Utrecht University

Faculty of Science

Master programme History and Philosophy of Science

Graduate Thesis

The Hard Problem for Physics

In Search of a Place for Consciousness in Physical Theories

Author
Otto Abel Lange

Supervisor
Dr. Guido Bacciagaluppi

Second reader
Dr. Niels van Miltenburg

April 1, 2019

Abstract

Although it is not endorsed by every theoretical physicist, I hold that - paraphrasing Abner Shimony - a theory of everything must 'close the circle'. That is, it must provide an account of the existence of consciousness in nature. Phenomenal consciousness, this intimate 'feel of me being me', is difficult to grasp in physical terms. David Chalmers famously distinguished the 'easy' cognitive problems of the mind from the true 'hard problem': physical theories describe worlds that could have been zombie-worlds as well, i.e. worlds occupied by creatures who act the way we do, but who do not share with us that special aspect we call phenomenal consciousness. That is to say, the facts of physics do not entail it.

In this thesis I explore how we should understand the problem from the perspectives of a discipline that seems to be concerned by its very nature with 'a view from the outside'. The burden of the hard problem of consciousness for physics seems to be concerned with putting a shift from the third– to the first person's perspective somewhere into physics. A deeper analysis of two concrete proposals for a physical theory of the mind reveals that specific interpretations of quantum mechanics take this first person's perspective as fundamental from the outset. It turns out that views on the ontology of quantum theory sometimes perfectly align with perspectives on the ontology of consciousness. I conclude that the essence of the hard problem for physics must not be concerned with an *explanation* of the existence of consciousness, but that it rather should be about the identification of metaphysically reasonable grounds to decide where to put it into nature.

Contents

Acknowledgments

It may appear that writing a thesis is a solitary activity, something that demands absolute isolation. Nothing could be further from the truth. A good research project needs inspiration for progress, it needs reflection as a prevention against taking the wrong turn, and it needs friends for finalization. I could not have written this text without the support of friends, colleagues, and family.

Of all the people who have helped me on this journey I first like to express my gratitude to my supervisor Guido Bacciagaluppi. I am thankful for the inspiring discussions we had, his critical reflections on aspects of my argumentation and for his 'antenna' for possible flaws. I also want to thank Niels van Miltenburg who, as a second reader, saved me from losing my balance in the slippery topics from the philosophy of mind.

Then, I want to thank my fellow students and others who are affiliated with the Masters programme HPS at Utrecht University. Special thanks go to all my friends who joined me on the trip to the Oxford conference in March 2017, and to those who together with me participated in the organization of the 19^{th} Foundations of Physics Conference in Utrecht.

A special thanks goes to Sir Roger Penrose. I had the pleasant opportunity to become his 'personal driver' during the Utrecht conference. We had nice conversations, not only on his ideas about physics, cosmology, Escher, and human consciousness, but also on his feelings concerning the position of philosophy of physics in general. These conversations were incredibly helpful to me in directing my research.

I also thank my colleagues at Utrecht University who patiently lent me an ear when I tried to explain that the hard problem of consciousness really is that hard.

And last, but absolutely not least, I owe much gratitude to my wife Petra and my daughters, Juno and Isabel, for their incredible patience.

To paraphrase Thomas Nagel, with 'the last word' I want to dedicate this thesis to my parents who, although they will not witness the end result, stood at the basis of my decisions to do the things I do.

Introduction

> *I regard consciousness as fundamental. I regard matter as a derivative of consciousness. We cannot get behind consciousness. Everything that we talk about, everything that we regard as existing postulates consciousness.*
>
> - Max Planck, *1933*

Many seem to hold that the world must be understood as ultimately constituted of only physical facts. Even more, also our perception of this world, that is, our awareness of it from a first person's perspective should be regarded as arising from physical aspects alone. This is a general description of the doctrine of *physicalism*, which promotes a monist stand regarding the relation between mind and matter: dualism is wrong in the sense that consciousness should not be considered as an independently existing substance beyond the material. Physical processes are all there is and our subjective experiences somehow arise from them. This text is concerned with an investigation of this claim from the perspective of physics itself.

It is generally agreed that the mind-body problem is one of the most profound of all issues in both philosophy and natural science. But a closer examination teaches that for a precise understanding of the problem many misconceptions need to be removed. I.e., it seems that the more precise one tries to formulate the problem, the more philosophy and physics seem to get caught up in confusing difficulties. In this thesis I will explore to what extent philosophical presumptions with regard to the mind-matter relation put a pressure on fundamental conceptual presumptions in physics. Such an exploration may reveal that both philosophy of mind and philosophy of physics can and must play a role in setting out the minimal requirements for the establishment of theories that claim to be 'about everything', the ultimate challenge for many theoretical physicists.

It is helpful to start with a sketch of some very intuitive thoughts about the connection between the human mind and physics. I expect these considerations to be familiar to everyone who once in a while contemplates her position as a small being in

an overwhelmingly extended but seemingly cold and mindless universe. Starting with an intuitive approach will be helpful in setting the stage, i.e. for the identification of the major issues, of possible background assumptions, and for the exact demarcation of the topic I want to engage with, the topic that is concerned with the search for a place for consciousness in physical theories. Such a quest will not solve the most difficult philosophical questions about human consciousness, but it is expected that it will at least give a deeper insight into what physics eventually should be about.

What am I? Whom could I ask this question? According to some popular accounts of modern neuroscience it is the operation of my neural system that I'm the sole result of. Obviously, the processes involved are highly complex and therefore not yet fully explained, yet in principle it is, according to these views, possible to comprehend how my awareness of 'me being me' arises from the physical brain processes involving neuron dynamics. And indeed, neuroscientists are often considered as the experts I could ask, since their field of expertise could in principle be used to explain everything that is going on in my head. However, perhaps eventually they will be able to describe what is physically going on in my head, but are they also the ones who could tell me what is going on in my *mind*? Are they the right experts to ask *what* I experience when I am aware of myself? Obviously not. The only person I could ask about the subjective experiences in my mind seems to be myself: I am my mind. To me it seems that I'm the only spectator within a stadium that is not directly accessible from the outside, and I have reasons, albeit no logical proof, that the same applies to other human beings as well. But even for myself it is difficult to understand *what* I am. It appears that my mind is a place where facts about the outside world are exposed. In my mind the world gets interpreted and yet I do not know how to interpret myself.

Suppose I'm having a nice cappuccino in the early morning sun somewhere in the Italian 'primavera'. My brain processes all the physical facts involved when I digest the smell and taste of my coffee, the impressions that come with the early sunrise above the mountains, communicated from the world via my retina into my inner self, packed as a fully private sensory picture. The qualitative aspects of these inner sensations feel to be coming from the momentary world around me, possibly mixed with memories from earlier experiences. In other words, the full picture in this instant of time seems to stem from the reality I somehow feel myself related to as a spectator. But I also feel like an *active* spectator, interacting with the stage. For instance, when I notice the sensible touch of the early spring I can deliberately focus my attention to continuation of this moment by turning my face towards the sun, i.e. my mind does not only receive signals from my senses, it also seems to control their usage at certain points.

I am aware of a fully internal assessment of this experience, but at the same time I feel I could try to externalize this feeling, perhaps by means of a description in natural language, possibly ready to be shared with a second person. It could be stored in a form of a personal persistent memory of a qualitative sensation. In that case I may be able to recall this experience for myself, possibly embellished with many qualitative aspects. Still, the only thing I feel really certain about is that the qualitative sensation of my experience is very intimate to me, intimate in the sense that it is my personal awareness of the knowledge I have about the context I seem to be situated in. Or to put it differently, it looks to me as if I'm constantly aware of the fact that I experience reality from a unique first person's perspective. But, however intimate, this perspective is highly mysterious to me at the same time. After all, paraphrasing Thomas Nagel, it is remarkable that I seem to be the only person who can answer the question '*what is it like to be me?*'[1] Perhaps I consider myself clever enough to answer it in colorful language, but the problem remains that I am the sole person who could do so.

As expounded by David Chalmers, the real 'hard problem of consciousness' is concerned with the explanation of the subjective aspect that we encounter when we process information coming from events in the world around us: "Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does."[2] The content of my conscious experiences seems partially to concern facts that are imposed upon me from the outside world, i.e. facts that eventually result from some sort of physical neurological processing of the changes in the world that are physically communicated via my senses. Therefore, it seems a natural idea that my subjective feelings are somehow connected to my physical brain during its processing of the physical impact from *my surroundings*. But, when I speak of surroundings, then I must ask myself the question 'what is surrounding what?' Is my mind situated in isolation within a physical world from which it is constantly triggered through physical signals that are processed to become meaningful?

The physicalist will probably hold that my brain hosts a *physical* feature through which physical facts represent themselves as subjective sensations. At the same time, it is well-known that our brain is a physically complex compound system, which implies that altering parts of it can have effects on different parts. So, whatever process in our neural system is assumed to bring forth the qualitative feelings we refer to as conscious experience, physical processes in the brain themselves can alter its overall state, possibly including the contents of my thoughts. In other words, the physical actions that trigger

[1] Nagel 1974, *What is it like to be a bat?*
[2] David J Chalmers 1995, *Facing up to the hard problem*, p.5.

thoughtful experiences do not necessarily reach me from the outside physical world, i.e. the world outside my physical brain. Therefore, without further looking at the outside world I could imagine my brain as an isolated *physical system that becomes aware of itself*. With this picture in mind we may paraphrase Nagel in a slightly different way by asking *'what is it like to be this system?'* Obviously, this is not the kind of questions physicists ask themselves when studying physical systems in the lab.

An apparent natural question for the physicalist is whether the qualitative feel of this system's subjective experiences has an influence on its state. Or to put it differently, could the qualitative feel of my experience be itself of influence on my subjective mental thoughts? Suppose it does not. Then the question arises what it means to consider it as physical.[3] In contrast, if it has an influence on the system's state then the effects should be observable. Now the difficulty for the physicist is how to explain a physical effect that is observed from a third-person's perspective as being caused by something physical that can only be observed from the first-person's perspective. So, without even having a clear picture of what 'physical' essentially implies, it seems that assigning physicality to a mental state introduces the problem of a necessary perspective-shift into physics.[4]

My final issue in this intuitive sketch is concerned with the *place* of the mind in the world: is the mind located in a physical world, and if so, how and to what extent? Where does it start and where does it end? Could it coincide with other minds? And perhaps most important for the current discussion, does it unambiguously coincide with physical phenomena, in which case one could imagine rephrasing the question in physical terms: where do processes that bring forth our mental events start and where do they end?[5]

What can be learned from the foregoing reflection? It reveals some issues that arise when one tries to understand the existence of subjective experiences from a picture of an apparently separate material world. Furthermore, what this contemplation especially shows is that all these issues are to some extent concerned with the confusion of perspectives. For example, when I speak of physical facts that come to me from the material

[3] For a discussion of difficulties with respect to a precise meaning of physicality in the context of physicalism, see Montero 2009, *What is the Physical?*

[4] The problem for physicalism may look even more profound. After all, ascribing causality to mental phenomena turns them into something real. To avoid a collapse into dualism these phenomena cannot be merely mental, but rather they must concern genuine physical facts.

[5] See Clark and D. Chalmers 1998, *The Extended Mind*. Andy Clark and David Chalmers promote what they call an 'active externalism' with regard to *cognitive* aspects of the functioning of our brain (p.7). In this view the content of our experiential *beliefs* is partially directly connected to artifacts in the physical outside world and our mental *self* should be considered as extended as well. Their paper starts with the promising question "Where does the mind stop and the rest of the world begin?" I doubt however whether they truly touch on the issue of the extension of the *mind* or whether they merely argue for an extended view on the brain.

world, I tacitly assume that I experience these facts from a first person's perspective, i.e. I consciously perceive them as pieces of a private exposition. At the same time, the world is described by physics from a third person's perspective. That is to say, the nature of physics is that it provides a view on things from the outside, descriptions that are publicly accessible. So, it seems natural to expect that the difficulty of putting consciousness somewhere into physical theories seems to be concerned with the difference between these two perspectives. As I already mentioned, physicists do not ask themselves what it is like to be the system they observe. But what about this question when the system is the human brain? The answer demands a shift in perspective, i.e. a shift from the third to the first person's view. But what I observed in my intuitive reasoning is that the answer will be private, which makes the question not really relevant for the physicist. What *is* relevant however is the question what the existence of two different perspectives implies for physics. Investigating them from the point of view of physical theories seems to me a requirement for finding a place to put consciousness into physical theories. This requirement basically demands an answer to the question whether physics can also, to some extent, provide a 'view on the inside'. Clearly, when one thinks about the observation and experience of facts in physics the focus will naturally be put on the role of the observer. Whereas quantum mechanics (perhaps) already demanded a view of the role of the observer as intertwined with the observed, studying consciousness from a physicist's perspective will likely beg for even more radical revisions of the basic perceptions of physics. After all, it seems that the consciousness of the observer should no longer be ignored, something which, as I already explained, introduces a first person's perspective on physical facts.

In the remainder of this thesis I will discuss what the burden of Chalmers's hard problem of consciousness is from the perspective of physics. In this respect I choose to refer to *the hard problem for physics*. At first sight, the foregoing discussion seems to show that the main difficulty is concerned with the seemingly absence of the first person's perspective in physical theories. However, we will see how a closer examination of two examples of physical theories of consciousness learns that this perspective is not totally absent in physics. Some interpretations of quantum mechanics rely on its existence and take is as a natural place to put consciousness into a physical picture of the world. I will show that there are reasonable options for doing so, but in the end all options depart from their own metaphysical assumptions. In this sense, my inquiry will reveal a natural interplay between philosophy and physics.

Structure of this text The first part of this thesis is concerned with the evaluation of the problem of consciousness from the perspective of physics. Chapter 1 is dedicated to a proper understanding of the problem. I discuss the most important arguments that are commonly used to show that consciousness is absent in physics. These are the zombie-argument and the explanatory gap.

In chapter 2 I delve deeper into the question of what is missing in physics to account for consciousness. The chapter starts with a discussion of Hempel's dilemma. I will look at options to mitigate its implications for the search for a physical theory of consciousness. From there on I focus on the essential aspects of the hard problem when it is presented to physics. A central topic in this part of the discussion is the absence of the first person's perspective in physical theories. At this stage I will introduce the notion of *first personal data*, a notion that will play a central role in the remainder of the text. A thorough investigation of what it means to deal with first personal data in an actual experimental setting reveals how the zombie-argument and the explanatory gap forbid an unambiguous interpretation of experimental data. I will discuss how this problem relates to the idea of immediate epistemic access, i.e. the need for a direct acquaintance with facts to assess them.

The second part of the thesis is dedicated to the application of what was discussed in part 1 to physics itself. This part starts with chapter 3, in which I discuss the role of the conscious observer. I provide expositions of von Neumann's interpretation of quantum mechanics, Stapp's psychophysical theory of the mind, and the many-minds view of Lockwood. All these discussions elaborate on the role of consciousness in physical observations.

The final chapter 4 is dedicated to a 're-assessment' of the hard problem, i.e. I will show how assumptions about the presence of consciousness can affect interpretations of quantum mechanics and vice versa.

A note on quotations All quotations in this text are taken literally from the original text, i.e. italics and quotation marks are represented as they appear in the source.

Part I

A hard problem for physics

Chapter 1

The mind in a physical world

All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.

- Immanuel Kant, *Critique of Pure Reason*

1.1 Demarcation on metaphysical grounds

The place of the conscious mind in a material world occupies both philosophers and scientists. Chalmers famously formulated the difficulty to explain why consciousness exists as 'the hard problem of consciousness'. Whatever one's position, it is safe to say that the issue is commonly understood as a both philosophically and scientifically profound problem, and no one seems to deny that it is far from settled. But, even with respect to what the problem entails one could already disagree. The most 'obvious' choice for *illusionists* like Dennett, Blackmore, and Frankish is to deny the problem: consciousness is illusory, i.e. phenomenal consciousness is an *introspective* illusion.[6] Obviously, who eschews the 'hard problem' in its familiar guise is left with a new intricate issue, namely, why does there seem to be a hard problem at all? Dennett often conforms to the 'magic metaphor' to emphasize the notion of illusion, but he acknowledges the new problem substitute: " [...] our burden is to figure out and explain how the 'magic is done."[7] Frankish refers to the altered challenge as the *illusion problem*: " [...] the challenge is to provide an account that explains how real and vivid phenomenal consciousness seems. This is the illusion problem."[8] And Blackmore adds: "For a drastic solution like 'it's all an illusion'

[6]Frankish 2016, *Illusionism as a Theory of Consciousness.*
[7]Dennett 2016, *Illusionism as the Obvious Default Theory of Consciousness.*
[8]Frankish 2016, *Illusionism as a Theory of Consciousness*, p.3.

even to be worth considering, there has to be a serious problem. There is. Essentially it is the ancient mind-body problem, which recurs in different guises in different times."[9] So, it seems that even the denial of consciousness as a real existent phenomenon does not relieve us from the mind-matter problem. For a true understanding of the world we are obliged to provide an explanation of why we perceive it in a specific way, i.e. to explain why our experiences seem to us the way they seem and why we have them in the first place.

1.1.1 The philosopher and the scientist

It must be noted that the relation between mind and matter can be considered from two seemingly opposite perspectives. On the one hand, the appearance of phenomenal consciousness, real or illusory, can be taken as a starting point for contemplation about what could be inferred about the physical world it supposedly relates to, and about our possible knowledge of that world. On the other hand, one could start with the results of present-day natural science and try to find out whether or how phenomenal consciousness could fit in the available scientific pictures of the world. It will be no surprise that the first perspective is taken by many philosophers (cf. McDowell, Sellars, Kant), whereas the opposite approach is most often chosen by natural scientists (cf. Penrose, Tegmark, Stapp). Of course, this is not uncommon in the human practice of seeking a total coherent picture of the world: we start digging intellectual tunnels from opposite directions and hope we'll all meet at some central point of understanding. Less common is a switch of perspective. That is to say, many keep digging their tunnels without bothering about the digging process in the opposite direction. As a result, one may fail in identifying a central meeting point because of deflections that inevitably result from the chosen 'digging tools', i.e. deflections with respect to the implicit understanding of the important concepts involved. Examples may be found in the ambiguous use of shared terminology, for instance as identified by the Israelian physicist Elitzur: "Although the term 'consciousness' is often used in some physical theories, almost nowhere in the physical literature has the fundamental problem associated with it been properly described."[10] This is not necessarily a form of careless practice from the side of the physicist, but often a prerequisite to keep science manageable. Sellars recognizes the issue and he acknowledges a role for the philosopher:

> The multiplication of sciences and disciplines is a familiar feature of the intellectual scene. Scarcely less familiar is the unification of this manifold which is taking place

[9] Blackmore 2002, *There is no stream of consciousness.*
[10] Elitzur 1989, *Consciousness and the Incompleteness of the Physical Explanation*, p.2.

