to which phenomenal-bonding is not co-consciousness. Any panpsychist working with phenomenal bonding will, after all, clearly need to make a decision as to whether the bonding relation is transitive, or not, and neither option will, according to Chalmers, help us understand how macro-subjects could arise.[358]

5. Ways Out of the Dilemma

Is there a way for the Russellian panpsychists to escape this dilemma? At this point the panpsychists could suggest that the co-consciousness relation is sometimes, but not always transitive. Such a suggestion would amount to the view that in some cases when Q_1 is co-conscious with Q_2, and Q_2 is co-conscious with Q_3, there will be a single subject experiencing all three phenomenal properties while in other cases this is not so. If the relation of co-consciousness is in this way two-fold, we may be able to find middle ground between – to use Chalmers's phrase – the Scylla of the cosmic subject, and the Charybdis of the vast number of quasi-micro-subjects. We could, after all, say that while in certain regions of brains the micro-subjects enter the relation of transitive co-consciousness and thus complex macro-subjects are generated, in other regions of the universe (e.g. in chairs, cars, mountains) micro-subjects enter instead the relation of non-transitive co-consciousness and, as a result, these micro-subjects do not form macro-subjects but rather the fragmentary quasi-micro-subjects.

In order, however, for this kind of proposal to be plausible, the panpsychists would need to clarify why the relation of co-consciousness is sometimes transitive and other times non-transitive and it is

[356]Chalmers (forthcoming).
[357]Chalmers (forthcoming).
[358]Chalmers (forthcoming, p. 24).

far from clear that they have resources to do that. At the very least, when we reflect on the relations between the phenomenal properties in our consciousness, we do not seem to discover two distinct kinds of co-consciousness. Without such an explanation, however, the proposal feels purely *ad hoc*.

What's more, the proposal faces some serious challenges. We saw that according to the Russellian conception of phenomenal bonding, co-consciousness is supposed to be a relational quiddity realising a fundamental physical relation. According to the current proposal, however, we are really dealing with two distinct kinds of co-consciousness – while one kind is transitive, the other kind is non-transitive. If these two kinds of co-consciousness differ when it comes to their nature, it is natural to expect that they also realise two different physical relations. It is far from clear, however, that there is a fundamental physical relation which appears only in those contexts in which micro-subjects bond to form macro-subjects, i.e. presumably only in certain brain regions, and does not appear elsewhere in the universe. There is in fact little reason to expect that, when it comes to *fundamental* physical relations, brains are in some interesting way different from chairs, computers or rocks. The interesting differences indeed seem to appear only on the higher levels of organisation, perhaps the level of neurons, perhaps that of microtubules, perhaps some other level yet.

Alternatively, the panpsychists could deny that there are two distinct kinds of co-consciousness, one transitive and one non-transitive, and hold instead that there is a single kind of co-consciousness, which may or may not be transitive, depending on the context in which the related micro-subjects happen to be in. One suggestion along these lines could be that iff the micro-subjects related by the relation of co-consciousness are densely cramped in a given region of space, then the relation is transitive, while if the concentration of the related entities is less dense, the relation is non-transitive. This suggestion amounts to the claim that the transitivity of co-consciousness requires some specific spatial relations between the micro-subjects. Such a suggestion, however, faces at least two kinds of worries. Firstly, it is not at all clear why density and associated spatial closeness of the related micro-subjects should matter as to the issue of transitivity or intransitivity of the relation of co-consciousness.[359] The second problem with this proposal is that it is doubtful that on the fundamental level the regions in which macro-subjects are supposedly generated out of myriads of micro-subjects are any more densely populated by micro-subjects than the areas in which no macro-subjects seem to be generated. There is then, I take it, no easy way for the panpsychist to justify the proposal that the co-consciousness relation is sometimes transitive while other times non-transitive.

[359]One could perhaps claim that spatial closeness of micro-subjects is a condition for the existence of the relation of co-consciousness (ignoring for the moment that the relation is supposed to realise a fundamental physical relation) but such a claim is significantly different from the current proposal.

Perhaps a more promising way in which the phenomenal bonding theorist may attempt to avoid both horns of Chalmers's dilemma is to suggest that the relation of co-consciousness is transitive but realises certain non-fundamental causal relations instead of fundamental micro-physical relations (causal, spatio-temporal etc.). Here the panpsychist can appeal to the above observed fact that it is, presumably, not the fundamental level where the brains are interestingly different from ordinary objects, say chairs, but rather one or more of the higher, non-fundamental levels, and suggest that the phenomenal bonding relation realises certain causal relations on one of those levels. One could appeal here, for example, to Giulio Tononi's integrated information theory (IIT) according to which certain non-fundamental, complex causal systems, perhaps not limited to human and animal brains, exhibit a high degree of the Φ (or *phi*) property which amounts to the level of information integration, i.e. roughly, causal interconnectedness, in a given system. The systems which exhibit a high level of Φ are, according to IIT the *loci* of higher levels of consciousness.[360] The phenomenal bonding panpsychist could suggest that the phenomenal bonding relation realises precisely the non-fundamental causal relations in a given system which are relevant from the point of view of information integration. Still, the appeal to IIT is not essential here as one could alternatively hold that phenomenal bonding realises some other, presumably biological, non-fundamental relation which only appears in conscious brains.

Proponents of the view that co-consciousness realises some such derivative, non-fundamental causal relations, will presumably need to address the question what quiddities are those which realise the fundamental causal relations grounding this and other derivative relations. Here one reply would be that fundamental causal relations are realised by the relation of *proto-co-consciouness*, a relation which is not itself co-consciousness, and thus does not bond micro-subjects, but which, when it is appropriately structured, such as, for example, in a system with a high level of Φ, gives rise to the relation of co-consciousness which is responsible for the fact that a macro-subject arises in these systems.