> by the building of scientific bridges between them. [...] What is not so obvious to the layman is that the task of 'seeing all things together' has itself been (paradoxically) broken down into specialities. And there *is* a place for specialization in philosophy.
>
> ...
>
> It is therefore, the 'eye on the whole' which distinguishes the philosophical enterprise. Otherwise, there is little to distinguish the philosopher from the persistently reflective specialist;[11]

According to Sellars, it is the special disciplines that "know their way around in their subject-matters." It is up to philosophers to decide to what extent they have to be familiar with these disciplines as well, in order to keep their eyes on the whole: "Yet if the philosopher cannot hope to know his way around in each discipline as does the specialist, there is a sense in which he can know his way around with respect to the subject-matter of that discipline, and must do so if he is to approximate to the philosophic aim."[12] I am tempted to add that, because of her 'eye on the whole', the philosopher is a domain specialist in a very peculiar way: she's expected to have a view on *which* subject-matters should be involved. That is to say, in the spirit of Sellars's picture and in the light of the present context the following natural question presents itself: Which discipline deals with the study of phenomenal consciousness? Who takes care of the subject-matters involved? One may be tempted to point to psychology or neuroscience, but when it really comes down to the explanation of the existence of experiential aspects, i.e. to the kind of 'why does this feel like to be me-questions', all known scientific disciplines grope in the dark. In fact, who are the subject-matter specialists with respect to Chalmers's hard problem of consciousness? Should it ultimately be physicists? Is that the conclusion from strong physicalism? I think the problem is so hard because there are no scientific specialists in these matters. Or should I rather say, there are none because the problem is so hard? After all, one may contend that consciousness cannot be studied scientifically because science rules out the first person's perspective.[13] Nevertheless, the philosopher and the scientist are on the same page in this regard, but a proactive role for the philosopher is highly demanded. Indeed, as Hilary Putnam reminds us: "But the putting-forward, not of detailed and scientifically "finished" hypotheses, but of schemata for hypotheses, has long been a function of philosophy."[14] When it comes down to physics

[11] Sellars 1963, *The Scientific Image of Man*, p.37.
[12] Ibid., p.36.
[13] Defenders of this position may claim that the 'tunnel' starting from the perspective of science will lead to nowhere. However, scientists working on theories of consciousness will not agree. Perhaps these theories fail to reach beyond merely postulating consciousness, still they may point at reasonable locations where a first person's perspective could fit in the scientific picture.
[14] Putnam 1992, *The Nature of Mental States*.

and consciousness, the philosopher's part should not be concerned only with interpreting newly found theoretical constructs, but it should certainly also be about where to find them. After all, the great challenge with respect to the mind-body problem is to decide whether this scientifically undeveloped area can in principle be conquered by a scientific discipline, or whether it will remain under the undisputed authority of pure metaphysics.

To this day there is no promising empirically verifiable candidate for a physical explanation of consciousness available. In this area of one of the toughest unresolved issues the physicist and the philosopher are sometimes crossing borders between their disciplines because they are studying shared subject-matters. An important shared topic concerns the question of what makes a phenomenon physical. Sometimes this issue is translated into questions about the contrasting class, that is, what does it mean to regard a substance as non-physical? Debates in the philosophy of mind about physicalism and dualism often boil down to disagreements about the relation between these two frequently vaguely understood notions. What will become clear in the course of this text is that this relation also figures in debates about philosophical interpretations of physical theories. I.e., later on we will observe how philosophically biased views on consciousness pervade deep into the heart of physics.

1.2 Available options and the horizon of science

The hard problem is not solved, i.e. no field within science does yet convincingly explain phenomenal consciousness, nor is even close. I take this as a fact. The truly difficult aspects of the mind-body problem are fully covered in metaphysical dust. In this section I will show why Chalmers's problem is so hard, i.e. I will discuss how the observation that physics nowhere entails the existence of consciousness threatens the physicalist's claim that ultimately mental phenomena arise from physical facts alone. This observation is packed into an argument that aims to show that there is no need for consciousness in a world described by physical theories. An understanding of this so-called 'zombie-argument' is necessary for the right appreciation of the problem for physics. But before turning to the argument I will briefly consider the available options for approaching consciousness from natural science. I will use a simple classification scheme that is described by Frankish and that covers the different metaphysical positions that are available with regard to the mind-body problem.

1.2.1 Consciousness and natural science

In order to structure his arguments in favor of illusionism, Frankish identifies three general types of theories of consciousness. Each type represents a different view on the possibility of explicating phenomenal consciousness through natural science.[15] Two of them define a realist position with respect to conscious phenomena, the third one is anti-realist. This division into three disjunctive classes is exhaustive in the sense that every view on the place of mind in the world will fit into one of these classes. For a simple explanation of Frankish's partitioning, let us assume we are given a specific scientific world picture. Then, when we encounter a phenomenon X we may ask ourselves the ontological question whether X is *real*, or only *apparent*. If X is apparent, i.e. an illusion, then we do not have to scientifically explain its existence. Obviously, it still leaves us with the difficulty of the explanation of how this illusion comes about, i.e. we are left with a variant of Frankish's illusion problem. If X is real we can decide that X is either explainable or unexplainable within our given scientific worldview. So, we end up with three types of possible theories: 1) X does not have to be explained from scientific theories, but its appearance must be clarified, 2) X must and can be explained from current scientific theories, and 3) X could perhaps be explained from scientific theories, but the available scientific framework is insufficient to do so. Note that the last category includes theories that deny explanation of X by science at all. When one takes X to be phenomenal consciousness, these three categories are respectively referred to by Frankish as 'illusionism', 'conservative realism', and 'radical realism'.

When Frankish coined the term radical realism, he referred to theories that take consciousness as a real, but from the perspectives of modern science, *anomalous* phenomenon. In other words, according to radical realists, scientific theories in their current form are unable to explain or predict the existence of it: " [...] there is radical realism, which treats phenomenal consciousness as real and inexplicable without radical theoretical innovation."[16] Important examples are varieties of dualism, and theories that call for a new physics, i.e. 'radical physicalism'. These are opposite positions in the light of facing the problem: is it scientifically soluble or not? Now, Chalmers describes physicalism as " [...] the doctrine that the physical facts about the world exhaust all the facts, in that every positive fact is entailed by the physical facts."[17] Moreover, he asserts that if the physical facts do not entail all facts about the world, then this basic claim of physicalism must be false. One can easily comprehend how forms of dualism arise

[15] Frankish 2016, *Ilussionism as a Theory of Consciousness.*
[16] Ibid., p.2.
[17] David J Chalmers 1996, *The Conscious Mind*, p.110.

from Chalmers's formulation of physicalism: there are non-physical phenomena in the world and consciousness is one of them. Indeed, as Chalmers argues, it is the failure of physicalism that forces us into a kind of dualism.

Robinson describes dualist theories as views that claim that " [...] the mental and the physical are both real and neither can be assimilated to the other."[18] This formulation emphasizes the presumed non-physical character of consciousness. It should be clear at the outset that dualistic perceptions of consciousness do not perceive it as illusory. The aspect of insolubility is packed up in the introduction of something 'beyond the physical', i.e. an entity that is obviously not entailed by physical facts. For dualists the mind-matter relation really is concerned with two *distinct* property classes. But what about the sustainability of this distinction in the light of possible future innovations in physics? After all, the dualist assumption that the present laws of physics cannot account for phenomenal consciousness rests on the view that it is a non-physical phenomenon. However, will this view survive if profoundly revised ideas about *what makes a phenomenon physical* were developed? It seems that the distinction between versions of dualism and varieties of physicalism depends on the question what physicality is about. This is more than a triviality, for what I want to refer to as *persistent dualism* – that is, strong versions of dualism that hold that phenomenal consciousness will *always* be non-physical, regardless of future scientific achievements – logically packs the conclusion that physics-based science can only describe a limited portion of reality. Thus, such a strong form of dualism conceptually marks out the realm of physics, something which is implicit in Robinson's reading.

1.2.2 Does the physical world need consciousness?

I will now turn to the question whether the world described by physics needs mental phenomena. The physicist Steven Weinberg seems confident that the gap between our physical world picture and a full description of reality may become very small in due time:

> It is not unreasonable to hope that when the objective correlatives to consciousness have been explained, somewhere in our explanations we shall be able to recognize something, some physical system for processing information, that corresponds to our experience of consciousness itself, to what Gilbert Ryle has called "the ghost in the machine". That may not be an explanation of consciousness, but it will be pretty close.[19]

[18] Robinson 2017, *Dualism*, in *Stanford Encyclopedia of Philosophy.*
[19] Weinberg 1994, *Dreams of a Final Theory*, p.45.

Weinberg appears to keep open the possibility that a gap, although perhaps a small one, in our full understanding of the world will remain. This does not imply that consciousness should be placed outside the physical realm, it rather puts a limit on what can in principle be explained in physical terms. This observation aligns with Chalmers's claim that, although consciousness does not *logically* supervene on the laws of physics, a form of *natural* supervenience should not be precluded. In fact, he claims that observed correlations between physical processes in our brains and our mental states provide strong reasons for the belief that our mentality somehow rests on physical facts, although it cannot be deduced from physical laws. He underpins his idea with the assumption that a full physical replication of a human will include her mind as well: "It remains as plausible as ever, for example, that if my physical structure were to be replicated by some creature in the actual world, my conscious experience would be replicated too."[20] Chalmers considers his view as a form of dualism that aligns well with natural science: "Although it is a variety of dualism, there is nothing antiscientific or supernatural about this view."

The Cambridge physicist Pippard, cited by Weinberg in his considerations about the mind, declared in his Eddington Memorial Lecture: "What is surely impossible is that a theoretical physicist, given unlimited computing power, should deduce from the laws of physics that a certain complex structure is aware of its own existence."[21] Thus, also Pippard puts a limitation on what could be achieved by deduction from physical laws, i.e. he delimits the scope of physics-based science by excluding it from studying consciousness from a first person's perspective. In fact, his observation relates to the difficulty that comes with the 'public-private' issue, i.e. the fact that science seems to be concerned with *public* rather than *private knowledge*:

> All too rarely do I find colleagues who will assent to the proposition (which I find irresistible) that the very ground-rules of science, its concern only for public knowledge, preclude its finding an explanation for my consciousness, the one phenomenon of which I am absolutely certain. Mostly they admit indeed that it will be a tough job, but like to believe that in due course the relationship of consciousness to brain activity will be made clear, and the ghost in the machine exorcised.[22]

Pippard expresses his belief that phenomenal consciousness presents a principal barrier

[20] David J Chalmers 1996, *The Conscious Mind*, p.110. Note that difficulties may arise from this somewhat rash use of terminology: if we want to replicate 'Chalmers's physical structure', what then do we need to replicate? It seems that the answers may vary from his body (including his brains), up to the entire universe.

[21] A. B. Pippard 1988, *The invincible ignorance of science.*

[22] A. B. Pippard 1992, *Counsel of despair.*

for what the laws of physics can logically account for. This barrier is metaphysically shaped by Kripke in the form of a a conceivability argument against physicalism, an argument that underlies the earlier mentioned zombie-argument.[23] Kripke formulated it as an attack on the *mind-brain identity thesis*, which generally holds that mental states are identical to neurological brain states.[24] Kripke's argument rests on the metaphysical necessity of truth:

> We ask whether something might have been true, or might have been false. Well, if something is false, it's obviously not necessarily true. If it is true, might it have been otherwise? Is it possible that, in this respect, the world should have been different from the way it is? If the answer is no, then this fact about the world is a necessary one. If the answer is yes, then this fact about the world is a contingent one.[25]

Now, the argument runs as follows. For the materialist the statement 'mental states are identical to brain states' is necessarily true, so it must be true in all logically conceivable worlds. However, there are logically conceivable worlds in which this fact does not hold, so it must be contingently true, which undermines the physicalist's position: the physical facts do not *necessarily* entail the identity of brain states and mental states, i.e. the latter do not follow from the first by necessity. This is underpinned by the conceivability, i.e. the logical possibility, of a world that is identical to ours, but fully inhabited by zombies, actors physically fully identical to us, but without consciousness.[26]. Another mode of explicating the argument is provided by Chalmers: "When God created the world, after ensuring that the physical facts held, *he had more work to do*".[27] That is, it is conceivable that God would have left a world physically identical to ours, albeit without any phenomenal consciousness in it, so in fact a 'zombie-world'.

Thus, the conceivability argument holds as the basic problem for physicalism that phenomenal truths do not supervene with metaphysical necessity on the physical truths.[28] The crux of this claim is that a basic commitment of physicalism must be false in the first place. This commitment, formulated first by Jackson and repeated by Lewis and Chalmers, is understood as follows:[29]

[23] Kripke 1972, *Naming and Necessity.*
[24] Smart 2017, *Brain/Mind Identity.*
[25] Kripke 1972, *Naming and Necessity*, p.198.
[26] For a wider exposition of the zombie argument, see for instance David J Chalmers 1996. A criticism of the argument is discussed in Kirk 1999. A discussion of philosophical zombies is to be found in Kirk 2015
[27] David J Chalmers 1996, *The Conscious Mind*, p.110.
[28] Stoljar 2005, *Physicalism and Phenomenal Concepts.*
[29] See Balog 1999, *Conceivability, Possibility, and the Mind-Body Problem.* This formulation of the claim is due to Jackson 1982 and used by Balog throughout her paper. The versions of Lewis and Chalmers are respectively to be found in Lewis 1983 and David J Chalmers 1996

1. *Two worlds are physical duplicates if and only if they agree on all the true statements expressed in the language of physics.*

2. *A minimal physical duplicate of a world is achieved when only its physical nature is replicated.*

Physicalism demands that

(C) Any world which is a minimal physical duplicate of our world is just a full copy of our world.

Obviously, the central point of the conceivability argument is to undermine the commitment C by showing that this is not a metaphysical necessity: the zombie-world is a minimal physical duplicate of our world, yet it is not a full copy.

Levine translated Kripke's metaphysical argument into an epistemological version, i.e. into a variation of the conceivability argument that is mostly referred to as the problem of the *explanatory gap*. His argument is not directed at the refutation of physicalism, but rather at the problematic determination of its truth. Indeed, although we cannot determine which psycho-physical identity statements are true, we also cannot conclude from Levine's argument that materialism is false. But psycho-physical identity statements at least leave an important 'explanatory gap':

> Unlike its functional role, the identification of the qualitative side of pain with C-fiber firing (or some property of C-fiber firing) leaves the connection between it and what we identify it with completely mysterious. One might say, it makes the way pain feels into merely a brute fact.[30]

These lines express the essence of the hard problem for physics, i.e. the issue I announced at the beginning of this section: from the perspective of physics conscious experience seems to be 'completely mysterious', a 'brute fact'. It is not entailed by physical theories, which makes it possible to conceive of zombie-worlds that are physically identical to ours. In fact, the absence of consciousness in physics explains why the problem is so hard.

[30] Levine 1983, *Materialism and Qualia.*

Chapter 2

Physics in a mindful world

Consciousness cannot be accounted for in physical terms. For consciousness is absolutely fundamental. It cannot be accounted for in terms of anything else.

- Erwin Schrödinger, *The Observer, 1931.*

2.1 Hempel's dilemma

In the previous section I observed how physics leaves an explanatory gap with respect to the existence of consciousness. So, if the materialist holds that the physical facts exhaust everything, it seems natural to ask the following: does the world entailed by these facts coincide with what can in principle be deduced from future laws of physics (and applicable initial conditions)? In other words, is it conceivable that the explanatory gap eventually vanishes? The difficulty for someone who wants to answer this question in the positive is that she cannot foresee what can be deduced from laws that are not yet discovered. Let us therefore question the physicalist's intuition: Which elements in the mind-matter relation are covered by physical facts? Physicalism-oriented scientists and philosophers alike would reply 'all'. They will assert that all aspects involved are in the core physical, i.e. every fact about the mind is entailed by physical facts. This is of course the thesis of physicalism, albeit phrased in a colloquial form. Dowell observes that a lot of disagreement has risen from the question of how to arrive at a less intuitive formulation or, in his own words, " [...] what exactly would have to be true for physicalism to be true?"[31]

[31] Dowell 2006, *Formulating the Thesis of Physicalism.*

Evidently, the possible truth of physicalism at least depends on the answer to *what makes a thing physical?* Or to put it differently, *what are physical facts?* An answer may naturally be expected from an appeal to physics. After all, the content of physicalism obviously must depend on how physics is interpreted. However, this is where the real trouble starts, for one bumps into what is called *Hempel's dilemma*, a dilemma that advances a problem of choice: *which physics* should the physicalist appeal to? Is it current physics, which is not generally taken as providing a full finalized description of the world? Or should she appeal to a future physics, of which the content is unknown? Carl G. Hempel presents the dilemma that carries his name with the following lines:

> I would add that the physicalistic claim that the language of physics can serve as a unitary language of science is inherently obscure: The language of what physics is meant? Surely not that of, say, 18^{th} century physics; for it contains terms like 'caloric fluid', whose use is governed by theoretical assumptions now thought false. Nor can the language of contemporary physics claim the role of unitary language, since it will no doubt undergo further changes, too. The thesis of physicalism would seem to require a language in which a *true* theory of all physical phenomena can be formulated. But it is quite unclear what is to be understood here by a physical phenomenon, especially in the context of a doctrine that has taken a determinedly linguistic turn.[32]

So, any physicalist's appeal to physics begs for clarification. The content of current physics is probably partially inaccurate and will be updated, and a future physics is reasonably expected to differ profoundly from its present-day predecessor. Further, it's also reasonable to assume that, although the content of physics will continuously increase, it will never be 'complete'. But the real issue for physicalism is not about the incompleteness of physics or the pursuit of an undisputed accuracy, but rather about the total absence of physical laws that can account for facts about mental phenomena. The question is whether the character of the mind-body problem obscures any view on what sort of physics is needed. In other words, it may turn out that a suitable future physics is not necessarily fully unimaginable to us if we can come up with some features it should minimally have. After all, when we have some characterization of a theory that could establish a convenient understanding of physicality then we do not have to know more about its future successors. Why is this so? The moment there is a physical theory that can account for consciousness there may still be plenty deep problems left, perhaps with respect to dark matter or the origin of the universe, or even to notions we do not know at all today. Yet, in such a theory statements involving consciousness are at least included,

[32] Hempel 1980, *Comments on Goodman.*

i.e. they are just as much subject to falsification as are other statements that are deduced from the theory. That is, the theory could be incorrect and the deduction of expressions about mentality could be misleading. However, *such a theory should at least provide a framework that is protected against conceivability arguments and freed from the burden of the explanatory gap.*[33] The problem is that we don't have such a theory. Now, the most important question is, although it is not available and we do not know its content, *if it is possible, what can still be inferred about such a theory*? Or to put it differently, to decide whether physicalism *can* be logically true it must be clarified what then is minimally needed from the side of physics. After all, the demand for a physicalist to come up with a definitive physics that can be appealed to is not reasonable. But I claim that the physicalist should at least be able to show that such a physics can exist, i.e. a physics that can account for the existence of consciousness, and that as such can provide the basis for understanding what physicality refers to. In a nutshell, I hold that, although the current framework of physics is insufficient to handle the mind-body problem and therefore incapable to explain what 'physical' means in the context of mental states, it is reasonable to demand from the physicalist a characterization of what exactly is missing. Instead of the mere acknowledgment of an explanatory gap, an analysis of this gap on physical grounds is required. That is to say, the explanatory gap itself must be expressed in the language of physics. If we can do so, then we'll have a target between the two horns of the dilemma, a target point that can be used in the assessment of proposals coming from the side of physics that claim to offer a solution to the mind-body problem. Thus, for a real understanding of physicality and for the resolution whether physicalism can be true, I once again suggest to turn the mind-body problem into a problem for physics. For finding out what form such a problem should have, it is helpful to have a closer look at the difficulty of the dilemma from the perspective of physics itself.