One question this kind of suggestion will lead to, is what the relation of proto-co-consciousness is supposed to be. While we arguably have a reasonably positive and determinate conception of co-consciousness, we seem to be quite at loss when it comes to the alleged relation of proto-co-consciousness. This relation should be viewed as a relation which is distinct from the relation of co-consciousness, but which grounds, in conjunction with structural facts about the given system, the relation of co-consciousness. Still, despite having this abstract, relational conception of co-consciousness, the panpsychists will need to admit ignorance as to the nature of this relation (with one possible exception which shall be mentioned later). This ignorance, however, should not be too

[360] See e.g. Tononi – Koch (2015).

surprising given that, as mentioned above, we do not even have a complete understanding of co-consciousness. While our conception of co-consciousness, as I see it, allows for the possibility that co-consciousness is not a fundamental relation and that it is somehow constituted by instances of the more fundamental relations of proto-co-consciousness, it does not provide us with any idea as to what the fundamental constituents of co-consciousness could be. However this may be, the mere fact that we are ignorant of the nature of proto-co-consciousness does not speak against the possibility of such a relation. Our ignorance concerning some relational (and perhaps even some monadic) quiddities is, after all, an integral part of phenomenal bonding Russellian panpsychism.

Still, it may sound suspicious that the panpsychists are helping themselves to an utterly unknown relation and it is important to be clear about what exactly the appeal to proto-co-consciousness means. Up until this point, the panpsychists attributed some properties and relations familiar from our mental lives to fundamental microphysical entities. While these attributions are controversial – although I hope to have repelled at least some of the doubt concerning them –, at least there the panpsychist was working with properties and relations which we are familiar or acquainted with. Can the panpsychist justify positing the relation of proto-co-consciousness whose nature is utterly unknown? I think the only sort of justification available to the panpsychist at this point is the explanatory power of the new posit. If the posit helps us explain the explanandum, its postulation can be seen as reasonably justified. Let me therefore consider now what benefits this posit brings to the constitutive Russellian panpsychist account.

The view according to which the relation of co-consciousness realises certain causal relations of a non-fundamental kind has the advantage that these specific non-fundamental causal relations will not hold among all entities in the universe. If one, for example, thinks about causal chains in a clock or a computer, it is easy to notice that the casual relations which have the most interest from the point of view of the question how the given mechanism works, will not be fundamental, nor will they be "ubiquitous" in the sense that they would hold between all components of the mechanism. Instead they hold only between specific components, connecting them in non-fundamental causal chains. As a result, the proponents of the view according to which the co-consciousness relation realises a specific kind of non-fundamental causal relations, are free to embrace the intuitive thesis that the relation of co-consciousness is transitive. Given, after all, that these causal relations hold only within a highly specific group of complex systems which include (perhaps exclusively) brains of higher animals, embracing transitivity will not lead to cosmopsychism. As we saw, after all, the cosmopsychist worry presupposed that co-consciousness realises a fundamental physical relation relating everything with everything (such as a spatio-temporal or causal relation). At the same time, transitivity of co-consciousness will explain why micro-subjects sometimes combine to form

macro-subjects. The view which I am proposing will then offer us a way out of the dilemma formulated by Chalmers as it will embrace the transitivity option without collapsing into cosmopsychism.

According to this view then, fundamental causality grounds, among other things, the derivative, higher-level causality within certain highly complex systems, and similarly proto-co-consciousness grounds co-consciousness. Co-consciousness then, according to this proposal, occurs if and only if there is a particular structure of the instances of the proto-co-consciousness relations. We can say then that facts about the nature of the proto-co-consciouness relation in conjunction with some structural facts concerning the instances of the proto-co-consciousness relation together metaphysically necessitate instantiations of co-consciousness. What kind of structure or organisation, however, is required in order for co-consciousness to arise out of many instances of proto-co-consciousness? Here the Russellian panpsychist can reply that it is the same sort of structure as the structure which instances of fundamental causality need to exhibit in order to necessitate the relevant higher-level causal relations within a specific kind of complex systems. This reply is, of course, unsurprising, given that proto-co-consciousness is the relational quiddity realising the fundamental causal relations.

One objection which could be raised against the recommended view tells us that derivative causality is just as ever-present as, presumably, fundamental causality so the cosmopsychism worry has not really been dispelled. Here the Russellian panpsychists should reply that the co-consciousness relation does not realise just any old derivative causal relation but merely certain causal relations in systems of a specific kind. It is at this point where the (non-mandatory) appeal to IIT can be made. If the panpsychists do appeal to IIT, they can say that it is only specific causal processes which are realised by the co-consciousness relation – namely, the causal processes which appear in systems exhibiting a high level of Φ and do not appear in systems exhibiting lower levels of Φ. The panpsychist can thus claim that it is only in systems exhibiting this specific kind of complex organisation that the instances of proto-co-consciousness combine so as to give rise to the relation of co-consciousness.

The organisational structure of the instances of the proto-co-consciousness relation then plays a crucial role in this approach – unless there is a particular kind of organisational structure, co-consciousness does not arise. This leads to the natural question of how could a mere structural or organisational difference play such a significant role? Here the panpsychists will, once again, need to admit ignorance. Perhaps if we had a full understanding of the nature of proto-co-consciousness, we could understand why under certain structurally defined conditions instances of the relation give rise to co-consciousness, but, regrettably, our grasp of proto-co-consciousness is too poor to allow

us to understand this issue. We are therefore left with finding out empirically when, i.e. under what physical, structurally defined, conditions, micro-subjects bond to form macro-subjects. The task of identifying these conditions should then be delegated by the Russellian panpsychist to the cognitive scientists who pursue the programme of finding the neural correlates of consciousness.

Another objection against this kind of proposal has been raised by Chalmers who suggests that it faces a new combination problem.[361] The problem, as Chalmers sees it, is that in order for the proposal to work, the instances of phenomenal proto-bonding will need to somehow combine to give rise to phenomenal bonding. Applied to the way I have phrased things, we can say that the problem raised by Chalmers amounts to the question of how instances of proto-co-consciousness could combine to yield co-consciousness.

Here I think proponents of panpsychism who appeal to the co-consciousness conception of phenomenal bonding can question Chalmers's claim that they face a new combination problem. The reason why we talk about the combination problem, after all, seems to consist in the fact that we believe that something we know about the entities which are to combine, prevents them from combining or at least prevents them from combining in the right sort of way. In this way, many, for example, believe that given what we know about phenomenal properties or micro-subjects, we have a good reason to deny that they could combine in the right sort of way.

Notice, however, that this worry does not seem to apply to the relation of proto-co-consciousness which we know very little about. It is, after all, a theoretical posit about which we suppose that (a) it realises fundamental causal relations, and that (b) its instances – when appropriately organised – constitute the relation of co-consciousness (which realises a particular kind of non-fundamental causal relations). Clearly neither (a) nor (b) give us any reason to think that the right kind of combination of instances of proto-co-consciousness could not generate co-consciousness. As we saw above, some thinkers conceive of the combination problem as a sort of conceivability argument.[362] Here the panpsychist could simply insist that truths about instances of proto-co-consciousness together with some structural truths will, given the definition of proto-co-consciousness, a priori entail truths about co-consciousness.

If these, admittedly speculative, considerations are sound, it seems that the notion of proto-co-consciousness turns out to be useful when it comes to attempting to solve the combination problem, or at least its part concerning the combination of micro-subjects. It must be admitted that the solution appeals to an unknown and somewhat obscure relation which renders the proposed account not quite satisfactory. A fully satisfactory treatment would need to clarify the nature of proto-co-

[361]Chalmers (forthcoming, p. 25)
[362]Goff (2009, pp. 296–297).

consciousness. I take it that such a clarification would enable us to understand why, given particular structurally defined conditions, instances of proto-co-consciousness ground instances of co-consciousness. Unfortunately, it is not clear that we will ever be able to achieve such a full understanding of proto-co-consciousness. As we saw, after all, our armchair reflection on the relation of co-consciousness, while it does not rule out that this relation is non-fundamental, gives us little clue as to what it is grounded in. Physical research will, as I see it, be of equally little help here as it will, for reasons discussed above, presumably reveal to us at most the micro-physical roles which the proto-co-consciousness relation is supposed to play without telling us much about the relational quiddity – proto-co-consciousness itself.

6. Other Combination Problems

The subject combination problem is not the only dimension of the combination problem for constitutive Russellian panpsychism, although it may easily be its most pressing dimension. Supposing that I have sufficiently addressed subject combination, it is now time to turn to the other dimensions. According to Chalmers, constitutive Russellian panpsychists also need to face the problem of quality combination and the problem of structure combination.[363] The quality combination problem, also called the "palette problem", is, according to Chalmers rather serious especially for the constitutive Russellian panpsychists who appeal to co-consciousness.[364] Here the thought is that given that there is a highly limited number of the fundamental, primitive phenomenal qualities, it is not clear how we could get the vast and diverse array of phenomenal qualities which we seem to experience given that the phenomenal bonding relation amounts to co-consciousness. This relation certainly does not seem to give us an option of "fusing" a few simple qualities into a wide and varied array of qualities. But, presumably, the approach offers us no other resource.

The thought that there is only a limited number of primitive phenomenal qualities has its roots in the Russellian feature of the proposed view. Given that there is presumably only a narrowly limited number of fundamental micro-physical roles, such as mass, spin, charge, etc., we can be led to thinking that there must only be a limited number of fundamental phenomenal properties, which are the quiddities realising these micro-physical roles. Here the thought is that phenomenally different quiddities also need to differ physically and will therefore need to play different physical roles.

The Russellian panpsychist could, of course, reject this principle and suggest in reply to the quality combination problem that, surely, various phenomenal qualities can realise the same micro-physical role. Perhaps, for example, the fundamental mass-role can be realised by any shade of phenomenal

[363]See Chalmers (forthcoming).
[364]Chalmers (2015, p. 274).

red, the fundamental charge role can be realised by a particular sort of sound qualia, for example a deep tone of any timbre, etc. This suggestion, however, goes against the plausible view that sometimes even small qualitative differences can have significantly different causal effects. The rejection of the principle, according to which phenomenal difference implies physical difference, would mean that many different fundamental phenomenal properties could in fact have the same causal profile, for example, the profile had by fundamental mass. Given that macro-causality is presumably grounded in micro-causality, we could then have phenomenally quite different states which would, nevertheless have the same causal effects. This consequence would, I believe, eradicate many of the benefits which Russellian panpsychism seems to have when it comes to being able to secure a place for consciousness in the causal processes happening in the physical world which I described in the previous chapter.

Is there another way for the Russellian panpsychists to tackle the quality-combination problem? Firstly, they could question the assumption that they need to work only with a highly limited number of primitive phenomenal properties. This assumption can be questioned with appeal to the fact that we do not know how many micro-physical roles the complete micro-physical theory will postulate. If, for example, the final theory is expressed in terms of energy, one could, inspired by the views of Lee Smolin, view different qualities as realising different measures of energy.[365] Such a view would, indeed, neatly complement the insights of Galen Strawson who identifies experience with energy.[366]

It could be objected against this kind of reply that it places too large a bet on one particular way the micro-physical world could turn out to be. Given that we cannot be sure as to how the micro-physical world will turn out to be, it is clearly a good strategy to suppose that it will turn out to be in the way which places most constraints on the metaphysical theory one proposes, and then attempt to accommodate the theory to that possibility. In other words, if the panpsychist is trying to to fit consciousness into the physical world, it is a good idea to work with those physical proposals which place more, rather than fewer, constraints on the way the physical world could turn out to be – in that way one can be later spared unpleasant surprises. Supposing then that completed physics will reveal that there are merely a few fundamental micro-physical properties which – as Russellian panpsychism tells us – are realised by merely a few phenomenal properties, we should ask how the Russellian panpsychists could explain the large number of macro-phenomenal properties which we experience.

One possibility here is to say that the relation of proto-co-consciousness somehow mixes or fuses

[365]See e.g. Smolin (2015, p. 99).
[366]Strawson (forthcoming).

the limited number of primitive phenomenal properties into the much larger number of less primitive phenomenal properties which we are familiar with from our experience. While it would seem implausible that the relation of co-consciousness, which we are acquainted with, could do this job of phenomenal mixing, there is no reason to think that the relation of proto-co-consciousness could not do this job. Indeed, it would seem that if the rich array of phenomenal properties we are acquainted with is somehow derived from a highly limited repertoire of primitive micro-phenomenal properties, the phenomenal mixing must somehow happen in virtue of some relation between the primitives. Here explanatory economy leads us to the speculative thought that perhaps it happens in virtue of the proto-co-consciousness relation.