To appreciate the burden of Hempel's dilemma, let us imagine an early 20^{th} century physicist, a scientist saddled with concepts that together constitute the full classical physics up to 1900. This is a physics in which 'things' are characterized by their possession of properties that can objectively be determined through experiment. Isham

[33] According to Bokulich 2011, the most common response on the dilemma downplays the relevance of the details of physics, i.e. the only relevance is concerned with " [...] a grasp of the mental. Physicalism will then amount to the claim that mental properties or truths are non-fundamental: they are nothing over and above the truths about and properties of the micro-entities that make up organisms." However, this is not my approach, i.e. I keep open the possibility that mental properties *are* fundamental and that physicalism proves to be false. My concern is with establishing a proper guideline for thinking about physicality. In that sense, I am not taking the second horn or sidestepping the dilemma as Bokulich characterizes the above response. For more discussions about the prospects *for physicalists* taking the first or the second horn see Crane and Mellor 1990, Wilson 2006, or Montero 1999.

provides a nice exposition of the conceptual presumptions that underlie classical physics when he addresses Heidegger's famous question 'what is a thing?'[34] He observes that classical physics " [...] concentrates on the bundle of properties, or attributes, that adhere to the thing and make it what it is." Physical operations in the form of measurements enable us to obtain knowledge about these properties, a view in accordance with the idea of the 'object-subject split of scientific methodology'. That is, we can partition the world into observer and observed parts, all subject to the same laws of physics. The properties of things, or in physical terms the observable quantities, can be internal – for instance the mass or the charge of a particle – or external like its velocity or position. An ideally precise measurement of an observable quantity must yield one specific value because the applicable quantity *has that value* at the time of measurement. Now, imagine that our physicist decides around 1900 to set up an experiment like the famous one performed later by Stern and Gerlach in 1922.[35] The setup of the experiment and the predictions of the outcome can be totally expressed in classical physical terms. Fully unexpectedly, our physicist will find out that, when a beam of silver atoms is sent through an inhomogeneous magnetic field in a specified direction – against all predictions from classical physics – two separate beams will come out, one 'up' and one 'down'.[36] This is strange from the classical perspective because a continuous spread of values was expected. Instead, spatial quantization is exhibited. But, a more profound problem manifests itself when our physicist decides to combine three of his setups in a particular sequence. Details of the experiment aside, the results will suggest that a measurement in a specific direction will destroy earlier obtained values for a different direction. In other words, characteristics of physical systems that were supposed to be autonomous, i.e. available for independent measurements in subsequent isolated experiments, seem to be influenced by the acts of measurement themselves. Of course, this puts a high pressure on the fundamental view on the subject-object split: the observer interferes with the observed. With hindsight we know that the resolution of the problem demanded not

[34] Isham 2001, *'What is a thing?'* in Lectures on Quantum Theory, pp.63-77.

[35] Stern and Gerlach performed their measurements to compare the Bohr-Sommerfeld theory of the atom with Larmor's classical picture. For a lucid discussion of the Stern-Gerlach experiment, see for instance Feynman, Leighton, and Sands 1965, III- Ch.5,6. For a historical background see Franklin and Perovic 2016, *Experiment in Physics*, App.5 and Brown, B. Pippard, and Pais 1995, *Twentieth Century Physics*, p.165

[36] One may wonder whether our hypothetical physicist will interpret the experimental results as the discovery of electron spin. As Franklin remarks:"The Stern-Gerlach experiment provides evidence for the existence of electron spin. These experimental results were first published in 1922, although the idea of electron spin wasn't proposed by Goudsmit and Uhlenbeck until 1925 (1925; 1926). One might say that electron spin was discovered before it was invented."Franklin and Perovic 2016, *Experiment in Physics*

only a highly revised physics, but also that the new physics was grounded on totally different conceptions. Before performing his experiment our physicist had no clue about the new physics and he had clear expectations about the outcomes of the measurements. After the experiment he had results that were not entailed by the physical facts and the laws as he knew them: he had to abandon his familiar physical framework and seek for laws that could account for the new observed facts. The old framework had risen from investigating the everyday world around us, the world of rocks, stones, and electromagnetic interactions. But the measurements performed at the atomic scale revealed a whole new class of physical facts. Thus, the domain that was successfully described by classical physics turned out to be a partial coverage of reality. The set of physical facts seemed to have expanded in a crucial way, i.e. the new facts were not entailed by the available physical laws. The question is, was our hypothetical physicist subject to Hempel's dilemma?

To see whether this is indeed the case it is important to decide if our physicist confronted an explanatory gap in the sense of Levine. Brief inspection will show that the experimental outcomes were anomalous in the light of classical physics. However, the explanatory gap in Levine's sense is not concerned with an anomaly, but rather with a lack of coverage, i.e. an absence of necessary entailment. That is, the domain of consciousness is fully precluded from the applicability of physics. One could object that the physicist's case didn't reveal an anomaly either because the experimental results came from a new domain of application, i.e. micro-physics at the atomic scale. However, the classical laws were assumed to apply at all scales, i.e. they entailed facts on the new experimental scale as well. So, the physicist ran into anomalous results in the sense that the outcomes deflected from what was predicted. Obviously, this is also supported by the fact that he started with outcome expectations in the first place. In conclusion, our physicist did not have to deal with an explanatory gap in the spirit of Levine. Therefore, it was evident for him that a solution had to be sought in the revision of standing ideas, i.e. there was no problem of choice between appealing to an old paradigm that had nothing to say about the new unfolded problem, and a new totally unimaginable physics. Thus, our physicist did not have to deal with the dilemma. Still, there is something interesting to learn from this false analogy. Namely that the difficulty with Hempel's dilemma for physicists results from the *absence of a true anomaly.* That is to say, there is no deviation from otherwise expected phenomena. Because current physics does not include any statements about consciousness there also seems to be no room for theory revision in this respect. Or to put it differently, we do not have any epistemic grounds to choose the first horn of Hempel's dilemma, i.e. take current physics as the

basis for physicalism, because current physics does not address issues of consciousness at all. But, we can say something about a physics that is supposed to do so. I will investigate what this is in the next section.

2.2 Defining the hard problem of physics

What must physics be about to allow the inclusion of consciousness in its domain of study? Physics is concerned with data in the form of physical facts that are accessible from a third person's perspective. On the other hand, conscious experiences are only perceived from a first person's perspective. So, for the inclusion of consciousness in its domain we demand that physics can account for the existence of these two perspectives. To clarify what this means, let me start with a paradigm example of a conscious experience, namely the awareness of being in pain. Suppose I cut my finger and that, triggered by the physical sensory input, I have the experience of having pain. The conscious experience is characterized by an internal aspect, the 'feeling of me being in pain'. In other words, I encounter this experience from a first person's perspective, i.e. from a perspective that is private in the sense that the pain experience is only accessible to me, the one person to whom this experience is exposed: I am the only one who knows *what it is like to have it*. The feel of my pain, i.e. this internally accessible aspect, I take to be a quale. Indeed, I will talk of qualia when I mean to refer to the qualitative internal presentation or 'feel' of something (notably an experience), the 'what is it like-aspect' of it.[37][38] So, qualia concern a form of representational data that are only available from a first person's perspective.[39] The physical manifestation of my pain can be perfectly described in neurological terms. This description presents my pain experience as a publicly assessable physical fact. So, it seems that a physics that covers the full pain experience

[37] There are different understandings of qualia around. Moreover, several authors deny that qualia are representations. For more on discussions about qualia, cf. Tye 2018 and Nida-Rümelin 2015

[38] I use the terms *presentation* and *representation* interchangeably. These terms should be taken literally, i.e. with the usage of the latter I do not mean to refer in any respect to *representational theories of consciousness*.

[39] Another familiar example concerns the conscious perception of colors. A lot has been written on imaginary Mary who perceives a red rose. In his exposition of the knowledge argument, Jackson in fact claims that a quale concerns a fact about the world that is not entailed by facts that are accessible from a third person's perspective. (Jackson 1986, *What Mary didn't know.*) When we read about Mary we are tempted to have an internal representation of what she will experience, i.e. we may think to have an idea about the *redness* in the picture sketched by Jackson. However, the only thing we can truly agree on with Mary is that the color label 'red' seems to refer to a common wavelength when we assign the color red. We cannot decide whether the internal representations of Mary and myself are the same when perception of colors is involved. This is an example of problems that arise when thinking about 'other minds'.

must account at least for the existence of two totally different modes of presentation.

It is useful to consider the idea of internal presentations in a slightly different way. I regard it as convenient to deploy the notion of *phenomenal concepts* (PCs) for this goal.[40] When talking about an internal presentation, i.e. an exposition of facts that are only accessible from a first person's perspective, natural questions will be concerned with their constituents and their accessibility. In other words, how should we understand the contents of qualia and why do we have access to them? A relatively new approach for dealing with qualia is the use of PCs. Levine characterizes PCs as follows: "Phenomenal knowledge is knowledge of conscious experience. Phenomenal concepts are concepts associated with that knowledge: those that express phenomenal qualities from the experiencing subject's perspective."[41] So, the moment we have a conscious experience we may think of PCs being deployed as the building blocks of an internal subjective representation of that experience. Balog points to the role of PCs in the epistemic access of qualia: "When we deploy phenomenal concepts introspectively for some phenomenally conscious experience as it occurs, we are said to be acquainted with our conscious experiences."[42] This acquaintance is not only private, but it also provides unique access to our conscious states: "We know our conscious states not by inference but by immediate acquaintance which gives us direct, unmediated, substantial insight into their nature."[43] In fact, PCs constitute conscious experiences. As a consequence, conscious experiences are characterized by PCs and the access to such an experience coincides with the access to the PCs. Formulated this way we may say that conscious experiences are internal presentations themselves.

The foregoing reveals two important aspects that somehow must be accounted for in a physical theory of consciousness. On the one hand we recognize two different perspectives from which facts are perceived, on the other hand we observe a significant difference in the epistemic access with regard to these perspectives. The familiar physical facts can be recorded, described, and (in principle) assessed by anyone at any time. This does not apply to mental phenomena. Indeed, we have knowledge about our conscious states because of our immediate acquaintance with them by the deployment of PCs.

[40] The *locus classicus* for the deployment of PCs in favor of physicalism is Loar 1990, *Phenomenal States*, which basically describes the so-called *Phenomenal Concept Strategy*. A discussion of this strategy is not relevant for the current context, but for critical expositions of it see for instance Stoljar 2005 and Demircioglu 2013

[41] Levine 2006, *Phenomenal Concepts and the Materialist Constraint*, Introduction.

[42] Balog 2009, *Phenomenal Concepts*. Balog refers to the term 'acquaintance' as it was introduced by Russell when he described the distinction between 'knowledge by acquaintance and knowledge by description'.

[43] Ibid., p.299.

That is, we have *direct* epistemic access to our experiences because of their internal-only presentation, something which is impossible for someone else. Balog refers to this aspect as 'asymmetric epistemology'.[44] So, our conscious experience can certainly not be assessed by anyone at any time. Even more, direct epistemic access means that even for myself accessibility is restricted to the sole moment I have the experience.

With all these considerations in mind we come a step closer to the hard problem of consciousness in a form that can be presented to physics. I observe that thus far two important classes of issues force themselves upon physics:

1. *The problem of internal presentation*
    - Is it physically conceivable that externally accessible facts are accompanied with an internal (only privately accessible) presentation?
    - Is it physically conceivable that internal presentations could have an effect on external (physical) facts?
2. *The problem of acquaintance*
    - How could direct epistemic access fit in a physical theory?

We are a step closer, but before bridging the gap towards a real problem for physics it is important to recall what current physics itself is about. Physics, as we know it, is concerned with both directly and indirectly observable *public* facts in the form of true statements about physical entities. Physical entities are for instance all sorts of particles, exotic and less exotic, relativistic fields, energy and mass distributions, etc. Because of the constant growth of the field the current list of physical entities is not expected to be exhaustive. The truth-value of physical statements is governed by physical laws and initial conditions, which together lay down the dynamical structures of physical processes. Obviously, the laws that describe these processes may be expected to change. With public facts I mean to refer to statements that can be shared among different observers without the need to perceive them from a first person's perspective in order to assess their possible truth. This is in a nutshell a common view on the content of physics. In this view there seems to be no place for something like an internal representation. Because all of physics is concerned with public facts, there also seems to be no problem of acquaintance, i.e. a role for direct epistemic access.

However, a closer look at how physics explains higher phenomena shows that things may be a bit more complicated. One aim of physics is to understand these phenomena

[44] Balog 2009, *Phenomenal Concepts.*

as the resultant of the behavior of the smaller physical entities that obey fundamental laws. This makes it for instance possible to understand complex biochemical processes as emerging from fundamental underlying physical processes. Clearly, this is the idea of reductionism, and its success cannot be overestimated. However, it is difficult to imagine how complex processes as the performance of Beethoven's 5^{th} Symphony could emerge from *meaningless* fundamental interactions in a comparable way. In fact, I do not think that the true difficulty is in the emergence itself, but rather in the fact that the spectator assigns meaning to the performance. If the music performed indeed emerges from purely physical facts, then we could regard the spectator as the observer of physical events. But, to make the music meaningful instead of merely an arbitrary sequence of waves, we seem to rely on an *interpreting observer*. Moreover, for the music to become meaningful the observer encounters an internal representation of having a listening experience. So, the question is how physics should explain that public facts about the superposition of sound waves lead to the experience of them via direct epistemic access. Obviously, there must be a shift in perspective involved, but this is clearly not the case in the standard account of reductionism I referred to above. In fact, qualia are not the 'standard' higher phenomena because they can only be studied by someone who is acquainted with them.

So, we have that immediate acquaintance with our experiences gives us private epistemic access to them. For example, the conscious observer of the performance of Beethoven's symphony is the only person who has epistemic access to his listening experience. Now, let me consider the act of observation in physics, for instance reading off the position of a pointer in a measurement apparatus. The observer records a value that can be shared as a public fact. But the observer's conscious perception of this fact is accompanied with an internal presentation, i.e. he is immediate acquainted with his experience of perception.[45] In fact, every conscious perception of a public fact, even the ones he recorded earlier, is presented internally as an experience. With this in mind, one could ask what it actually means to consult public facts from a third person's perspective. It seems that these facts at least provide input for internal presentations, but in the end the observer is only aware of non-public facts. This is what we would expect, because the third person's perspective is not accompanied with acquaintance. All we seem to know about public facts is that different observers agree on their subjective internal presentations when consulting them. Without going into details of realism vs anti-realism debates, it appears that consciousness has already a peculiar role to play in physics. That is to say, public facts are shielded from the *direct awareness* of the

[45] It is important to note that the internal presentation is not about the pointer's value, but rather about the act of perceiving this value.

observer. In the hope to consciously perceive them, the observer is confronted with the necessity of an intervention, i.e. an observation that leads to a conscious experience.

The foregoing line of reasoning may sound speculative, but it hides a pressing question: What should we think of physics without a conscious observer? This question was already raised in the 1930s when physicists were thinking about what quantum mechanics tells us about the world. In the next chapter I will explore assumptions about a role for conscious observers in quantum theory. The point I want to make here is that the picture of physics as providing a *publicly accessible* picture of reality contains some hidden assumptions from the start. Many (but certainly not all) physicists may be convinced of the idea that they are observing objectively existing physical facts, still it seems that the only thing they can feel certain about is their immediate acquaintance with an experience of observation and the apparent agreement about it with other observers. In fact, all conscious observations are experienced from a first person's perspective. The problem is to understand why there is indeed agreement about these observations. This is not a plea for a reinterpretation of the content of physics, but rather it shows that questions about internal representation and acquaintance were implicit aspects of physics from the beginning. In other words, there is something to say about a role for consciousness in physics, and there must be something to mention about the presentation of facts that are only available from a first person's perspective. In this light it is good to recall Thomas Nagel:

> What is needed is something we do not have: a theory of conscious organisms as physical systems composed of chemical elements and occupying space, which also have an individual perspective on the world, and in some cases a capacity for self-awareness as well.[46]

Nagel's 'individual perspective on the world' coincides with the introspectively accessible presentation of facts. Clearly, a successful physical theory that includes consciousness must have a place for these internal representations, i.e. for *first personal data*. When I speak of first personal data I refer to facts that are only accessible from a first person's perspective. In this sense I am really referring to private knowledge, or in Nagel's words, an individual perspective. As I already observed, this perspective has been with physicists all the time, but thus far it has not been explained in a satisfactory manner. (Although certain physicists claim they have.) The burden for physics is to give this perspective a place in physical theories. It is an open question whether this should lead to a new physics, or rather a different interpretation of physics as it is. David Chalmers contends that any appeal to a new physics will fail in explaining consciousness:

[46] Nagel 1989, *The View from Nowhere*, p.51.