How, however, could a single relation between a few primitive micro-phenomenal properties produce the experienced richness of macro-phenomenal properties? Here the thought is that perhaps the relation of proto-co-consciousness comes in a large number of different intensities or levels, instead of being a matter of on / off. While the relation, as I conceive of it here, realises fundamental causality, fundamental causal relations presumably occur between entities in various spatio-temporal relations and it is natural to think that the spatial distance between two micro-subjects could correlate with intensity of proto-co-consciousness relating these micro-subjects. If proto-co-consciousness comes in many different intensities or levels, it is to be expected that there will be many different ways in which the primitive phenomenal properties can be mixed or fused into more complex ones. Here one can speculate that if, for example, phenomenal properties A and B are related by a higher intensity of co-consciousness than the pairs of phenomenal properties B and C, and A and C, the resulting complex phenomenal quality will have a lot of A and B phenomenal elements with little or none of the C element added into the mixture. In this way a small number of primitive phenomenal elements can, presumably be mixed into a large number of phenomenal properties. Of course, fuller understanding of this sort of phenomenal mixing is prevented by a lack of understanding of what exactly the primitive phenomenal elements – the micro-phenomenal properties – are. While we have theoretical reasons to think that they are phenomenal, we know little else about them.

Such a suggestion can clearly be viewed as one which puts way too much emphasis on the relation of proto-co-consciousness. This special relation would now be viewed as not only grounding the familiar relation of co-consciousness but also as somehow mixing the limited number of primitive micro-phenomenal properties into a much larger number of further, derivative phenomenal properties.

It may certainly sound somewhat odd that there be a fundamental relation which performs these two different functions. Surely, one would expect that the relation of proto-co-consciousness which is

supposedly responsible for the phenomenal mixing of the items it relates, the micro-phenomenal properties, into a single complex macro-phenomenal property could not ground (together with structural facts) a relation of co-consciousness which precisely does not mix its *relata* in this way. Intuitively, after all, when I watch the computer screen and listen to a Rolling Stones song at the same time, there is a clear sense in which I have two distinct experiences, if perhaps influenced by one another, and not some sort of fused or mixed experience. Still, the Russellian panpsychist could simply insist here that this is because in the one case we are dealing with the relation of proto-co-consciousness while in the other case we are dealing with the familiar relation of co-consciousness. These, however, are two distinct relations, although the former grounds the latter.

Finally, let me briefly address the structure combination problem which can also be seen as pressing with respect to Russellian panpsychism. This problem arises from the fact that, according to constitutive Russellian panpsychism, micro-phenomenal properties, being quiddities, realise the micro-physical structure and it is not clear how this structure can constitute the macro-phenomenal structure of our consciousness. Here I think the proponents of the view recommended here can appeal to the fact that the crucial structure for them will presumably be the non-fundamental structure of the non-fundamental causal relation, which is realised by the co-consciousness relation. They will thus need to insist that there is a level in the brain which is structurally isomorphic with the macro-phenomenal structure one is aware of. At the same time, the panpsychists can presumably allow that experience has more structure than one is aware of and will be able to hold that much of the micro-physical structure is thus not part of our conscious field. The combination of these two lines of thought can presumably point the panpsychists in the right direction as to replying to the structure combination problem.

7. Conclusion

What then is the final view I am recommending here? It is a variety of Russellian panpsychism which views the co-consciousness relation as a key to phenomenal bonding and formation of macro-subjects. The proposal views the co-consciousness relation as a non-fundamental relational quiddity which realises certain non-fundamental, i.e. derivative causal relations occurring in human and higher-animal brains (and perhaps in some other highly causally complex systems as well). Being non-fundamental, the co-consciousness relation must be grounded in some more fundamental relation which, I have suggested, we can call proto-co-consciousness. Apart from grounding the co-consciousness relation, the proto-co-consciousness relation also realises some fundamental relations holding between fundamental microphysical entities. I have, moreover, attempted to show the kind

of reply which the proponent of proto-co-consciousness can offer to the quality combination and the structure combination problems.

It must, of course, be admitted that the account of consciousness I have attempted to offer here fails to be fully perspicuous or transparent. I think the first source of opacity is the notion of co-consciousness. There we saw that while we have a sort of intuitive grasp of what co-consciousness is, there is a real sense in which we do not really understand this relation. Still, given that the relation offers us a model of integrating various phenomenal properties – and probably the only model of integrating phenomenal properties we have –, I take it that the appeal to co-consciousness is justified. Of course, more opacity came into my account with the notion of proto-co-consciousness – there we deal with a purely speculative posit with few constraints as to what it is or what job it can do in the presented theory. Still, I take it that this posit is justified by the need to find a fundamental relation which would ground the derivative relation of co-consciousness. It is clear that much of the proposed solution to the combination problem is speculative but perhaps when it comes to truly hard problems, speculation is the best we can do.

Abstract:

This thesis attempts to provide a philosophical answer to the question of how phenomenal consciousness, or experience, can exist in the physical world, i.e. in the world as it is described by science. The thesis has three parts: In the first part (chapter 1) I explicate the concept of phenomenal consciousness and contrast it with other concepts of consciousness common in the literature. Moreover, I suggest that the project pursued in this thesis can be naturally viewed as a part of the more general project of trying to find a stereoscopic view of man, taken by Wilfrid Sellars to be a crucial task for contemporary philosophy.

In the second part of the thesis (chapters 2 to 4) I offer a detailed evaluation of the attempts at a materialist reduction of consciousness. While in chapter 2 I explore and critique the approach of a priori physicalism (Dennett, Lewis, Rey, etc.), in chapters 3 and 4, I focus on the more recent doctrine of a posteriori physicalism and especially its most prominent variety called the *phenomenal concept strategy* (Loar, Papineau, Levin, Schroer, etc.). One problem with a posteriori physicalism is that, as Nida-Rümelin, Goff and others argue, the view cannot make sense of the plausible thesis that our phenomenal concepts a priori provide us with rich knowledge of the nature of their referents. I offer a new version of this type of argument, suggesting that the *non-structural translucency claim,* according to which our phenomenal concepts a priori reveal to us more than merely structural knowledge of their referents, is incompatible with a posteriori physicalism.

In the third part of the thesis (chapters 5 to 7), I explore the prospects of the main non-reductive approaches to consciousness. First, in chapter 5 I focus on strong emergentism, the view that consciousness is an ontologically new property which arises in certain highly complex physical systems. While I initially defend the view against arguments of Nagel and Strawson, I conclude that strong emergentists cannot make sense of consciousness being causally efficacious with respect to the physical world, and, as I argue, even of consciousness being caused or determined by purely physical underlying processes. In chapter 6, I argue for constitutive Russellian panpsychism which is a version of Russellian monism, a non-reductive view of consciousness inspired by Russell's philosophy of science. According to Russellian panpsychism, phenomenal properties are the quiddities that realise the micro-physical roles which physics informs us about. In the course of defending constitutive Russellian panpsychism, I offer new arguments against the main competing Russellian views: emergent Russellian panpsychism and panqualityism. Perhaps the most serious problem for constitutive Russellian panpsychism is the combination problem, i.e. the question, expressed e.g. by William James, as to whether and how micro-phenomenal properties can combine to collectively produce macro-phenomenal properties. In the final chapter I try to show how we can start solving this problem by appealing to the intuitive notion of co-consciousness.