> The trouble is that the basic elements of physical theories seem always to come down to two things: the structure and dynamics of physical processes. [...] But from structure and dynamics, we can only get more structure and dynamics. [...] No set of facts about physical structure and dynamics can add up to a fact about phenomenology.[47]

If all new physical theories indeed basically were to come down to the extension of structure and dynamics, then I agree with Chalmers. But I do not see why physics could not be extended in different directions. For instance, a different view on the role of the observer could contribute to a solution of the problem. As I already pointed out, the question is whether this should be referred to as a 'new physics' or rather as a 'new interpretation of physics'. However, a dispute on this issue could be closed if a theory that starts from a profoundly new role for the observer can produce new observable facts.[48] So, for now I will not follow Chalmers and I leave this question as it is. That is, I grant physics the chance to come up with a scientific explanation of consciousness.

It seems it should be possible by now to bridge the gap and to present the hard problem of consciousness in a form that could be digested by physics. What the hypothetical physicist in section 2.1 did not foresee was that he had to give up the subject-object split. In other words, the idea of an objective non-interfering outside observer had to be abandoned in the realm of atomic physics because at the quantum level the observer necessarily interferes with the observed and experiments partially enforce their own results.[49]This is truly a paramount change for a physicist who expects to obtain an objective picture of nature by probing it through experimentation. But from the former discussion about internal representations I infer that in a successful physical theory of consciousness an even more profound transition is required. That is, if physics needs to include our own mental phenomena in its domain of study, then it seems required that

> Req_1: *The* ***conscious*** *observer must become part of the observed.*

Since consciousness can not be observed without direct epistemic access, one may wonder how the conscious observer should be put in a typical physical experiment, i.e. an observation in physics. In fact, the issue of acquaintance already demands that in such a case the observer must observe herself. In the next section I will use a thought experiment

[47] David J Chalmers 1996, *The Conscious Mind*, p.118.

[48] In the next section I will use a thought experiment to illuminate some of the profound difficulties that come with the search for such facts.

[49] This is the traditional way of putting it. It must be noted that various approaches to the foundations of quantum mechanics apply a somewhat different understanding of the fundamental change physics had to undergo, but all agree that some fundamental change took place.

to find out which difficulties arise in such a scenario. But we can already anticipate these problems by using a slightly different formulation:

> Req_2: *The observation must include the role of the* ***consciousness*** *of the observer herself.*

So, where the transition from classical to quantum physics was concerned with difficulties with respect to the separation of observed and observer, a physical theory about consciousness must either explain how the observer is conscious or, if this can not be clarified in physical terms, what role consciousness plays within physical observations. I regard the formulated demands for physics as 'candidate' requirements. After all, it is an open question whether physics can undergo the suggested transitions. This is something I will explore in the remainder of this text.

There is a subtle but at the same time important difference between the two formulations. The first focuses on how to get the consciousness of the observer into the domain of study, i.e. it emphasizes the problem of the needed shift from the third to a first person's perspective, i.e. the problem of the inaccessibility of first personal data. It appeals to the familiar understanding of the hard problem in the form of the explanatory gap, i.e. why should physical processes give rise to consciousness? Or to put it differently, how could a seemingly inanimate physical system become a conscious observer of its surroundings? The second formulation departs from the act of observation and seeks an answer to the question why consciousness should be involved. As already mentioned, I will explore the difficulties with respect to the first formulation in the next section. An investigation of the consequences of the second formulation will follow in part II of this thesis. For now it is important to understand what the distinction between the two entails. When I state that 'the observation must include the role of the *consciousness* of the observer herself' I mean to refer to the idea that the classical subject-object split totally vanishes in the sense that, not only the observer becomes necessarily part of the observed system, but that her consciousness is pulled within as well. That is to say, the fact that the observer is conscious may play an essential role in physical theories. Ideas about an integration of a conscious observer into physical theories are to be found in some (but certainly not all) theories that try to explain the measurement problem in quantum physics. Propositions in this realm are not starting from pictures of what conscious is about, but rather they stem from considerations about what quantum measurement is about. Examples along these lines are to be found in the ideas of von Neumann, London and Bauer, Wigner, Stapp, and Lockwood. The opposite formulation, i.e. 'the *conscious* observer must become part of the observed' takes consciousness per se as the point of departure. I.e.,

this approach tries to pull the topic of mental phenomena into the domain of physics. Obviously, this means that an explanation for the existence of first personal data must play a significant role from the start. Examples of theories that try to *explain* consciousness from within physics are to be found with Penrose and Tegmark. Of course, the distinction I make between these separate approaches is not always that sharp. In the end, all physical theories about consciousness come down to an explanation of the role of mental phenomena, but for an understanding why such a role is proposed I consider it helpful to acknowledge this distinction.

To end this section, it is clear that the major transition that is needed must involve somehow the inclusion of first personal data. Difficulties for such an inclusion come from our acquaintance with these data and from their role as internal representations. Both these aspects can be implicitly packed in an epistemological form that can be used for a deeper inspection of the first requirement above. This will be the topic of the next section.

2.3 The trouble with first personal data

The first candidate requirement from the previous section holds that a physics that can account for the existence of consciousness must approach the conscious observer as part of the observed. It is a requirement for theories that claim to provide a physics-based explanation of how mental phenomena arise. In physics we explain phenomena by observing all the physical facts that somehow seem related to them. So, the question is, can we explain consciousness by observing relevant facts? In the previous section I explained that we have only direct epistemic access to our experiences. Therefore, it seems already that consciousness is a totally different phenomenon when compared with the usual ones we study in physics. That is to say, to investigate it an observer must have immediate access to the phenomenon, something which is obviously reserved for the observer who will be part of the observed. However, the question is whether a decision about the *presence* of a conscious experience always demands direct access to the experience itself. After all, even without knowledge about the exact contents of an experience, we may already be able to identify the public facts involved when someone encounters it. This could help us understand the physical circumstances that give rise to consciousness. This strategy motivates the following epistemological form that can help to obtain a deeper insight into the first requirement above:

Q1: Can the presence of first personal data be explained or inferred from physical experiments?

Q2: Is it possible to observe – without presuming its presence – first personal data in physical experiments?

Q3: Do we need access to first personal data to physically explain its presence?

Q2 and Q3 are specific sub-questions that will naturally be addressed in an inquiry into Q1. I will now turn to an imagined experiment to find out which difficulties physicists will run into when they want to devise experiments to handle these epistemological questions. Afterwards I will provide the answers that I infer from the thought experiment.

The conscious observer observed: a thought experiment Imagine that I want to find out whether I can identify physical roots of human consciousness by experiment. Clearly, the first personal data necessary for my research can only be obtained from a first person's perspective, so, as the observer I am restricted to use my own phenomenal experiences. After all, what applies to me applies to other conscious humans as well: I am the only person who knows what it feels like to be me, and this is exactly the kind of phenomenon I want to study. That is, I want to find out what physical mechanism could give rise to *this* intimate feeling. The limitation on the accessibility of first personal data may at first sight seem to give rise to a principal obstacle: although I have a strong belief in the existence of other minds, the only one I seem to be able to study is my own. However, since I *have* access to my own mental phenomena there appears to be no reason for me not to proceed with the investigation. After all, when I can solve the issue for myself then this sort of introspective research could be done by 'other minds' as well.

Clearly then, the most natural place to start my research is indeed in my own brain. Comparable to Mary in Jackson's seminal paper, I am a researcher who has collected everything there is to be known from physical and neuroscientific facts that could be relevant for the understanding of the dynamics of my own brain.[50] Mary was subjected to the question whether the first personal data related to her experience of seeing red were entailed by everything she already knew from public knowledge. It is my goal to

[50] Jackson 1986, *What Mary didn't know*. Jackson describes vividly a version of the *knowledge argument* against physicalism. His argumentation is based on the observation that hypothetical Mary, who has been locked up her entire life in a black and white room, but who is assumed to know everything there is to know about advanced brain science, and who is therefore considered to be capable to describe everything that goes on in her mind, still learns a new – for her unpredictable – fact about the world when she experiences seeing a red rose for the first time. Thus, the knowledge argument is a variation on the idea that not all facts about reality as we can perceive it are covered by the physical facts. The argument is criticized in Dennett 1991, *Consciousness Explained*.

find out whether this question can answered by experiment, i.e. I am concerned with the issue whether these first personal data *can observably be identified by the physical facts I have access to.* Indeed, I already know that I am acquainted with my first personal data, but can I identify them without this acquaintance, that is, by the public facts that I learned from my intensive studies in brain science, combined with the physical facts that I could discover about my brain? I devise an experimental setup in which I connect some highly advanced scanning devices to my head. When I say 'highly advanced' I mean to suggest that the apparatuses in my setup are capable of registering every physical event in my brain, including individual neuron firings, the dynamics of electrical synapses interactions, energy transports, and so on. The overall setup allows me to observe my brain activities the moment I have subjective experiences.[51]An obvious question now will be: *'What is it that I am actually observing in the brain?'*

To get some insight into this question it is good to acknowledge the difference between my setup and the situation in standard cognitive brain research. In a 'normal' setup a brain researcher will not have direct (private) access to phenomenal data. That is, in a typical neuroscientific research setting a third person's brain activity is observed while he or she executes a task script or is exposed to circumstances that are expected to stimulate specific experiences. Conscious experiences are either assumed to take place or reported by the test person.[52] The essence for now is that in these standard setups data about experiences are only available as public knowledge, i.e. as data in a form that can be reported and shared. The setup in my own experiment is aimed at getting

[51]The imagined setup may be regarded as a simplified example of Feigl's thought experiment with the *'autocerebroscope'.* Meehl refers to this thought experiment as an example of 'empirical metaphysics', an effort to use an imaginary device to find support for the mind-body identity thesis. For an exposition of the autocerebroscope, see Meehl 1966, *The Compleat Autocerebroscopist: A Thought-Experiment on Professor Feigl's Mind-Body Identity Thesis.* In a 1998 interview the late Marvin Minsky expressed his belief that the imaginary device will become real one day: *"In another 20 or 50 years, you'll be able to put a cap on your head that will show what every neuron is doing. (This is Dan Dennett's "autocerebroscope.") [...] Then, for the first time, we'll become capable of some"genuine introspection." For the first time we'll be really self-conscious. Only then will we be able to wean ourselves from dualism."* (Brockman 1998, A Talk With Marvin Minsky [2.26.98] at www.edge.org, accessed on 02-19-2019)

[52]A very recent and advanced example of this kind of research in cognitive neuroscience is to be found in Demertzi et al. 2019. In this study it was "*determined whether dynamic signal coordination provides specific and generalizable patterns pertaining to conscious and unconscious states after brain damage.*" The authors claim that their proposed model "*can account for modes of conscious and unconscious information processing.*" However, the cohort study is based on data coming from brain measurements on both patients in *cognitive* conscious– and *cognitive* unconscious states. The authors' claim that they identified brain patterns relating to consciousness therefore seems to rest on the assumption that first person's awareness of having an experience always accompanies the *cognitive* experience itself. In Chalmers's terminology this kind of research would be concerned with the 'easy problems' of consciousness (David J Chalmers 1995).

round this limitation by the establishment of a direct feedback scenario. This can only be achieved when the observer coincides with the observed conscious mind. There is a direct feedback in the sense that my subjective experience and the related physical brain states are synchronized in the conscious process of observation itself. Recall that it is my aim to observe my brain while it produces the feeling of me being me. Therefore, I presume that the relevant public and private forms reveal themselves to me at the same time. Is this presumed co-occurrence necessary for my research? Yes, it is, as I will now explain.

I observe that I could possibly recall at a later moment what an experience was about. However, the experience itself is not accessible anymore. Or in other words, I cannot be acquainted afterwards with the first personal data I want to observe in my study. To see this, note that a conscious experience consists of our acquaintance with the deployment of phenomenal concepts. In a 'recall' of an experience we may apply the same PCs again. However, the experience we are acquainted with at that later time is the 'experience of recalling' itself. In other words, it is the internal representation of a different event.[53] So, for the inspection of the first personal data I have a single moment available, the moment of acquaintance. But why could I not have a look at the public data at some later time? After all, I could for instance attach a clock to my measurement apparatus. Now, instead of monitoring my brain states I chose to focus myself on the clock. Of course, the case will be different because the measured brain activity is now correlated to my conscious experience of keeping time. But at least I should be able now to decide *afterwards* which brain image is correlated to which former conscious experience. For example, it seems obvious that my conscious observation of the clock pointing at time t *coincided* with the brain image taken at time t. However, this is not necessarily true unless co-occurrence is presumed. Now, the problem is that without this presumption I do not see any possibility to decide afterwards which recorded brain state is a correlate of my current private mental experience. After all, the nature of first personal data forbids me to observe on an arbitrary scan of my brain the contents of my subjective feelings. The co-occurrence of the private and public data concerns the most simple relation in time that is available to connect physical brain states and subjective experiences. Any other form of relation will make my experimental setup useless unless I have an explicit understanding of this relation in both public and first personal data. But I will only have this understanding after I have 'solved the mind-body problem', something which

[53] Of course, it could be that the physical state of my brain looks the same in both cases. However, I have no epistemic means to confirm this. The distinction between the remembrance of awareness and its actual *first-time* operation aligns with the difference between basic- and non-basic applications of phenomenal concepts. See Balog 2009 for more details.

I designed my experiment for in the first place. So, for my experimental setup to make sense at all, the presumption of co-occurrence of correlated private and public data is a necessity.

Now, suppose that I am monitoring my brain's activity and almost at the same time observe a red rose somewhere in my room. There is an awareness accompanying both these events that affirms to me that I consciously experience them. However, I expect that, at the moment I observe my brain activity, the synchronized feedback in my setup ensures me that what I actually observe is the physical correlate of the current conscious event, in this case 'the physical correlate of me having the experience of consciously observing my brain state'. Am I trapped in my experiment because I have to acknowledge that I can only observe brain correlates of experiences of myself observing this same brain correlate? Combined with the fact that only (conscious) observers have direct access to private phenomena, it seems I have run into a profound obstacle: *the only brain states that are available for direct physical observation and of which I know that they are conscious states, are those that apply to the experience of observation itself, i.e. the observation done by the observed.*[54] For a way out of this trap it seems necessary to have a phenomenal experience accessible for a different observer at a different time, that is, through the shift to a third person's perspective. Or to put it differently, I need to be able to translate my first personal data into public data. However, the essence of first personal data is that we cannot make it public for reuse this way. Indeed, as explained above, the only access to an experience is provided through actually having it. Thus, the answer to my question about what I am actually observing must simply be *the publicly accessible brain activity data that are assumed to be a correlate of my experience of the observation itself.*

Although I am left with only experimental access to a single specific state of consciousness, I can still pursue my investigation for this unique case. What can be learned then from the simultaneously accessible public and first personal data? To see this I hypothesize that the unique conscious state can be isolated in the following sense. I assume that before and after the observing experience I can put my brain in a state of deep sleep, i.e. a state in which I am not aware of me being me. I implemented a sophis-

[54] This issue is briefly touched by Hut & Shepard in an 'autocerebroscope-like' example. They claim that "this may lead to Gödelian paradoxes." The authors ascribe this observation to the fact that both mathematics and consciousness can be described *self-reflexively.* I.e., they have in common that they fully rely on internal concepts for the understanding of facts about their domains. As opposed to physics, which depends on mathematics as its language, both mathematics and consciousness are *self-reflexive* in the sense that "math, however, can be directly modeled by math, and consciousness can be directly studied by consciousness." Hut and Shepard 1996, *Turning 'The Hard Problem' Upside Down & Sideways*, p.16

ticated piece of software that allows me to program precisely the intervals of deep sleep. With the aid of this software, which has been coined a *zombie-switch*, I have full control over the length of my conscious moments.[55] So, I start observing my physical brain and the zombie-switch will interfere once and a while by putting me in an unconscious state or pulling me out of it. Now, what I observe during the intervening conscious moments in the simultaneously generated public and first personal data *could* be a perfect correlation between physical brain processes and phenomenal aspects.[56] Indeed, a physicalist will claim that this is the case. He will explain that the 'perfect correlation' is due to the fact that I am observing the same physical phenomena from both a first– and a third person's perspective at the same time. Moreover, there are no first personal data, there is merely an additional perspective for the perception of physical facts.[57] A dualist however would not approve this conclusion. She would tell that I could make an exact physical brain copy based on the observed public data. Doing my experiment with this copy would not make sense because I could not decide whether (the same) first personal data would coexist with the copy of my public data. That is, in an experiment with a physical copy of my own brain I would lose the necessary first person's perspective of observation, and the observed would no longer coincide with the observer. As a result, it is conceivable that a copy of my physical brain states is not accompanied with the mental facts that I perceive as first personal data in my own mind. In fact, she would hold that I cannot even be sure that my zombie-switch will still do the job because I cannot know whether there is 'something that can be switched off'. Clearly, the dualist presents a practical implementation of the zombie-argument.

Such a discussion between a physicalist and a dualist will not bring me any further towards my goal. There merely seems to be a difference of opinion with respect to the *interpretation* of my public brain data, i.e. the data that look the same to both of them. In an effort to get out of this impasse, I ask myself in what way the observed public data and my first personal data must at least be related to allow for observable public facts. In a scientific context where only public knowledge is involved, correlations in the

[55] It was Guido Bacciagaluppi who suggested the term 'zombie-switch' in this respect. The original idea was based on a lab colleague who was able to flip my un-/conscious states.