Key words:

consciousness, experience, phenomenal character, qualia, phenomenal concepts, physicalism, panpsychism, Russellian monism, panqualityism, emergence, strong necessity, conceivability argument, reductionism

Abstrakt

Tato práce má za cíl poskytnout filosofickou odpověď na otázku, jak může fenomenální vědomí neboli prožívání (*experience*) existovat ve fyzickém světě, tj. ve světě, který nám popisují přírodní vědy. Práce sestává ze tří částí: v první části (kap. 1) pojednávám o pojmu fenomenálního vědomí a uvádím jej do protikladu k jiným pojmům vědomí běžným v literatuře. Dále ukazuji, že projekt, o který mi jde v této práci, je přirozené chápat jako součást obecnějšího projektu hledání stereoskopického pohledu na člověka, který Wilfrid Sellars považoval za klíčový úkol pro filosofii své doby.

V druhé části práce (kap. 2 až 4) předkládám detailní zhodnocení pokusů o materialistickou redukci vědomí. Zatímco v kap. 2 zkoumám a podrobuji kritice apriorní fyzikalismus (Dennett, Lewis, Rey atd.), v kap. 3 a 4 se zaměřuji na novější, aposteriorní fyzikalismus a především na jeho vlivnou variantu zvanou *strategie fenomenálních pojmů* (Loar, Papineau, Levinová, Schroer atd.). Jedna obtíž, které aposteriorní fyzikalismus čelí, spočívá v tom, že, jak se pokoušejí ukázat Nida-Rümelinová, Goff a další, toto stanovisko není slučitelné s přesvědčivou tezí, že nám naše fenomenální pojmy poskytují bohaté poznání povahy svých referentů. Předkládám novou verzi argumentu tohoto typu, která vychází z myšlenky, že *tvrzení ne-strukturní průsvitnosti,* podle něhož nám naše fenomenální pojmy vyjevují více než jen strukturní poznatky o svých referentech, je neslučitelné s aposteriorním fyzikalismem.

Ve třetí části této práce (kap. 5 až 7), zkoumám hlavní nereduktivní přístupy k vědomí. Nejdříve se v 5. kapitole zaměřuji na silný emergentismus, stanovisko, podle něhož je vědomí ontologicky nová vlastnost, která vzniká v některých vysoce komplexních fyzických systémech. Toto stanovisko sice nejdříve obhajuji proti argumentům Nagela a Strawsona, dospívám však k závěru, že silní emergentisté nedokážou přesvědčivě vysvětlit, jak může vědomí mít kauzální účinky ve fyzickém světě, a – jak se pokouším ukázat – ani, jak může vědomí být způsobováno či determinováno čistě fyzickými procesy. V kapitole 6 argumentuji pro konstitutivní russellovský panpsychismus, který představuje variantu russellovského monismu, nereduktivního přístupu k vědomí inspirovaného Russellovou filosofií vědy. Podle russellovského panpsychismu máme fenomenální vlastnosti chápat jako quidity, které realizují mikrofyzické role, o nichž nás informuje fyzika. V rámci obhajoby konstitutivního russellovského panpsychismu nabízím nové argumenty proti jeho hlavním konkurenčním stanoviskům: emergentnímu russellovskému panpsychismu a pankvalitismu. Zřejmě nejzávažnější problém pro konstitutivní russellovský panpsychismus představuje problém kombinace, tj. otázka, vyjádřená již Williamem Jamesem, zda a jak se mohou mikrofenomenální vlastnosti kombinovat a tím společně utvářet makrofenomenální vlastnosti. V závěrečné kapitole se pokouším ukázat, jak tento tento problém začít řešit poukazem na intuitivní pojem spoluvědomí.

Klíčová slova:

vědomí, prožívání fenomenální povaha, qualia, fenomenální pojmy, fyzikalismus, panpsychismus, russellovský monismus, pankvalitismus, emergence, silná nutnost, argument z myslitelnosti, redukcionismus

Bibliography

Alexander, Samuel (1920) *Space, Time, and Deity. The Gifford Lectures in Glasgow 1916-1918*. *Vol. II.* New York: The Humanities Press.

Alter, T. – Nagasawa, Y. (eds.) (2015) *Consciousness in the Physical World. Perspectives on Russellian Monism.* Oxford University Press.

Alter, T. – Walter, S. (eds.) (2007) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* Oxford University Press.

Armstrong, D. (1968) *A Materialist Theory of Mind.* London: Routledge and Kegan Paul.

Armstrong, D. (1983) *What Is a Law of Nature?* Cambridge University Press.

Barnes, E. (2012) "Emergence and Fundamentality". *Mind* 121, no. 484, pp. 873–901.

Basile, P. (2010) "It Must be True – But How Can it Be? Some Remarks on Panpsychism and Mental Composition". *Royal Institute of Philosophy Supplement* 67, pp. 93–112.

Bedau, M. A. (1997) "Weak Emergence", in: J. Tomberlin (ed.), *Philosophical Perspectives: Mind, Causation, and the World*, *Vol. 11*. Malden: Blackwell, pp. 375-399.

Bennett, J. (1984) *A Study Of Spinoza's Ethics.* Indianapolis: Hackett.

Berkeley, G. (1721/1965) "De Motu", in: D. A. Armstrong (ed.), *Berkeley's Philosophical Writings,* Macmillan, New York.

Bird, A. (2005) "The Dispositionalist Conception of Laws". *Foundations of Science* 10, pp. 353–370.

Blamauer, M. (2011) *The Mental As Fundamental. New Perspectives on Panpsychism.* Heusenstamm: Ontos Verlag.

Block, N. (1994) "Qualia", in: S. Guttenplan (ed.), *A Companion to the Philosophy of Mind.* Oxford: Blackwell. URL: <http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Qualia_from_Guttenplan.pdf>

Block, N. (1995) "On a Confusion about a Function of Consciousness", *Behavioural and Brain Sciences* 18, pp. 227-287.

Block, N. (2004) "Qualia", in: R. L. Gregory (ed.) *The Oxford Companion to the Mind.* Oxford University Press.

Block, N. (2007) "Wittgenstein and Qualia", *Philosophical Perspectives* 21, pp. 73-115.

Block, N. – Stalnaker, R. (1999) "Conceptual Analysis, Dualism and the Explanatory Gap", *Philosophical Review* 108, pp. 1–46.

Bourget, D. – Chalmers, D. (2014) “What Do Philosophers Believe?”, *Philosophical Studies* 170, pp. 465–500.

Braddon-Mitchell, D. – Jackson, F. (2007) *The Philosophy of Mind and Cognition. An Introduction*. Malden: Blackwell.

Broad, C. D. (1925) *The Mind and its Place in Nature.* New York: Harcourt, Brace & Company.

Caston, V. (1997) “Epiphenomenals, Ancient and Modern”. *Philosophical Review* 106, pp. 309–363.

Chalmers, D. J. (1995) “Facing Up to the Problem of Consciousness”, *Journal of Consciousness Studies* 2, no. 3, pp. 200-19.

Chalmers, D. J. (1996) *The Conscious Mind. In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. J. (1999) “Materialism and the Metaphysics of Modality”. *Philosophy and Phenomenological Research* 59, no. 2, pp. 473–496.

Chalmers, D. J. (2002) “Does Conceivability Entail Possibility?”, in: T. Gendler – J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford University Press.

Chalmers, D. J. (ed.) (2002) *Philosophy of Mind. Classical and Contemporary Readings*, New York: Oxford University Press, pp. 329–334. Cambridge: Harvard University Press.

Chalmers, D. J. (2003) “The Content and Epistemology of Phenomenal Belief”, in: Q. Smith – A. Jokic (eds.), *Consciousness: New Philosophical Perspectives*, Oxford University Press.

Chalmers, D. J. (2006) “Strong and Weak Emergence”, in: Clayton, P. – Davies, P. (eds.), *The Re-Emergence of Emergence. The Emergentist Hypothesis from Science to Religion.* Oxford University Press, pp. 244–254.

Chalmers, D. J. (2006b) “The Foundations of Two-Dimensional Semantics”, in: M. Garcia-Carpintero – J. Macia (eds.), *Two-Dimensional Semantics.* Oxford University Press, pp. 55–140.

Chalmers, D. J. (2007) “Phenomenal Concepts and the Explanatory Gap”, in: Alter –Walter (eds.), pp. 167–194.

Chalmers, D. J. (2010) *The Character of Consciousness.* Oxford: Oxford University Press.

Chalmers, D. J. (2012) *Constructing the World.* Oxford University Press.

Chalmers, D. J. (2015) “Panpsychism and Panprotopsychism” In: Alter – Nagasawa (2015), pp. 246–276.

Chalmers, D. J. (forthcoming) “The Combination Problem for Panpsychism”, in: L. Jaskolla – G. Brüntrup (eds.), *Panpsychism.* Oxford University Press.

Chalmers, D. J. - Jackson, F. (2010) “Conceptual Analysis and Reductive Explanation”, in:

Chalmers, pp. 207–247.

Churchland, P. (1997) “The Hornswoggle Problem”, in: J. Shear (ed.), *Explaining Consciousness. The Hard Problem,* Cambridge: MIT Press, pp. 37–44.

Coleman, S. (2009) “Mind under Matter”, in: D. Skrbina (ed.) *Mind that Abides. Panpsychism in the New Millennium.* Amsterdam: John Benjamins, pp. 83–107.

Coleman, S. (forthcoming) “Panpsychism and Neutral Monism: How to Make Up One's Mind”, in: L. Jaskolla – G. Brüntrup (eds.), *Panpsychism.* Oxford University Press.

Dainton, B. (2011) “Review of Consciousness and Its Place in Nature”. *Philosophy and Phenomenological Research* 83, no. 1, pp. 238–261.

Damnjanovic, Nic. (2012) “Revelation and Physicalism” *dialectica* 66, no. 1, pp. 69–91.

Dennett, D. C. (1991) *Consciousness Explained.* New York: Back Bay Books.

Dennett, D. C. (1995) “The unimagined preposterousness of zombies”. *Journal of Consciousness Studies* 2, no. 4, pp. 322-326.

Dennett, D. C. (2002) “Quining Qualia”, in: D. J. Chalmers (ed.), *Philosophy of Mind. Classical and Contemporary Readings* New York: Oxford University Press, pp. 226–246.

Dennett, D. C. (2005) *Sweet Dreams. Philosophical Obstacles to a Science of Consciousness.* Cambridge: MIT Press.

Descartes, R. (1641/1996) *Meditations on First Philosophy*. Trans. and ed. J. Cottingham. New York: Cambridge University Press.

deVries, W. (2015) "Wilfrid Sellars", in: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition).* URL = <http://plato.stanford.edu/archives/fall2015/entries/sellars/>.

Diaz-Leon, E. (2013) “Do A Posteriori Physicalists Get our Phenomenal Concepts Wrong?”. *Ratio (new series)* 27, pp. 1–16.

Feigl, H. (1971) “Some Crucial Issues of Mind-Body Monism”. *Synthese* 22, no. 3, pp. 295–312.

Fine, K. (1995) “Essence and Modality”. *Philosophical Perspectives* 8, pp. 1–16.

Goff, P. (2006) “Experiences Don't Sum”, in: Strawson, G. et al. *Consciousness and Its Place in Nature. Does Physicalism Entail Panpsychism?* Exeter: Imprint Academic, pp. 53–61.

Goff, P. (2009) “Why Panpsychism Doesn't Help Us Explain Consciousness”. *dialectica* 63, no. 3, pp. 289–311.

Goff, P. (2010) “Ghosts and Sparse Properties: Why Physicalists Have More to Fear From Ghosts than Zombies”. *Philosophy and Phenomenological Research* 81, no. 1, pp. 119–139.

Goff, P. (2011) “A Posteriori Physicalists Get Our Phenomenal Concepts Wrong”, *Australasian Journal of Philosophy* 89, no. 2, pp. 191–209.

Goff, P. (2015a) “Against Constitutive Russellian Monism”, in: Alter – Nagasawa, pp. 370–

400.

Goff, P. (2015b) “Real Acquaintance and Physicalism”, in: P. Coates – S. Coleman (eds.), *Phenomenal Qualities. Sense, Perception, and Consciousness.* Oxford: Oxford University Press, pp. 121–141.