[56] Note that to make this picture viable I still assume the co-occurrence of correlated public and first personal data. If the public and private facts are not synchronized in time, i.e. when there is a delay involved in the accessibility of one of the two classes of facts, then I would clearly observe brain activity related to unconscious events. The non-synchronicity of physical brain states and *cognitive* events is still highly debated within discussions about free will. For claims about a significant delay of the cognitive state that accompanies a seemingly action of free will, see Libet et al. 1983 and Libet 2006.

[57] This is the physicalist's reply to the knowledge argument. Later in this text at page 37 I will show how Carruthers applies it in a defense of the so-called phenomenal concept strategy. (See also Carruthers and Veillet 2007, *The Phenomenal Concept Strategy*, p.215)

data are supposed to be revealed in a publicly observable causal chain of events.[58] So, to allow my experiment to come up with results I presume at least that the existence of first personal data is efficacious. That is, my personal awareness of having an experience has to have an effect on the observed public data, i.e. on the physical events in my brain. However, a new profound problem reveals itself: *I cannot compare my experimental results with a case in which first personal data are absent.* After all, as argued by my hypothetical dualist, when I investigate a physical copy of my brain I can no longer decide whether phenomenal consciousness is absent or not, because the brain is 'not mine'. That is, I cannot be sure about the absence of consciousness because I have shifted from a first person's perspective to a third person's one, a perspective from which I no longer have simultaneous access to both private and public data. I seem to be trapped: in a first experiment I need to be involved as a conscious observer because I am the sole person who has access to the first personal data that is investigated. At the same time, I cannot do a second experiment in which phenomenal experiences are not involved because therefore I need to decide that there are none, but that can only be decided from a first person's perspective. That is, the non-existence of phenomenal first personal data is itself concerned with a form of inaccessible data. This is why zombies are conceivable. Clearly, the explanatory gap manifests itself full out: I can look at any physical detail in my brain, nowhere will I observe the intimate awareness that accompanies my cognitive states and that I perceive from a first person's perspective *unless I assume it is there.* It seems that the physicalist is doomed to adhere to a form of *strong emergence*: facts concerning consciousness *may* result from the underlying physical facts, but they are not even in principle observably deducible from them.[59]

Aftermath This simplified image of looking at the physics in the brain at least suggests that there are insurmountable difficulties for observing or identifying phenomenal first personal data within an experiment. Physical experiments as we know them are all about public knowledge. Are we asking the impossible from physics when we want to have a

[58] Whether this is the case in all practical circumstances is doubtful. Many physical facts are observed *indirectly* on the basis of a presumed relation with *directly* observable facts. Examples of these are the existence of black holes and the smallest elementary particles. The essence for now is that the *causal closure of physics* is assumed to hold for the current context. Intuitively this principle states that every public physical fact has a physical cause. So, when a private fact *has* a role to play in the physical causal chain of events, then this fact must itself stem from a physical cause. More in-depth discussions about the causal closure of physics and the role it plays as a physicalist's argument against interactive dualism are to be found in Lowe 2000 and Montero 2003. A critical analysis of the principle is found in Elitzur 1989.

[59] For a detailed discussion about strong and weak emergence, see David J Chalmers 2006, *Strong and Weak Emergence.*

physical explanation of consciousness, something which demands direct access to first personal data for the observer? Before contemplating this issue I will turn to my earlier announced answers to the three questions that motivated the thought experiment:

Q1: Can the presence of first personal data be explained or inferred from physical experiments?

A1: *No. Because it is not possible to decide in an experiment whether consciousness is absent, it is not possible to compare experiments with and without simultaneous access to both first personal and public data. As a result, it is not possible to decide whether the experiential aspect itself is efficacious or not. This all results from the fact that only a single* conscious *observer can do an experiment in which private and public data are involved at the same time.*

Q2: Is it possible to observe – without presuming its presence – first personal data in physical experiments?

A2: *No. Only a physicalist will observe first personal data because of her metaphysical belief.*

Q3: Do we need access to first personal data to physically explain their presence?

A3: *Yes. The explanatory gap forbids its inference from public knowledge. The conscious observer herself, being part of the experiment, can confirm the existence of the phenomenal first personal data. However, she cannot physically explain it from the publicly observable physical processes, but she can postulate it in a theory about both private and public knowledge.*

From an epistemological point of view the deployment of empirical research to solve the (true) hard problem of consciousness, i.e. the scientific explanation of the awareness of 'me having this experience', is reserved for the conscious observer who investigates her own experiences. So, when involving consciousness in physical theories, science becomes necessarily introspective. For some this may sound like an 'obvious truth', for others this cannot be because it feels unsatisfactorily that consciousness is fundamental and immune to reduction. I hold that both the conceivability argument and the explanatory gap forbid an explanation of the existence of phenomenal first personal data from publicly

accessible physical facts. In that sense I consider consciousness as fundamental. But this does not imply that I think that consciousness cannot play a role in physical theories. The foregoing discussion reveals that a discussion about the issue whether mental phenomena play a role in physics will essentially be a discussion about the universality of the *causal closure* or *completeness of physics*. The answer to Q3 above suggests that I cannot experimentally identify the causal effects of my first personal data. But philosophy shows that I can use reason to find arguments for postulating them. In this respect, it is illuminating to compare two opposite stands with regard to the knowledge argument.

In a defense of the *phenomenal concept strategy* as a successful argumentation in favor of physicalism, Carruthers explains why Jackson's Mary may not encounter any new physical facts about the world:

> So when she leaves her room, she does acquire the capacity to think some new thoughts (these are thoughts involving phenomenal concepts). Hence she also learns some new facts (in the sense of acquiring some new true thoughts). But for all that the argument shows, these new thoughts might just concern the very same physical facts that she already knew, only differently represented (now represented by means of phenomenal concepts).[60]

The essence of the argument is that the causal chain of physical events is not influenced by the availability of an *alternative representation* of the physical facts. That is, the existence of the first person's perspective does not in itself change the facts as they are perceived from the third person's perspective. In short, Mary's awareness of seeing the red rose has no causal effect on the physical chain that she would describe on the basis of her knowledge. Her new knowledge is merely concerned with a new *representation* of physical facts. An interesting counterargument is put forward by Sir Roger Penrose. He deploys a version of the knowledge argument as a support for the physical efficacy of human consciousness:

> I would contend that the evolutionary development, through natural selection, of the ability to think consciously indicates that consciousness is playing an *active* role and has provided an evolutionary advantage to those possessing it. For various reasons I find it hard to believe that conscious awareness is merely a concomitant of sufficiently complex modes of thinking - and it seems to me clear that consciousness is itself *functional*. [...] Indeed, if consciousness had no operational effect on behaviour, then conscious beings would never voice their puzzlement about the conscious state and would behave just like unconscious mechanisms 'untroubled' by such irrelevancies![61]

[60] Carruthers and Veillet 2007, *The Phenomenal Concept Strategy*, p.215.
[61] Penrose 1987, *Quantum Physics and Conscious Thought*, p.116.

As it seems, the fact that I am writing this text on issues surrounding 'what it is like to be me' must in itself be considered as an observable physical consequence of the private facts about my mind. It is puzzlement about my acquaintance with my conscious experiences that brings me to this writing and to the choice of words I decide to put on paper. So, according to this line of reasoning, having an experience will produce new physical facts. These words by Penrose are also cited by Elitzur, who recognizes in them a severe challenge to the 'passivity dogma'.[62] He translates Penrose's words into what he calls the 'Bafflement Argument' against the 'Qualia Inaction Postulate': "The fact that humans are baffled by the Percepts-Qualia Nonidentity, and express this bafflement by their observable behavior, is a case where qualia per se - as nonidentical with percepts - play a causal role in a physical process."[63] And he extends the argument with a profound consequence for physics: "The first thesis is that consciousness is not passive but rather a part of the causation on behavior. The second and consequent thesis is that physics, being unable to describe consciousness, is inherently incomplete."[64] Now, Elitzur claims that the Bafflement Argument can be used to establish a testable case for the physicalist-dualist dispute, thereby turning the debate into an empirical issue. This claim rests on his observation that physicalists will understand the bafflement as due to false beliefs. As a result, bafflement must be explained by the physicalist on physical grounds: "When future neurophysiology becomes advanced enough to point out the neural correlates of false beliefs, a specific correlate of this kind would be found to underlie the bafflement about qualia."[65] The Bafflement Argument itself leads to the opposite prediction: "No neural pattern underlying a false belief will be found to underlie adherence to dualism."[66] Without going into the details of the reasoning, Elitzur contends to have a proof that the physicalist's view that bafflement rests on 'misperception' must be false. He calls it the *Asymmetry Proof*: "If a quale is identical with its percept, then its appearance as nonidentical must be due to misperception. But misperception, being a special kind of perception, occurs in accordance with physical law. Hence, upon reflection, it must turn out to be obligatory. Qualia, in contrast, can be conceived of as altogether absent."[67] Although Elitzur claims this proof to be new, I hold that this approach concerns a variation of the zombie-argument about conceivability. Why then putting forward this argumentation by Elitzur? The reason is that it exemplifies the

[62]Elitzur 1989, *Consciousness and the Incompleteness of the Physical Explanation of Behavior*, p.10.
[63]Elitzur 2009, *Consciousness Makes a Difference: A Reluctant Dualist's Confession*, p.14.
[64]Elitzur 1989, p.10.
[65]Elitzur 2009, *Consciousness Makes a Difference: A Reluctant Dualist's Confession*, p.15.
[66]Ibid.
[67]Ibid.

difficulties involved with respect to first personal data. First, his claim that the dispute is transformed into an empirical case is not justified. As illustrated by the thought experiment above, 'a future neurophysiology would not point at the neural correlates of false beliefs, *unless one assumes it does.*' Indeed, for the identification of bafflement a first person's perspective is needed. But again, the nature of experiments where first personal data are involved forbids a comparison with cases where these data are absent. So, Elitzur's claim about a testable hypothesis will bring him into the same difficulties that come with the autocerebroscope. Then secondly, the physicalist will again object that bafflement is not due to misperception, but rather stems from a different representation of physical facts, namely, a view from a first person's perspective. And of course, finally, the physicalist will contend that he saved the completeness of physics.

In conclusion, the trouble with the very nature of first personal data is that it sabotages every experimental effort to explain 'why it is like this to be me'. The demand for immediate epistemic access to compare public facts and to correlate them with unique first personal data frustrates every effort to a successful identification of mental phenomena in a physical experiment. Therefore, the first candidate requirement for a physics that includes consciousness is doomed to lead us nowhere. That is to say, when we make the observer part of the observed we will still not be able to decide by experiment whether she is conscious or not. In fact, we could put a zombie in as well. The thought experiment reveals that both physicalists and dualists can maintain their own interpretations of what it observed in such an experiment. The conclusion must be that the available philosophical views are not threatened by a possible experimental confirmation of the presence or absence of consciousness.[68] Therefore, it seems that the first requirement for a physics that entails consciousness must be abandoned: the explanatory gap forbids an explanation of how consciousness arises in an observer. In the second part of this text I will explore the options for the second candidate requirement. That is to say, I will try to find out how consciousness as a fundamental phenomenon can be assigned a role in a coherent physical theory.

[68] It seems to me that the physicalist's option naturally must be based on the idea of strong emergence. The autocerebroscope thought experiment leaves no room for weak emergence, i.e. the first personal data will even not *eventually* be explained from public facts.

Part II

Solving the Hard Problem of Physics

Chapter 3

The conscious observer

> *What is needed is something we do not have: a theory of conscious organisms as physical systems composed of chemical elements and occupying space, which also have an individual perspective on the world, and in some cases a capacity for self-awareness as well.*
>
> - Thomas Nagel, *A View from Nowhere, 1958*

Recall that I proposed two slightly different candidate requirements for the transition that is needed in physics to include consciousness in its domain (see page 27, *Re1* and *Req2*). The second of these demands that a physical theory of consciousness must be able to address the role of the consciousness of the observer in an observation. Stated this way, the hard problem briefly comes down to the fact that the *consciousness of the observer is involved in the act of observation*, while at the same time this role for consciousness is the explanandum. This chapter is concerned with consciousness as seen from the perspective of physical observations.

As I already mentioned before, it is hard to imagine what to think about *physics* without a role for a conscious observer. After all, what we expect from physics is a description of the world, as good as it gets, which can be consumed in the conscious mind of an observer. Of course, such a formulation already begs for clarification: 'Why should the observer be conscious?' Could non-conscious arbitrary physical objects not act as 'observers' that likewise digest facts that are provided through physical interactions with the outside world? Not quite. It is hard to see how a rock will act as an observer in the sense that it observes and interprets physical processes. Or to put it differently, we do not feel that there is an explanatory gap in this case. Intuitively we think of an observer as an entity that has both *intentions* and a *'feel' that accompanies an internal presentation*

of facts. That is, it deploys an intention when it initiates an act of observation, and it is acquainted to an internal representation of facts that are obtained through the act of observation.[69] But wait, could an unconscious artificial cognitive system not act as an observer then? Again, when I apply my intuitive idea of an observer, then the answer is no. Surely, it is conceivable that intentions can be programmed as seemingly intentional actions. But because the system is unconscious it lacks the acquaintance to an internal mode of representation. Or, in the terminology from the former section, it lacks first personal data because it is unconscious by definition. Therefore, I take it as a fact that there is a natural role for a conscious observer in physics from the start: a world without consciousness would be a world without physics, a world without *self-interpretation*. Or to put it differently, there is no physics in a zombie-world! Such a world would be a world without first personal data, a world without any first person's perspective on its facts. After all, this is how zombies must be understood in terms of the conceivability arguments. The essence of physics is that it provides its participants with input for an internal representation of knowledge.[70]

3.1 Classical physics

Thus, I hold that the conscious observer always has a role to play in physics, but what is this role? The sketch I gave in the former paragraph was rather intuitive, but it appeals to an observation that was already common in classical physics or, as Jammer puts it:

> Classical physics described physical reality as composed of entities devoid of sensuous qualities (extended bodies moving in space or fields); but the theory achieved its validity only by virtue of the fact that its predictions could be tested, an operation which, in the last analysis, had to involve human consciousness.[71]

It seems that a recognition of the role of human consciousness in the validation of theories could have justified a different treatment of it. However, because physics could deal very well with ordinary objects there may have been a natural inclination to by-pass the less appealing topic of consciousness. Shimony puts it this way:

[69] Obviously, these facts are assumed to come from an outside world, but all it can be really certain about is only the internal representation itself.

[70] It must be noted that I refer to physics when I think of the *activity* of studying and describing physical processes and physical facts. Although *doing physics* is precluded in the zombie-world, there is of course a place for physical facts in such a world.

[71] Jammer 1974, *Philosophy of Quantum Mechanics, The Interpretations of Quantum Mechanics in Historical Perspective*, p.471.

> Thus a relation between the physical *Weltbild* and experience could be established, even though the process whereby the observer performed his act of recognition was very obscure. [...] The "bifurcation of nature" – as Whitehead named the extreme separation of physical and mental entities in nature – may have been a scandal from the standpoint of metaphysics, but it became a convenient working arrangement from the standpoint of physical theory.[72]

So, the role of consciousness in classical physics was – perhaps on pragmatic grounds – restricted to the confirmation of theories in which it had no place itself. Indeed, as Jammer points out, although ontological and epistemological issues with respect to the relationship between physical objects and human consciousness "seemed to be involved", there was an approach to measurement that made it possible to ignore them completely. To see this, it is necessary to understand the classical interpretation of physical measurement.[73]

On page 20 I referred to the idea of the 'object-subject split' in classical physics, i.e. the view that the world can be partitioned into an observer and observed parts. Although the complementary partitions are all subject to the same laws of physics, in the classical view it is assumed that these laws can be understood as devoid of any role for a conscious observer. That is to say, the observed part of the world can be described without the need of the presence of an observer, i.e. a description that applies solely to the unconscious observed part. And even more importantly, the observer can have knowledge about this description. How then could an observer obtain this knowledge? Clearly, this is where the act of *measurement* comes in. A classical physical observation comes down to an interaction of an observed system with a measuring device, plus an interaction of the latter with the conscious observer. So, in classical physics it was already acknowledged that a full observation should therefore necessarily involve a physical measurement I_1

$$I_1 = I_{S \leftrightarrow M} \tag{3.1}$$

between an observed system S and a measuring device M, and a *psychophysical* interaction I_2

$$I_2 = I_{M \leftrightarrow C} \tag{3.2}$$

between the measuring device M and the observer C.
A crucial assumption in classical physics was that the order of magnitude of the influence of M on S, i.e. the effect of $I_{M \to S}$, was so small in comparison with the impact of $I_{S \to M}$

[72] Shimony 1963, *Role of the Observer in Quantum Theory*, p.755-756.
[73] The explanation along these lines is provided by Jammer 1974, p.471-472

that it could be neglected. In fact, the impact of $I_{S\to M}$, for example a result in the form of a pointer position on a measuring device, was the main concern of the observation. It was further assumed that I_2 was "extraneous to physical theory". These two assumptions allowed the freedom "to 'objectivize' classical physics, that is, to treat its processes as independent of observation and to ignore the role of the observer."[74]

Together with the deterministic character of the classical physical laws, this objectivization of physics made it possible in principle to describe the true state of a system at any moment in time. That is, the objectivization allowed for the discovery of observation-independent initial values, and the deterministic laws fixed future and past states on the basis of these initial conditions.

3.2 Consciousness and wave function collapse

In quantum mechanics, i.e. in the physics of atomic processes, equation (3.1) can no longer be interpreted in the classical way. Or as Bohr used to put it, the finite size of Planck's constant h forbids a neglect of the impact of $I_{M\to S}$ in comparison with $I_{S\to M}$. Without this neglect the ignorance of the role of the observer is no longer justified. I.e., the classical strategy of objectivization must be dismissed: quantum mechanics introduced a *non-separability of observed and observer*. Although Bohr himself never talked in terms of the consciousness of the observer, a follow-up question could be whether the observer's consciousness should still be regarded as foreign to physical theories. In other words, is there a place for consciousness within an act of observation? I think that with the rise of quantum mechanics this question has become highly relevant. Why? From the early days of quantum physics it was recognized that the new theory of physics had to have a profound impact on the interpretation of experiments. There was an epistemological aspect involved from the start. Already in the early days of the Copenhagen Interpretation Bohr, Heisenberg, and Born, amongst others, were concerned with the issue of what measurements actually inform the observer about.[75] For Bohr comple-

[74] Jammer 1974, p.472

[75] There exists an extensive literature on the history of the Copenhagen Interpretation. It is generally agreed that the name does not refer in an unambiguous way to what quantum theory is about. Howard 2004 holds that the term Copenhagen Interpretation was coined by Heisenberg in the 1950s. Its connotation is linked to a positivist's view in which Bohr and Heisenberg were assumed to play central roles. However, from 1926 onwards Heisenberg and Bohr strongly disagreed with respect to crucial aspects of the theory. Howard observes that *" [...] what is called the Copenhagen interpretation corresponds only in part to Bohr's view, here termed the complementarity interpretation. Most importantly, Bohr's complementarity interpretation makes no mention of wave packet collapse or any of the other silliness that follows therefrom, such as a privileged role for the subjective consciousness of the observer. Bohr was also in no way a positivist."*ibid., p.669 I will not go into details of the fuzzy understanding of the

mentarity played a crucial role. At the atomic scale we are confronted with phenomena that we cannot know about without the application of an experimental context. Observable entities that demand mutually exclusive contexts for the measurement of their values are complementary in the sense that they cannot be investigated within the same experimental setup. A paradigm example concerns the impossibility of getting knowledge about an electron's position and momentum within a single measurement. So, complementarity holds that the observer cannot know all aspects of reality at once. However, the knowledge that can be obtained through all the mutually exclusive measurements " [...] exhausts all possible objective knowledge of the object."[76] Thus, quantum theory tells us in this respect what knowledge the observer could obtain by measurement.

3.2.1 Orthodox quantum theory and the measurement problem

Before turning to the role of the observer, and consequently to the act of measurement, it is good to summarize what the (mostly) undisputed part of orthodox quantum theory is about. The central feature of a system from the perspective of quantum mechanics is its *physical state*. In classical mechanics the state of a system of N particles with D degrees of freedom at time t can be identified by a point in a $2N \times D$-dimensional *phase-space*. E.g., if the positions and momenta of the individual particles are $\mathbf{q}$ and $\mathbf{p}$, then the phase-space Ω is defined as

$$\Omega \subset \mathbb{R}^{6N} = \{\mathbf{x} = (\mathbf{q}, \mathbf{p}) \mid \mathbf{p} \in \mathbb{R}^{3N} \ , \ \mathbf{q} \in \mathbb{R}^{3N}\} \tag{3.3}$$

The deterministic equations of motion of classical mechanics describe the evolution of the system's state as a trajectory of a point $\mathbf{x}$ through phase-space. The state of the system is now fully specified. In quantum mechanics the state of a system S is specified by a vector $|\Psi\rangle$ in a linear vector space, i.e. a Hilbert space $\mathcal{H}$. Now, the mathematical framework of quantum theory provides 'algorithms' for answering two physical questions about S:[77]

1. *quantization algorithm*
   How do we get from classical observables to observables in quantum mechanics, and which possible values for an observable Q in quantum mechanics can be obtained upon measurement?

Copenhagen Interpretation. More on the issue is to be found in Faye 2014

[76] Ibid., p.16.

[77] For further details about different formulations of these algorithms (e.g. Dirac, von Neumann), see Redhead 1987, *Incompleteness, Nonlocality, and Realism*

2. *statistical algorithm*
   Given a state of system S, what is the probability of yielding a value q when measuring an observable Q?

An important feature of the theory is that the superposition principle holds for state descriptions. I.e., If $|\psi_1\rangle$ and $|\psi_2\rangle$ are possible states of a system S, then so is

$$|\psi\rangle = \alpha\,|\psi_1\rangle + \beta\,|\psi_2\rangle \ \text{, with } \alpha, \beta \in \mathbb{C} \tag{3.4}$$

The state description or *wave function* $|\Psi\rangle$ of a system S contains all information about physical quantities of S that can be known through experiment. These quantities – or *observables* – are identified with operators Q on the Hilbert space $\mathcal{H}$. The quantization algorithm ensures that a measurement of an observable Q in a finite-dimensional Hilbert space $\mathcal{H}$ will yield one of the *eigenvalues* $q_1, q_2, q_3, ..q_n$ belonging to the *eigenvectors* or *eigenstates* of Q.[78] The set of eigenstates is complete in the sense that a set of eigenvectors $|q_1\rangle, |q_2\rangle, |q_3\rangle, .. |q_n\rangle$ belonging to Q can be chosen in such a way that they form an orthonormal basis of $\mathcal{H}$, i.e. every state $|\Psi\rangle$ can be expressed as a linear combination of these vectors $|q_i\rangle$. Furthermore, a measurement of Q yields the value q_i if and only if the state $|\Psi\rangle$ is an eigenstate (with corresponding eigenvalue q_i) of the system S (the *eigenvalue-eigenstate link*). Generally, if the state $|\Psi\rangle$ of S is

$$|\Psi\rangle = \sum_{i=1}^{N} c_i\,|q_i\rangle\,, \tag{3.5}$$

then the statistical algorithm, which is based on Born's rule, gives (in the discrete non-degenerate case) for the probability that a measurement of Q yields the value q_i

$$Prob(q_i)_Q^{|\Psi\rangle} = |c_j|^2 = |\,\langle q_j|\Psi\rangle\,|^2 \tag{3.6}$$

Finally, the time-evolution of the undisturbed system S in state $|\Psi\rangle$ at $t = t_0$ is given by the time-dependent *Schrödinger-equation*:

$$i\hbar\frac{\partial\,|\Psi(t)\rangle}{\partial t} = \hat{H}\,|\Psi(t)\rangle\,, \tag{3.7}$$

in which H is the energy operator or the Hamiltonian of the system. The linear Schrödinger-equation describes a deterministic law that governs a reversible process and

[78] In this brief overview I do not consider notions like *degeneracy* or *infinite-dimensional* Hilbert spaces. These issues do not play a role in the current context of discussion.

that may be considered as the counterpart of Newton's classical laws of motion. From (3.7) it follows that the undisturbed system S will evolve as

$$|\Psi(t)\rangle = e^{(i/\hbar)H(t-t_0)}.|\Psi(t_0)\rangle \tag{3.8}$$

Note that in general $|\Psi(t)\rangle$ and $|\Psi(t_0)\rangle$ will assign different probabilities to the values q_i. This is why one refers to the state's dynamical evolution in time.

This brief exposition summarizes the main ingredients of quantum theory with respect to the description of physical states of undisturbed systems. But physics is about observation through measurement. What the rules above do not state is what happens when a measurement of Q is performed on a system S that is not in one of the eigenstates connected with Q. Before turning to von Neumann's account of measurement and the role of the (conscious) observer therein, suppose a system S is not in an eigenstate of the observable Q, but has rather evolved according to (3.8) into a (pure) state ψ of the form (3.4). This superposition then introduces an indefiniteness of the value of Q, something which is common for observables like position and momentum in the realm of atomic physics. On the other hand, we observe that macroscopic objects are constituted of small atoms that obey the laws of quantum mechanics. Therefore, we must infer that everyday objects like trees and rocks in fact should be regarded as large composite quantum systems. However, as conscious observers of the everyday world around us we do not perceive any superposition. Rather, we observe that trees and rocks have definite positions. In a nutshell, the fact that the description of quantum theory given above cannot account for our everyday observations is the basis of the notorious *measurement problem*. In measurement terms the problem comes down to the requirement that measurements yield definite results. This requirement is referred to as the *objectivation condition*.[79] Butterfield argues that the measurement problem in quantum mechanics has a natural relation to the role of human experience, and consequently to the role of the human mind:

> Any description of the world that someone advocates as being complete, [...] must 'close the circle': it must include an account of how we come by that description. In particular, any physical theory that claims such completeness must account for our experience as observers.[80]

He connects this observation to the measurement problem by approaching the latter as a problem of experience:

[79] Busch, Lahti, and Mittelstaedt 1996, *The Quantum Theory of Measurement*, p.73.

[80] Butterfield and Fleming 1995, *Quantum Theory and the Mind*. Butterfield paraphrases Shimony where he speaks of 'close the circle'.

> But tables and chairs surely have definite positions etc.; at least, we experience them as doing so. So, if QT is to account for our experience, it must either secure such definiteness, or at least explain the appearance of it.[80]

In fact, Butterfield's observation is a crucial part of his defence of the relevance of the mind-matter relation to the interpretation of quantum theory. Stapp claims that quantum theory can be formulated in such a way that it "reduces the problem of quantum measurement to the problem of the dynamical connection of mind to brain", a perception that obviously brings us very close to the topic of this text.[81] As I defended earlier, I hold that physics is ultimately about conscious observation. Therefore, a theory that strives after completeness must indeed account for the subjective representation of its facts.

3.2.2 Von Neumann's chain

It is time to extend the limited quantum machinery sketched thus far by the introduction of von Neumann's *projection postulate* or the *collapse of the wave function*. The mathematical foundations of quantum theory, with the inclusion of the act of measurement, were laid down by von Neumann in 1932.[82][83] Von Neumann identifies two separate processes involved in the evolution of a system, which he refers to as *'process 1'* and *'process 2'*. Process 2 is the standard deterministic evolution that is governed by the Schrödinger equation. It is the unitary evolution that follows (3.8) and it describes the system when it is undisturbed, i.e. as long as no measurement is performed. Now, the problem is that process 2 alone suggests an incomplete picture since it does not seem to provide the definiteness we experience around us or in experiments. In fact, as Jammer points out, " [...] if the whole physical universe were composed only of microphysical entities, as it should be according to the atomic theory, it would be a universe of evolving potentialities (time-dependent ψ-functions) but not real events."[84] What is additionally needed to account for the world as we perceive it, is *either a process or an interpretation, that ensures that it looks to the observer as if the system's state was reduced to one of its eigenstates*. After all, in such an eigenstate the eigenvalue-eigenstate link tells us that one or more observable values will come up with probability 1 and we will not observe

[81] Stapp 1995, *The Integration of Mind into Physics*, p.5.
[82] Von Neumann 1955, *Mathematical Foundations of Quantum Mechanics*.
[83] In 1933, one year after von Neumann, Pauli also published a theory of measurement. His analysis of the *measurement problem* was much alike von Neumann's. (Busch, Lahti, and Mittelstaedt 1996, *The Quantum Theory of Measurement*, p.100)
[84] Jammer 1974, p.474.

any superposition. Whether the system's state really is reduced to one of its eigenstates is open for discussion.

In the quantum framework established by von Neumann reduction really takes place. Besides process 2 he introduces what he calls process 1, i.e. a wave function collapse as an additional ingredient in the evolution of states. With the introduction of processes 1 and 2 von Neumann distinguishes "two fundamentally different types of interventions": "the arbitrary changes by measurements" and "the automatic changes which occur with passage of time".[85] In doing so he extends the orthodox interpretation into a theory of measurement. Letting the technical details of the non-causality and irreversibility of process 1 aside, it is for the discussion about conscious observation highly relevant to have a closer look at von Neumann's interpretation of this process. Shimony refers to this process as a 'transition of type 1'. Seven years after the introduction by von Neumann it was also described by London and Bauer, who remarked: "A measurement is achieved only when the position of the pointer has been *observed*. [...] We note the essential role played by the consciousness of the observer in this transition from the mixture to the pure case. Without his effective intervention he would never obtain a new ψ function."[86] Shimony observes that both von Neumann and London and Bauer regard type 1 transitions as discontinuous steps that are due to the act of measurement. Furthermore, all of them presume that, as expressed in the words of London and Bauer, measurement is the "registration of the result in a consciousness".[87] Indeed, it appears that their interpretation of process 1 actually puts human consciousness explicitly into quantum theory.[88] To see in what way consciousness enters the theory, it is important to acknowledge von Neumann's perception of the place of the *physical* observer in the world:

> The theory of the measurement is a statement concerning $S+M$, and should describe how the state of S is related to certain properties of the state of M (namely, the positions of a certain pointer, since the observer reads these). Moreover, it is rather arbitrary whether or not one includes the observer in M, and replaces the relation between the S state and the pointer positions in M by the relations of this state and the chemical changes in the observer's eye or even in his brain (i.e., to that which

[85] Von Neumann 1955, V. p.351.
[86] London and Bauer 1983, *The Theory of Observation Quantum Mechanics*, p.251.
[87] Shimony 1963, *Role of the Observer in Quantum Theory*, p.757.
[88] This is a point where he certainly deviates from Heisenberg's later reading of the Copenhagen Interpretation. In 1958 Heisenberg writes: *"Certainly quantum theory does not contain genuine subjective features, it does not introduce the mind of the physicist as a part of the atomic event."* Heisenberg 1958, *The Copenhagen Interpretation of Quantum Theory* in *Physics and Philosophy*,. p.23

> he has "seen" or "perceived").[89]

Von Neumann seems to put the measurement device and the physical parts of the observer on a par. The 'arbitrariness' regarding the inclusion of the observer in M is important and rests on the interpretation of what is often referred to as the *von Neumann chain*.[90] To get the idea of this chain, suppose the measurement of an observable O on a system S can yield two values, '0' and '1'. The pointer of a measuring apparatus M will, after a well-performed measurement of O, point at one of these two values. Initially the apparatus and S are not in contact. When S starts in one of its eigenstates belonging to O, say the state that will yield value '1' with probability 1, then interaction between S and M will lead to

$$|1\rangle_S |ready\rangle_M \xrightarrow{\text{interaction}} |1\rangle_S |\text{`1'}\rangle_M \tag{3.9}$$

However, if S starts out in a superposition of the two eigenstates then things look rather different:

$$(\alpha |1\rangle_S + \beta |0\rangle_S) |ready\rangle_M \xrightarrow{\text{interaction}} \alpha |1\rangle_S |\text{`1'}\rangle_M + \beta |0\rangle_S |\text{`0'}\rangle_M \tag{3.10}$$

The combined system $S + M$ is in a superposition and it takes a reduction process, i.e. a process 1, to get definite values for O and the pointer position of M. According to the earlier presumption this requires a registration of the result in consciousness.[91] Why? This is where von Neumann's 'chain' comes in. Suppose that we add an additional measuring apparatus M' to the system $S + M$. Then the following will happen during interaction:

$$(\alpha |1\rangle_S |\text{`1'}\rangle_M + \beta |0\rangle_S |\text{`0'}\rangle_M) |ready\rangle_{M'} \xrightarrow{\text{interaction}} \alpha |1\rangle_S |\text{`1'}\rangle_M |\text{`1'}\rangle_{M'} + \beta |0\rangle_S |\text{`0'}\rangle_M |\text{`0'}\rangle_{M'} \tag{3.11}$$

Thus, extending the chain with intermediate measuring apparatuses will lead to an increasing chain of superpositions that is only known to end with the conscious registration of an observer. Or, as von Neumann puts it:

> [...] no matter how far we calculate – to the mercury vessel, to the scale of the thermometer, to the retina, or into the brain, at some time we must say: and this is perceived by the observer. That is, we must always divide the world into two parts,

[89] Von Neumann 1955, V. p.352.
[90] d'Espagnat 2013, *On Physics and Philosophy*, p.111.
[91] Shimony 1963, *Role of the Observer in Quantum Theory*, p.758.

> the one being the observed system, the other the observer.[92]

The subjective experience itself is foreign to the physical environment:

> [...] the subjective perception is a new entity relative to the physical environment and is not reducible to the latter. [...] subjective perception leads us into the intellectual inner life of the individual, which is extra-observational by its very nature (since it must be taken for granted by any conceivable observation or experiment).[93]

To connect this subjective perception with the observer a crucial role must be played by the *principle of psycho-physical parallelism*. He applies it to pave the way for some form of dualism:

> [...] it is a fundamental requirement of the scientific viewpoint – the so-called principle of the psycho-physical parallelism – that it must be possible so to describe the extra-physical process of the subjective perception as if it were in reality in the physical world – i.e., to assign to its parts equivalent physical processes in the objective environment, in ordinary space.[94]

In other words, what is subjectively perceived must coincide with the outcome of a physical process 1. To avoid a violation of the psycho-physical parallelism principle von Neumann shows that process 1 can be applied anywhere in the sequence $\sum_i M_i$ of measurement devices, without altering the observed values for the observables of S.

What we have seen in this brief exposition is that von Neumann divides the physical world into an observed part and an observer. Crucially, the observer is constituted of an arbitrary (physical) subset of a von Neumann measuring chain, plus an extra-physical part that is responsible for subjective perception. This latter part, consciousness, registers the outcome of a state reduction when a process 1 is applied. A few things must be noted. First, von Neumann is silent on the question of what initiates a process 1. An act of measurement may seem a voluntary act from the side of the observer, but nowhere he explicitly assigns an active role to consciousness in the establishment of a wave function reduction. The non-conscious part of the observer clearly (physically) interacts with its environment, which leads to an update of the ψ-function. What then triggers the overall compound system of $S + \sum_i M_i$ to a wave function collapse and to an update of the knowledge on the subjective side?[95] Secondly, and noticeably left out in von Neumann's analysis, there is the issue of *intersubjective agreement*: if a conscious

[92] Von Neumann 1955, p.419-420.
[93] Ibid., V.I. p.418.
[94] Ibid., V.I. p.419.
[95] Although not fully worked out, London and Bauer are a bit more specific in this respect. Shimony recognizes " [...] some important, though incompletely developed, propositions regarding the place

mind is responsible for the establishment of definite measurement outcomes, why then should two distinct observers agree on their results? It is remarkable that von Neumann left the issue as an 'exercise to the reader'.[96] There are two options available, each of them problematic in its own respect. The first option is that two observers both independently register the same reduction. But then their agreement can only be based on something like a "pre-established harmony".[97] Obviously, this option will leave us with a huge mysterious explanatory gap. The other option is that only one observer is truly responsible for the reduction, thereby leaving a reduced wave function as a 'no-choice option' for the other.[98] According to Shimony this option leaves us with the unpleasant position of a "single ultimate subject". I.e., he holds that this option must imply that only a single mind has the ability to reduce superpositions.[99] I do not think that this is a necessary conclusion because the reductive powers could still be distributed among multiple subjects who act according a 'first-comes-first' principle. However, as rightly pointed out by Shimony, this leaves difficulties with regard to causal relations between the observations that are at odds with relativity principles. Putting the intersubjectivity agreement aside, the idea of considering von Neumann's subjective perception on a par with a 'mind' seems to me problematic from the start, something which brings me to the third point of notice: How must we understand in von Neumann's picture the special compositional character of our physical brain? Or to put it differently, why do we have the feeling that the reduction that is registered by the subjective 'I' operates on a *chain that is terminated by the physical constituents of the human brain*? After all, if the subject has the ability to reduce a superposition, then we could imagine that it reduces superpositions that do not entail the state of human brain cells. Even more, we

---

of the mind in nature." (Shimony 1963, *Role of the Observer in Quantum Theory*, p.759) L&B do not regard the conscious observer as extra-physical, but rather they assign him the "faculty of introspection". (London and Bauer 1983, *The Theory of Observation in Quantum Mechanics*, p.252) This provides the observer the possibility to "keep track from moment to moment of his own state". And "By virtue of this "immanent knowledge" he attributes to himself the right to create his own objectivity – that is, to cut the chain of statistical correlations [...]" The authors regard this as "creative action", something which they label as "making objective". (ibid.) It must be noted that L&B suggest that the conscious mind itself can be in a state of superposition before actual reduction occurs. For a further discussion of this view, see Shimony 1963, pp.759-763

[96] Von Neumann 1955, p.445.