Goff, P. (forthcoming 1) “The Phenomenal Bonding Solution to the Combination Problem”, in: L. Jaskolla – G. Brüntrup (eds.), *Panpsychism*, Oxford University Press.

Goff, P. (forthcoming 2) *Consciousness and Fundamental Reality.* URL*:* <http://www.philipgoffphilosophy.com/publications.html>

Goff, P. – Papineau, D. (2014) „What's Wrong with Strong Necessities?“. *Philosophical Studies* 167, no. 3, pp. 749–62.

Hill, J. (2012) “How Hume Became the 'New Hume': A Developmental Approach”. *The Journal of Scottish Philosophy* 10, no. 2, pp. 163–181.

Hříbek, T. (2016) “Mají zvířata vědomí?”. *Filosofický časopis* 64, no. 1, pp. 3–22.

Jackson, F. (1982) “Epiphenomenal Qualia”. *Philosophical Quarterly* 32, pp. 127–136.

Jackson, F. (2003) “The Knowledge Argument” In: *Richmond Journal of Philosophy* 1, no. 3, pp. 6–10.

James, W. (1890) *The Principles of Psychology. Vol. 1.* New York: Henry Holt and Company.

Kim, J. (1998) *Mind in a Physical World. An Essay on the Mind-Body Problem and Mental Causation.* Cambridge: MIT Press.

Kim, J. (1999) “Making Sense of Emergence”. *Philosophical Studies* 95, pp. 3–36.

Kim, J. (2006) “Being Realistic About Emergence”, in: P. Clayton – P. Davies (eds.), *The Re-Emergence of Emergence. The Emergentist Hypothesis from Science to Religion.* Oxford University Press, pp. 189–202.

Kripke, S. (1980) *Naming and Necessity.* Cambridge: Harvard University Press.

Kripke, S. (2002) “Naming and Necessity”, in: D. J. Chalmers (ed.), *Philosophy of Mind. Classical and Contemporary Readings.* Oxford University Press, pp. 329–334.

Ladyman, J. (1998) “What is Structural Realism?”. *Studies in the History of Philosophy of Science* 29, no. 3, pp. 409–424.

Ladyman, J. (2009) “Structural Realism”, in: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition).* URL = <http://plato.stanford.edu/archives/sum2009/entries/structural-realism/>.

Leibniz, G. W. (1714/1898) “The Monadology”, in: *The Monadology and Other Philosophical Writings.* Oxford: Claredon Press.

Langton, R. (1998) *Kantian Humility: Our Ignorance of Things in Themselves*. Oxford: Clarendon

Press.

Langton, R. – Lewis, D. (1998) “Defining 'Intrinsic'”. *Philosophy and Phenomenological Research* 58, no. 2., pp. 233–245.

Levin, J. (2002) “Is conceptual analyses needed for the reduction of qualitative states?”. *Philosophy and Phenomenological Research* 64, no. 3, pp. 571–91.

Levine, J. (2001) *Purple Haze. The Puzzle of Consciousness*. Oxford University Press.

Levine, J. (2007) “Phenomenal Concepts and the Materialist Constraint”, in: Alter – Walter, pp. 145–66.

Lewis, C. I. (1929) *Mind and the World Order. Outline of a Theory of Knowledge.* New York: Charles Scribner's Sons.

Lewis, D. (1966) “An Argument for the Identity Theory”. *Journal of Philosophy* 63, no. 1, pp.17–25.

Lewis, D. (1995) “Should a Materialist Believe in Qualia?” *Australasian Journal of Philosophy* 73, no. 1., pp. 140–144.

Lewis, D. (2009) “Ramseyan Humility”, in: D. Braddon-Mitchell – R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism.* Cambridge: MIT Press, pp. 203–222.

Loar, B. (1999) “David Chalmers's The Conscious Mind”, *Philosophy and Phenomenological Research* 59, no. 2, pp. 465–472.

Loar, B. (2002) “Phenomenal States – Second Version”, in: D. J. Chalmers (ed.) *Philosophy of Mind. Classical and Contemporary Readings.* Oxford University Press, pp. 295– 310.

McGinn, C. (1989) “Can We Solve the Mind-Body Problem?”. *Mind* 98, no. 391, pp. 349–366.

Mihálik, J. (2013) “Russellův neutrální monismus a problém vědomí”, in: M. Soutor – T. Marvan – L. Dostálová (eds.), *Studie k filosofii B. Russella.* Praha: Filosofia.

Müller, T. – Watzka, H. (eds.) (2011) *Ein Universum voller 'Geiststaub'?. Der Panpsychismus in der aktuellen Geist-Gehirn-Debatte.* Paderborn: Mentis Verlag.

Nagel, T. (1974) “What Is It Like to Be a Bat?”. *Philosophical Review* 83, no. 4, pp. 435–450.

Nagel, T. (1979) *Mortal Questions.* Cambridge University Press.

Nagel, T. (1986) *The View from Nowhere.* Oxford University Press.

Newman, M. H. A. (1928) “Mr. Russell's causal theory of perception”. *Mind* 37, pp. 137–148.

Ney, A. (2015) “A Physicalist Critique of Russellian Monism”, in: Alter – Nagasawa, pp. 346–369.

Nida-Rümelin, M. (2007a) “Dualist Emergentism”, in: B. P. McLaughlin – J. Cohen (eds.) *Contemporary Debates in Philosophy of Mind.* Malden: Blackwell Publishing, pp. 269–286.

Nida-Rümelin, M. (2007b) “Grasping Phenomenal Properties”, in: Alter –Walter (2007), pp. 307–

338.

O'Connor, T. - Wong, H. Y. (2005) “The Metaphysics of Emergence”. *Noûs* 39, pp. 658–678.

O'Connor, T., Wong, H. Y. (2012) “Emergent Properties”, in: E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition).* URL = <http://plato.stanford.edu/archives/spr2012/entries/properties-emergent/>.

Parkes, G. (2009) “The Awareness of Rock. East-Asian Understanding and Implications”, in: Skrbina (2009), pp. 325–340.

Papineau, D. (2001) “The Rise of Physicalism”, in: C. Gillett – B. Loewer (eds.) *Physicalism and Its Discontents.* Cambridge University Press, pp. 3–36.

Papineau, D. (2002) *Thinking About Consciousness.* New York: Oxford University Press.