[97] Shimony 1963, p.767.

[98] When Schrödingers's famous cat consciously observes what happens in the box, then the physical superposition of 'dead and alive' is reduced and the rest of the conscious world is left with a no-choice option. The problem that is left concerns the question of what happens when no such observation takes place inside the box.

[99] Shimony (Shimony 1963, p.767) points out that this conclusion is in agreement with Wigner's analysis of the thought-experiment with his 'famous friend'. (Wigner 1995, *Remarks on the Mind-Body Question*)

could think of a subject reducing the state in which another brain is involved. Clearly, this implication sounds weird, but the essence holds: if there is a correlation between the brain and the mind, then von Neumann's chain can not be an arbitrary one. I.e., human subjective perception is related to a process 1 in which the brain is involved. As a final point I notice that it is unclear how we precisely must understand the notion of subjective perception. Von Neumann seems only to refer to psychological aspects or our 'inner-life', but from the perspective of part I of this text we should be able to add some more to this point. Generally, before the registration of a measurement outcome, the whole chain is in a superposition with indefinite values of observables. Now, recall that I talked of physical facts as seen from a first person's perspective, the first personal data that accompany conscious experiences. So, the question is how we can deploy the first person's perspective in von Neumann's picture, i.e. how can we interpret subjective perception when the terminology I used in the first part is applied? Let's find out. When a process 1 has occured, then there is an internal representation of facts that were not actual before this representation. In fact, because of the character of process 1, representing first personal data coincides with the actualization of the physical facts they represent! Or in other words, without conscious representation there are no determinate physical facts. This seems like a curious observation, because it suggests that consciousness actualizes the world into what it seems to us. Whether there is truly an active role for consciousness in this sense remains unclear, but we must at least conclude that consciousness is postulated as a fundamental non-physical entity in the view of von Neumann. Of course, he did not have a full-blown theory about human consciousness in mind when he laid down the mathematical framework for orthodox quantum mechanics. But Stapp observes in von Neumann's analysis of the measurement chain an important suggestion to work out such a theory.

3.2.3 Stapp's mindful universe

In the previous sections I sketched how consciousness – in the form of an aspect of the observer – entered physics as the source of wave function collapse. Now it is time to have a closer look at an actual theory of the mind that elaborates on this role. Henry Stapp proposes such a full theory of consciousness, i.e. a view that is essentially based on the role of the quantum wave reduction postulate. To see how everything I discussed in part 1 fits in such a theory, I will go a bit deeper into this theory, a theory that is basically an extension of the ideas proposed by von Neumann in 1932. The purpose of the following exposition is to find out if and how quantum state reduction can be

related to the 'hard problem for physics', i.e. to find out whether it can provide a reasonable footing for putting the consciousness of the observer, and thereby the related first person's perspective, into a consistent physical theory. Before I go into an analysis from the perspectives I sketched in part I, I will first emphasize the motivations behind the theory and set out its main structure.

Philosophical basis

Over the last three decades Henry Stapp has written many articles about his quantum theory of the mind. Taken together these texts reveal a gradual development in the form of many refined ideas surrounding a central fixed core. Sometimes there is a shift in focus, for instance from the role of Heisenberg's ontology and his idea of *res potentia* towards the role of choice in the universe, and from the central position of von Neumann's process 1 to the role of decoherence and the quantum Zeno effect. Still, however extensive his writings, it is possible to grasp a coherent picture of Stapp's ideas. To understand Stapp's theory about the mind, it's helpful to acknowledge his philosophical point of departure. A correct appraisal of the implicit context of his theory appears inevitable for understanding his arguments. I.e., his perception of orthodox quantum physics seems to align very well with his philosophical basis. Indeed, for example both in *Mind, Matter and Quantum Mechanics* and in *Mindful Universe*,[100] two recent books that assemble many of his earlier writings, Stapp balances often between the philosophy of William James and the psycho-physical approach to quantum physics by von Neumann, thereby establishing links for motivation, justification, and confirmation of his own ideas. One might say that Henry Stapp stands in a particular tradition of physical thought. His writings often refer to the 'founding fathers' of quantum physics, i.e. to physicists like Heisenberg, Bohr, Pauli, Dirac, Wigner, and von Neumann.[101] On the philosophical side he is especially indebted to William James and Alfred North Whitehead. In the preface of his book *Mind, Matter and Quantum Mechanics* Stapp calls his proposed solution the *Heisenberg/James model*, because it "unifies Werner Heisenberg's conception of matter with William James's idea of mind". Despite the fact that James had to develop his conceptions about the human mind against the background of classical physics, there are many passages in his *The Principles of Psychology* that seem to beg for physical principles only available after the rise of quantum physics.[102] And this is exactly why

[100] Stapp 2009, *Mind, Matter and Quantum Mechanics*, Stapp 2011, *Mindful Universe*

[101] Perhaps not totally unexpectedly, since, in the 1950s as a theoretical physicist he worked closely together with influential people like Wolfgang Pauli, Werner Heisenberg, and John Archibald Wheeler.

[102] Although the 19$^{\text{th}}$-century philosopher and psychologist William James published his monumental *The Principles of Psychology* in 1890 (James 1890), his influence on contemporary psychology endures.

Stapp links the perspectives of James with the concepts of orthodox quantum mechanics. His relation to James's work is saliently summarized in the following remark in the first chapter of *Mind, Matter and Quantum Mechanics*: "The main conclusion of the present work is that James's ideas about mind and its connection to brain accord beautifully with the contemporary laws of physics."[103]

Linking James and von Neumann

For a true understanding of the role of James's ideas within Stapp's framework, it is crucial to identify their most important elements, which have – according to Stapp – surprising, yet unrecognized, counterparts in von Neumann's formulation of orthodox quantum mechanics. A thorough analysis of Stapp's account of James's insights shows how he applies quantum theory to 'implement James's program', i.e. how he essentially aims to demonstrate to which extent the orthodox interpretation provides solutions for the problems of consciousness that were already identified by James in the context of classical physics. A major observation made by James was that the mind was wrongfully left out in natural science and, according to Stapp, rehabilitation necessarily demanded a new physics: " [...] the problem of the connection of conscious process to brain process was irresolvable within the framework of the classical physics of his day. He foresaw, accordingly, important changes in physics."[104] So, the new physics should account for essential claims of James's theory about the human mind. I will set out how Stapp connects these claims to physical counterparts in orthodox quantum theory, a theory he occasionally refers to as *von Neumann-Wigner quantum theory.*[105] But first, I will start with a brief exposition of what I consider to be James's three most important claims.[106]

Also, his influence on the works of Henry Stapp cannot be overestimated. As a philosopher James was of great importance for Niels Bohr and Wolfgang Pauli, something Stapp seems to have inherited from these founding fathers.

[103] Stapp 2009, *Mind, Matter and Quantum Mechanics*, p.12.

[104] Ibid., p.13.

[105] Stapp 2000, *Decoherence, Quantum Zeno Effect, and the Efficacy of Mental Effort*, p.2.

[106] William James was a psychologist and a philosopher. In many of his writings he approaches the issues of the mind from the perspective of psychology and its relation to natural science. With this in mind it may look as if the overall content of the three presented claims is mainly concerned with psychological aspects of cognition and less with the mysterious character of the first person's perspective, i.e. the 'real hard problem'. James talks about *voluntary selection* and conscious *thoughts*, notions that seem more related to the *content* of our experiences and not so much with the question of why we perceive them the way we do. However, it must be noted that Stapp is not ignorant of the hard problem. In fact, the experiential aspect is fully involved in the core of the strategy he deploys to give physical support to these claims. Whether this aspect plays a necessary role can only be assessed after an exposition of the main elements of Stapp's theory.

Causal efficacy: the mind as a selecting agency James explicitly rejects the conception that the human mind merely acts as a classical physical automaton, subjected to an uncompromising physical causal closure. The automaton picture involves strict determinism and as such offers no place for an efficacious mind whatsoever. On the contrary, James identifies multiple reasons why one should assign efficacy to the mind in relation to the natural world, one of them being the intuition that through our minds we as humans make deliberately *selections out of multiple options* to influence the chain of events in the world: " [...] consciousness is at all times primarily a *selecting agency.* [...] The item emphasized is always in close connection with some *interest* felt by consciousness to be paramount at the time."[107]

Stream of conscious thoughts A second core aspect of consciousness stems from the observation that "No one ever had a simple sensation by itself."[108] Conscious thoughts seem to be elements of a 'continuous' stream. The sense of continuity proposed by James is characterized by:

1. That even where there is a time-gap the consciousness after it feels as if it belonged together with the consciousness before it, as another part of the same self;

2. That the changes from one moment to another in the quality of the consciousness are never absolutely abrupt.[109]

The most relevant implication of the continuity aspect is for Stapp the importance of 'process dynamics' and a role for 'sustainability'. Elements like *memory*, *intentional (volitional) effort*, and *templates for action* enter his theory as demands for the continuity in our streams of thoughts. Stapp observes the following apparent contradiction:

> A fundamental feature of experience is the feel of the 'flow of consciousness', or the 'perception of time'. On the other hand, each actual event is ontologically distinct from all others, and its feel is the feel of itself alone. Thus the 'present' mental event is the feel exclusively of the 'present' physical event; it has no access to past physical events.[110]

Stapp solves this apparent paradox by adopting James's notion of 'specious present'.[111] He observes that "According to this picture, each immediately present mental event

[107] James 1890, *The Principles of Psychology*, p.139.
[108] ibid., p.224
[109] ibid., p.237. Enumeration in original.
[110] Stapp 2004a, p.131.
[111] James 1890 James himself (p.609) actually refers to a term introduced by Edmund R. Clay in *The*

contains within itself a sequence of parts perceived as 'temporally' ordered. [...] Thus the feel of the new event will have components that correspond to components of earlier events."[112]

The unity of conscious thought James applies the notion of *unity* or *wholeness* to conscious thoughts in the following sense: " [...] however complex the object may be, the thought of it is one undivided state of consciousness"[113] And also to its physical correlate, the brain: " [...] the whole brain must act together if certain thoughts are to occur. The consciousness, which is itself an integral thing not made of parts, 'corresponds' to the entire activity of the brain, whatever that may be, at the moment."[114] For both James and Stapp the appearance of unity in both the physical brain and mental phenomena is a major obstacle for classical physics: how can one conceive of meaningful patterns in a picture where the whole machinery is dictated by the interaction between autonomous elementary points of mass? James expresses his concerns as follows:

> But the molecular fact is the only genuine physical fact - whereupon we seem, if we are to have an elementary psycho-physic law at all, thrust right back upon something like the mind-stuff theory, for the molecular fact, being an element of the 'brain', would seem naturally to correspond, not to the total thoughts, but to elements in the thought.[115]

James asserts that only 'genuinely physical facts' can explain the emergence of wholeness, something which the classical physics at the time obviously seemed incapable of.

Stapp's psycho-physical theory

To understand how these three observations about consciousness enter the physical theory it is necessary to start with Stapp's picture of the evolution of the physical universe, a view that is based on what he refers to as the *Heisenberg ontology*. In this evolution three processes are involved, the first one being von Neumann's process 2, i.e. the deterministic evolution of the universe's state in accordance with the relativistic form of

alternative: A study in psychology (1882). At several occasions James applies the term to explain the essence of a single moment of consciousness in a flow of time. Examples are "The specious present has, in addition, a vaguely vanishing backward and forward fringe; but its nucleus is probably the dozen seconds or less that have just elapsed."(p.613) and "But the original paragon and prototype of all conceived times is the specious present, the short duration of which we are immediately and incessantly sensible."(p.631)

[112] Stapp 2004a, p.132

[113] James 1890, p.276.

[114] Ibid., p.177.

[115] Ibid., p.108.

Schrödinger's equation from quantum field theory.[116] The time evolution of the wave function that describes the *Heisenberg state* of the universe follows in the familiar form:

$$S(t + \Delta t) = e^{-iH\Delta t} S(t) e^{iH\Delta t} \tag{3.12}$$

The state S does not represent the physical universe, but rather a collection of "objective tendencies", or "propensities", for the occurrence of an "impending *actual event*".[117] Such a "Heisenberg actual event" is concerned with a collapse of the wave function and, as we are about to see, it falls apart into two separate processes. The process governed by the Schrödinger equation only holds between these Heisenberg events and it has a strict localized character, i.e. all causal connections are due to interactions between "neighbouring localized microscopic elements". Then, in accordance with Heisenberg's idea of *actualization*, he conceives these events, or *quantum jumps*, as the *actual things in nature*.[118] Importantly, these quantum jumps have both "local and global aspects". The global aspect is packed in the idea that an event acts "over a *macroscopic* domain in an *integrative* fashion." This presumption makes it possible to introduce the ideas of *wholeness* and *unit* into the theory, something that was impossible in the picture of classical physics:

> This essential quality of the actual event to grasp as a *unit*, and actualize as a whole, an entire high-level pattern of activity injects into the quantum universe an integrative aspect wholly lacking in the classical conception of nature. This fundamentally *integrative* action of the Heisenberg actual event is the foundation of the quantum theory of consciousness developed here.[119]

Of course, this aspect of the quantum jumps is supposed to align with James's claim about the unity of conscious thought.[120] In this sense Stapp will consider quantum jumps as the 'genuinely physical facts' that James demanded. The integrative aspect gives support to the idea that many macroscopic phenomena will indeed be perceived as 'integrated high-level actions', for example the "firing of a Geiger counter". The global

[116] Stapp 2001, *Quantum Theory and the Role of Mind in Nature*, p.1483.
[117] Stapp 2009, *Mind, Matter, and Quantum Mechanics*, p.122.
[118] Ibid., p.41.
[119] Ibid., p.123.
[120] It must be noted that it also plays an essential role in Stapp's view on neurological aspects of mind-brain interaction. Without going into further details, it is for the current discussion sufficient to mention that Stapp holds that macroscopic brain events have a compositional structure that is isomorphic to the structure of the corresponding human experience. Through the idea of a *body-world schema*, i.e. a continuously updated representation of a body and its environment, he tries to explain the role and character of "top-level processes". Along these lines he claims to explain why we can raise our arm without instructing every individual muscle. For further details, see ibid., pp.124-129

character of the jumps comes from the fact that the whole Heisenberg state will be influenced, i.e. each actual event induces a global change in the tendencies for the next one.

The most important element for the current discussion is that Stapp assigns to the Heisenberg events an 'experiential aspect': "The latter is called the *feel* of this event, and it can be considered to be the aspect of the actual event that gives it its status as an intrinsic actuality."[121] Combined with his consideration that quantum jumps are the actual things in nature, it seems that he comes with a rather strong claim, namely that the full actualization of nature is accompanied with a subjective aspect. To acknowledge the reasons for his assumption it must be noted that Stapp detects two problems in the Heisenberg ontology. The first one he calls the "runaway ontology": when the actual things are only characterized by shifts in tendencies, then we will end up with no more than an infinite sequence of updated possibilities for future actualities. This observation echoes von Neumann's words, i.e. the need for a termination of a measurement chain. To come to the postulation of the experiential aspect he combines this observation with the second problem: " [...] the omission from the description of nature of the one thing really known to exist: human thought."[122] So, at this point we observe that Stapp seems to be in line with the interpretation by von Neumann that I discussed earlier. In fact, I mentioned earlier that von Neumann's picture of process 1 must imply that the representation of first personal data coincides with the actualization of the physical facts they represent. This implication also holds in Stapp's approach when we read that the experiential aspect gives an actual event 'its status as an intrinsic actuality'. Stapp claims that *all* jumps have a subjective aspect.[123] His argument is based on considerations about parsimony and assigns a unique position to the mind-brain interface: "There is no empirical evidence for the occurence of jumps at any place other than the mind-brain interface. Hence there is no scientific basis for introducing other jumps." An important reason for Wigner to leave this same position was because of the role of *decoherence* effects. The idea is that, because of the loss of interference terms in superpositions of quantum systems due to their interaction with the complex environment, quantum theory is in general inapplicable for the description of compound macroscopic phenomena.[124] Zurek points out that decoherence aspects apply to states

[121] Ibid., p.123.
[122] Ibid., p.123.
[123] Heisenberg presumes that they also occur in animals and inanimate objects. He thus seems to allow that the von Neumann chain not necessarily terminates in the human brain. Wigner initially assumed that quantum jumps always involved brain action, but he later changed his mind on this issue. See Butterfield and Fleming 1995, p.130 and Stapp 1995, p.2
[124] ibid., *The Integration of Mind into Physics*, p.2. For an introduction to the role of decoherence in

of the brain as well: "[...] the process of decoherence we have described above is bound to affect the states of the brain: [...] Decoherence, or more to the point, environment-induced superselection, applies to our own "state of mind"."[125] Stapp fully endorses this idea, and he agrees with Tegmark that theories of consciousness that depend on quantum coherence over large parts of the brain are problematic. In fact, he claims that decoherence effects are an important ingredient of his theory of mind:

> [...] the only quantum effects that survive decoherence are those associated with very close neighbors. Thus the quantum state of the brain is effectively, to very good approximation, simply a collection of alternative possible classically described brains. [...] The only macroscopic quantum effect that appears to survive the decoherence effects is the quantum Zeno effect.[126]

This observation is essential for two reasons. First, in order to make quantum jumps relevant in a theory of mind that can adhere to James's idea of unity, he proposes a mind-brain interaction that is based on a mixture of closely resembling essentially classical macroscopic brain states. He assumes that quantum decoherence is the cause for the decomposition into these states. Then, secondly, a central claim in the theory of Stapp is that the mind acts effectively on this mixture via the quantum Zeno effect. To get the idea and its relevance in the overall picture, I will now return to the two processes that together constitute a Heisenberg event.