Papineau, D. (2006) “Comments on Galen Strawson”. *Journal of Consciousness Studies* 13, no. 6, pp. 100–9.

Papineau, D. (2007) “Phenomenal and Perceptual Concepts”, in: Alter – Walter, pp. 111–144.

Rey, G. (1995) “Towards a Projectivist Account of Conscious Experience”, in: T. Metzinger (ed.) *Conscious Experience.* Paderborn: Schöningh, pp. 123–142.

Rosenberg, G. (2004) *A Place for Consciousness: Probing the Deep Structure of the Natural World.* Oxford University Press.

Rosenthal, D. (1991) “The Independence of Consciousness and Sensory Quality”, in: E. Villanueva (ed.), *Consciousness: Philosophical Issues I.* Atascadero: Ridgeview Publishing Company pp. 15-36.

Rosenthal, D. (1993) “State Consciousness and Transitive Consciousness”. *Consciousness and Cognition* 2*,* no. 4, pp. 355–363.

Russell, B. (1954) *The Analysis of Matter.* Dover Publications, New York.

Russell, B. (2015) “Excerpts from *Analysis of Matter* (1927), *Human Knowledge: Its Scope and Limits* (1948), *Portraits from Memory* (1956), and *My Philosophical Development* (1959)”, in: Alter – Nagasawa, pp. 29–57.

Schroer, R. (2010) “Where’s the beef? Phenomenal concepts as both demonstrative and substantial”. *The Australasian Journal of Philosophy* 88, no. 3, pp. 505–22.

Seager, W. (1995) “Consciousness, Information and Panpsychism”. in: J. Shear (ed.), *Explaining Consciousness – The ‘Hard Problem’.* Cambridge: The MIT Press.

Seager, W. (2006a) “The ‘Intrinsic Nature’ Argument for Panpsychism”. in: Strawson, G. et al. *Consciousness and Its Place in Nature. Does Physicalism Entail Panpsychism?* Exeter: Imprint Academic, pp. 129–145.

Seager, W. (2006b) “Rosenberg, Reducibility and Consciousness”. *Psyche* 12, no. 5, pp. 1–15.

Seager, W. (2010) “Panpsychism, Combination and Combinatorial Infusion” *Mind & Matter* 8, no. 2, pp. 167–184.

Searle, J. R. (1997) *The Mystery of Consciousness.* New York: The New York Review of Books.

Searle, J. R. (1998) “How to Study Consciousness Scientifically”, in: S. R. Hameroff – A. W. Kaszniak – A. C. Scott (eds.), *Toward a Science of Consciousness II.* Cambridge: MIT Press.

Searle, J. R. (1993) “The Problem of Consciousness“. *Social Research* 60, no. 1, pp. 3–16.

Sellars, W. (1991a) “Philosophy and the Scientific Image of Man”, in: *Science, Perception and Reality.* Astacadero: Ridgeview Publishing Company, pp. 1–40.

Sellars, W. (1991b) “Empiricism and the Philosophy of Mind”, in: *Science, Perception and Reality.* Astacadero: Ridgeview Publishing Company, pp. 127–224.

Shani, I. (2015) “Cosmopsychism: A Holistic Approach to the Metaphysics of Experience”, *Philosophical Papers* 44, no. 3, pp. 389–437.

Shoemaker, S. (1980) “Causality and Properties”, in: *Identity, Cause, and Mind* Cambridge University Press.

Shoemaker, S. (1982) “The Inverted Spectrum”. *Journal of Philosophy* 79, pp. 357-81.

Skrbina, D. (2005) *Panpsychism in the West*. Cambridge: MIT Press.

Skrbina, D. (ed.) (2009) *Mind that Abides. Panpsychism in the New Millennium.* Amsterdam: John Benjamins.

Smart, J. J. C. (2002) “Sensations and Brain Processes”. in: Chalmers, pp. 60–68.

Smolin, L. (2015) “Temporal Naturalism”. *Studies in History and Philosophy of Modern Physics* 52, pp. 86–102.

Stoljar, D. (2001) “Two Conceptions of the Physical”. *Philosophy and Phenomenological Research* 62, pp. 253–81.

Stoljar, D. (2005) “Physicalism and Phenomenal Concepts”. *Mind and Language* 20, no. 5, pp. 469–494.

Strawson, G (1994) *Mental Reality.* Cambridge: MIT Press.

Strawson, G. (2006) “Realistic Monism: Why Physicalism Entails Panpsychism?”. *Journal of Consciousness Studies* 13, no. 10–11, pp. 3–31.

Strawson, G. (2010) *Mental Reality (2nd edition).* Cambridge: MIT Press.

Strawson, G. (2011) “Cognitive Phenomenology: Real Life”, in T. Bayne – M. Montague (eds.), *Cognitive Phenomenology*, Oxford University Press.

Strawson, G. (forthcoming) “Mind and Being: The Primacy of Panpsychism”, in: L. Jaskolla – G. Brüntrup (eds.), *Panpsychism,* Oxford University Press.

Stubenberg, L. (2010) “Neutral Monism”, in: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*.
URL = <http://plato.stanford.edu/archives/spr2010/entries/neutral-monism/>.

Taylor, J. H. (2013) “Physicalism and Phenomenal Concepts: Bringing Ontology and Philosophy of Mind Together”. *Philosophia* 41, pp. 1283–1297.

Tollar, V. (2012) „Panpsychistické motivy ve filosofii Ladislava Hejdánka“, in: E. Kohák, J. Trnka (eds.) *Hledání české filosofie. Soubor studií.* Praha: Filosofia, pp. 287–300.

Tononi G. – Koch C. (2015) “Consciousness: Here, There and Everywhere?”. *Philosophical Transactions of the Royal Society B* 370: 20140167. URL: <http://dx.doi.org/10.1098/rstb.2014.0167>

Tye, M. (2009) “Qualia”, in: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*. URL: <http://plato.stanford.edu/archives/sum2009/entries/qualia/>.

Van Gulick, R. (2001) “Reduction, Emergence and Other Recent Options on the Mind-Body Problem: A Philosophical Overview”, in: A. Freeman (ed.), *The Emergence of Consciousness*. Thorverton: Imprint Academic.

Van Cleve, J. (1990) “Mind-Dust or Magic? Panpsychism Versus Emergence”. *Philosophical Perspectives* 4, Action Theory and Philosophy of Mind, pp. 215–226.