A Heisenberg event is constituted of what Stapp calls "willful" and "statistically lawful aspects", which are expressed in two very different processes. The first aspect is controlled by von Neumann's process 1, also referred to as the "Heisenberg Choice". The second is governed by Dirac's "choice on the part of nature", referred to by Stapp as "Dirac Choice" or "Process 3".[127] Expressed in von Neumann's mathematical framework, i.e. with the use of projection operators, process 1 looks like

$$S \to S' = PSP + (1-P)S(1-P) \tag{3.13}$$

This process is an intentional mental act on the part of the mind-brain system that produces a jump by probing an aspect of nature through posing a 'Yes-No question'. With P a projection operator with eigenvalues 1 and 0, PSP represents the probability

Quantum Theory, see Bacciagaluppi 2016

[125] Zurek 2007, *Decoherence and the Transition from Quantum to Classical-revisited*, p.21 A practical calculation of the scale of decoherence effects in the brain is provided in Tegmark 2000

[126] Stapp 2011, p.51. The target of both Stapp's and Tegmark's decoherence-based criticism is the Hameroff-Penrose theory of consciousness, the theory of *orchestrated reduction*. (See Hameroff and Penrose 2014, *Consciousness in the universe. A review of the 'Orch OR' theory.*)

[127] Stapp 2004b, *Quantum Leaps in Philosophy of Mind, Reply to Bourget's Critique*, p.2.

of a "Yes" feedback, $(1-P)S(1-P)$ of "Not-Yes".[128] The second process involved in the quantum event, process 3 in Stapp's terminology, is Dirac's reduction of the quantum state:[129]

$$\begin{aligned} &S' = PSP \text{ with probability } \frac{trPS}{trS} \\ &\quad \text{or} \\ &S' = (1-P)S(1-P) \text{ with probability } \frac{tr(1-P)S}{trS} \end{aligned} \tag{3.14}$$

Stapp seems to separate von Neumann's collapse postulate into process 1 *plus* state reduction. The difference is essential, since P depends on the *intention of the agent*, whereas the actual outcome is offered by nature and based on the quantum statistical rules. In other words, the conscious agent *freely* chooses what to probe, the response is a choice by nature.[130] It is important to note that Stapp seems not always to be clear about where he exactly puts the aforementioned experiential aspect, i.e. the 'feel' of the Heisenberg event or quantum jump. What he does say is that each *reduction* has this aspect. He immediately proceeds by stating that "each thought involves an effort to attend to something – i.e., to pose a question – followed by a registration of the answer."[131] So, a thought seems to be a composition of a voluntary 'Yes-No question' and a choice by nature in the form of a wave function collapse. But then, in the same paragraph, he refers to a sequence of thoughts as a stream of 'consciousness'. Thus, it is not totally clear whether the experiential aspect is supposed only to apply to the jump $S \to S'$, i.e. the global state reduction, or whether process 1, as it does in von Neumann's reading, has a 'feel' as well. My impression is that Stapp wants to assign the experiential aspect to the entire Heisenberg event, just like von Neumann did with process 1. Stapp's decision to split this event into the Heisenberg Choice and the Dirac Choice seems to support his idea of a distinction between an *active* and a *passive* role of the mind: the experimenter chooses what to attend to, i.e. "[...] which question he wants to ask about the physical world. This is the active role of mind. The second aspect is the recognition, or coming to know, the answer that nature returns. This is the passive role of mind."[132] With the identification of the active and passive roles of the mind it

[128] See Bourget 2004, *Quantum Leaps in Philosophy of Mind, a Critique of Stapp's Theory*, p.2. This idea is based on the observation that "basic empirical questions always take a yes-or-no form and must be formulated in terms of observable quantities, e.g. 'is the particle at position p?'"

[129] See Dirac 1981, *The Principles of Quantum Mechanics*, p.35

[130] See Stapp 2009, pp.215-217, Stapp 2000, pp.2-4, Stapp 2001, p.1483

[131] Stapp 2000, p.3.

[132] Stapp 2001, p.1486. See also Schwartz, Stapp, and Beauregard 2005, *Model of mind-brain interaction*,

has become easier to explain why and how the quantum Zeno effect enters the theory. Recall from James's first claim about the mind as a willful selecting agency that 'the item emphasized is always in close connection with some interest felt by consciousness to be paramount at the time'. And in relation to the second claim, the continuity of conscious thought, Stapp remarks that 'the feel of the new event will have components that correspond to components of earlier events'. So, although the active role of the mind is concerned with a deliberate choice for a 'Yes-No question', it stands in a relation to 'some interest' or *attention*. That is, attention must lead to continuity in the stream of similar conscious thoughts by posing closely related questions to nature. However, the passive role of the mind seems to introduce randomness: the statistical laws allow nature to come up randomly with 'Yes' or 'No'. So, how could the active role of the mind despite of this randomness become responsible for the establishment of a continuous stream of partially overlapping conscious thoughts, i.e. thoughts that have some similarity? This is where the essential role of the quantum Zeno effect comes in: "A well-known non-classical feature of quantum theory provides, however, a way to overcome this problem [the randomness in nature's choice], and convert the available free choices into effective mental causation."[133]

Suppose a 'Yes-No question' was posed according to (3.13) (process 1), and nature came up with 'Yes'. Then, according to (3.14) (process 3) the state has become $PS(t)P$. If the same 'question P' is repeated after a time interval Δt, then (3.12) gives

$$S(t+\Delta t) = Pe^{-iH\Delta t}PS(t)Pe^{iH\Delta t}P + (1-P)e^{-iH\Delta t}PS(t)Pe^{iH\Delta t}(1-P) \tag{3.15}$$

For Δt sufficiently small a series expansion shows that

$$\begin{aligned} e^{-iH\Delta t} &= 1 - iH\Delta t - \frac{1}{2}H\Delta t^2 ... \\ e^{iH\Delta t} &= 1 + iH\Delta t - \frac{1}{2}H\Delta t^2 ... \end{aligned} \tag{3.16}$$

With $(\Delta t)^2$ very small, the fact that $PP = P$, and the linearity of (3.15), it turns out that the 'leakage' of $PS(t)P$ from P to $(1-P)$ becomes very small, i.e. the second term is effectively washed out up to an expression in the second order of Δt. In effect, this implies that the probabilities of getting 'Yes' when P is repeatedly probed with very small time intervals all approach 1. This 'holding in place' of the outcomes is the

p.12. To add a bit more to the confusion, in this latter text he states that "*It is useful to classify process 1 events as either 'active' or 'passive'.*"

[133] Stapp 2011, p.35.

basis for the deployment of attention and it is a crucial element in Stapp's theory for the support of the notion of continuity in conscious thoughts: "In this model the brain does practically everything, but mind, by means of the limited effect of consenting to the rapid re-posing of the question already constructed and briefly presented by brain, can influence brain activity by causing this activity to stay focused on the presented course of action."[134] According to Stapp it is "just as effective for a statistical mixture $S(T)$ of quasi-classical states as for a pure state: the decoherence generated by interaction with the environment does not weaken this quantum effect."[135]

Stapp and the hard problem of consciousness

The brief sketch I provided in the previous section reveals that Stapp proposes an all-embracing theory about the mind, i.e. a theory that claims not only to explain the existence of consciousness, but that also can account for issues like the *binding problem* and free will.[136] His overall program is an elaboration of von Neumann's psycho-physical parallelism. Is is an effort to work out the psychological observations of William James by attaching them to quantum physics. Several authors have expressed their doubts about the correctness of some of Stapp's physical arguments. The criticisms are mainly aimed at elements in the theory that rely on these arguments to connect the functioning of our brain to functional aspects of experience. Besides the functional aspects of the mind, Stapp suggests that there is an answer from the part of physics to Chalmers's hard problem. That is to say, one should consider the wave function collapse postulate as the basis for giving consciousness a place in nature. My main concern is with this specific aspect and with his claims about having solved the hard problem.

The brief overview of Stapp's picture is sufficient to find out whether he succeeds in

[134] Stapp 2001, p.1489.

[135] Several objections on the basis of decoherence have been made against this point, see for example Bourget 2004 and Georgiev 2012

[136] The binding problem is concerned with the question of why our experiences appear to be tied together. I.e., our conscious states seem to encompass multiple aspects at once. For example, when we experience seeing a red rose we encounter qualia like the 'redness' of the flower, its position in our surroundings, and the meaningful form as a flower it has for us. Bayne and Chalmers (Bayne and David J. Chalmers 2003, *The Unity of Consciousness*) describe how the binding problem can be understood as two strongly related problems. The first is concerned with the problem of "how a system such as the brain manages to bring together two separately represented pieces of information." This is referred to as the *cognitive binding problem*. The second binding problem is that of "explaining how it is that we perceptually experience separate pieces of information as bound together in pertaining to the same object. This is the problem of explaining objectual phenomenal unity." (p.9) For the current discussion it suffices to observe that the binding problem is connected with James's claim about unity. Stapp aims to solve the issue via the integrative character of the Heisenberg events: non-localized events that cover separate connected macroscopic parts of the brain at once.

proposing a reasonable candidate theory to provide what Chalmers calls *bridging* principles between consciousness and science as we know it, i.e. principles that can explain "how experience arises from physical processes."[137] As I mentioned before, Stapp's articles on the mind-matter subject, varying from those in the early nineties up to the most recent ones, reveal a gradually unfolding picture. Many of his writings put a specific focus on physical aspects of the binding problem or to issues concerning free will, like the role of the quantum Zeno effect. Fortunately, Stapp has also written a single article entitled *The Hard Problem: A Quantum Approach*, which summarizes all the arguments he has presented in favor of his idea that the Wigner/von Neumann theory solves the hard problem as it was formulated by Chalmers.[138] This article is one of his most philosophy-oriented papers. In fact, in this text he explicitly pleads for the introduction of the philosophy of mind into the debate about the interpretation of quantum mechanics:

> All interpretations agree on the need for a dualistic ontology, with one aspect being the quantum analog of *matter*, and the other aspect pertaining to *experiences*. Thus the whole debate among quantum theorists is essentially a debate about the mind-matter connection. This debate is precisely where an input from philosophy of mind should enter. To wait until the quantum debate is over is to miss the whole mind-matter ball game.[139]

This is clearly a very strong claim and it demands some closer consideration. According to Stapp, the two parts of this dualistic ontology refer to two kinds of 'beingness'. The first is the quantum mechanical analog of 'matter' in classical mechanics. It is governed by the deterministic evolution of the wave function, and it is "represented as an aggregation of localized properties". I.e., just like in classical mechanics these properties are determined by neighboring properties. The second part of the ontology involves the kind of beingness that "pertains to *choices* between alternative possible *experiences*." Stapp's proposal is interspersed with the terms 'choice' and 'experience' and it is important to realize that these terms have specific connotations in his theory. As we may recall from the previous section, his quantum model of the mind relies on 'Yes-No questions' and an active agent posing them. However, when Stapp refers to 'all interpretations of quantum mechanics' we must be careful not to regard 'choice' as an act of some agency. After all, such an idea is not common in all interpretations. Likewise, the idea of 'experience' suggests that some aspect of awareness of an observer is at play,

[137] David J Chalmers 1995, *Facing Up to the Hard Problem of Consciousness*, p.18.
[138] Stapp 1996, *The Hard Problem: A Quantum Approach.*
[139] Ibid., p.4.

something which, again, clearly does not apply to all interpretations of quantum theory. So, such connotations should be avoided and I think that in the statement above we better should interpret these terms respectively as 'options' and 'perceptions'. With these replacements it appears easier to understand his claim. In such a relaxed form the statement holds that debates about the interpretation of quantum mechanics are focused on explaining the distinction between

- The world described as states that are constituted of 'options' and that evolve in accordance with the Schrödinger equation.
- The world as we 'perceive' it.

It must be noted that this dualistic ontology seems not applicable to GRW since wave function collapse in that theory is a spontaneous process in which the perception of facts does not have a role to play. But the exceptions aside, in this relaxed sense I agree at least partially with Stapp that there *seems to be* a dualistic aspect involved in orthodox quantum theory from the outset, that is to say, a natural distinction between globally accessible descriptions of the material world and a subjective assessment of facts.[140] Or in other words, public facts that are accessible from a third person's perspective and internally represented facts that are only available from a first person's perspective.[141] The dualistic aspect could then follow from the idea that the public facts predicted by the Schrödinger equation do not always coincide with the subjectively perceived ones. For example, in a two-slit interference experiment a public fact can hold that the electron can pass each of the slits with an objective probability, while a first personal fact presents itself to someone as an awareness of the fact that it actually passed one. Clearly, in following von Neumann's idea that an observation should be considered as the conscious termination of a chain of entangled measurement apparatuses, Stapp chooses to connect the second part of the dualistic ontology to the mind of the observer. He forcefully contends that all major interpretations of quantum mechanics endorse the dualistic ontology, i.e. all major versions naturally interpret it as a 'mind-versus-matter'

[140] Whether GRW should be regarded as a version of orthodox quantum mechanics seems open for discussion. See Wallace 2016, *What is orthodox quantum mechanics?*

[141] The impression that we only perceive physical facts from a first person's perspective, plus the idea that we never seem to perceive a state of superposition, forms the rationale behind the collapse-consciousness claim. In the end, everything an observer knows concerns first personal facts. He or she doesn't even know whether there are public facts. Or to make things even more confusing: how should one characterize a fact as public? Public facts may be shared amongst observers, but in the end they have to be digested as first personal data. A first personal fact on the other hand may be 'easily' characterized as 'there is something it is like to perceive this fact'.

aspect.[142] It seems to me that under the assumption of the collapse postulate Stapp's idea of mind-versus-matter could entail promising options for putting the mind somewhere in the picture of the world. However, there seems to be no pressing reason to equate the quantum dualistic ontology with a mind-versus-matter distinction in which the mind is *conscious*. It is easy to see this when we ask ourselves the following question:

Could a zombie live in a quantum world?

The answer seems clear to me: in the two-slit experiment the zombie can both observe the same probabilities as a conscious observer does and register the same outcome *without any need for a conscious experience of the act of registering*. Furthermore, the zombie could fully acknowledge the same quantum dualistic ontology, including an endorsement of the mind-versus-matter aspect.[143] Like many other authors Stapp seems to be tempted to implicitly assign consciousness to 'any form of experience', but in doing so he is mixing up Chalmers's hard and 'easy' problems. The hard problem exposes itself through the conceivability arguments and the explanatory gap. Considering these will again reveal the essential difficulties. When Stapp refers to the mind-matter distinction in the dualistic ontology he points to the difference between for example the description of particles and the perception of them. And I agree, this distinction seems to be prominent in quantum mechanics. But the zombie-argument shows that there is still no reason to assume that perceiving determinate facts as a result from state reduction must go along with a 'feel'. It seems that consciousness remains mysterious as ever and that Stapp's approach will not bring us beyond mere speculation about its existence.

Still, there is room for mitigating such a negative impression. That is, *if consciousness should be assigned a place in reality, one can at least ask whether the second part of the quantum dualistic ontology points to the most reasonable option for doing so.* In other words, if Stapp's two kinds of 'beingness' coincide with the distinction between public and first personal data, is it then reasonable to assume that consciousness is concerned with the part of the ontology where facts are actualized via Heisenberg's quantum events? In that case it is imaginable that one pursues the matter along these lines:

[142] Stapp analyzes the ontology from the perspectives of Bohr's complementarity, Bohmian mechanics, and the many-minds picture. In Bohr's view the dualistic character is constrained to the difference between the world that is described by the wave function and the physicist's 'experience' of the measurement results, thereby not referring to anything like the human mind or consciousness per se. Stapp does not touch on modified versions of quantum mechanics like for instance the spontaneous-collapse or Ghirardi-Rimini-Weber theory (GRW). As already mentioned, the dualistic ontology seems to me not applicable in such a case.

[143] Obviously, for the zombie the mind is a cognitive system with all aspects of binding and possibly free will. However, this mind will not have the consciousness element.

1. Consciousness is real, but it is fundamental in the sense that its existence cannot be explained from any physical facts
2. Therefore, consciousness can be postulated, but it can not be explained
3. The collapse postulate holds
4. We only consciously observe determinate facts that result from state reduction
5. Because of reasons of parsimony, actualization of facts through quantum events provide the only reasonable place to postulate consciousness because reduced states concern the only things we consciously experience.

I think this is exactly what Stapp's theory in fact holds with respect to consciousness. That is, the Heisenberg event provides the only natural place for an aspect like consciousness. On the basis of the zombie-argument one may be tempted to conclude that consciousness still could have been left out. However, if consciousness *is* real (1) and the Heisenberg quantum events provide the *only* possible place to put consciousness into reality (3,4), then there is no reason not to do so. It seems to me that this line of reasoning is needed for Stapp to put any claim on 'some kind' of solution of the hard problem. Although he is not explicit about 'postulating' consciousness itself, It seems that he partially follows this path. To see this, let us first observe his following statement:

> Wigner (1961) and von Neumann (1932), noting that there is nothing in the purely material aspect of nature that singles out where the actual events occur, suggest that these events should occur at the points where consciousness enters: i.e., in conjunction with conscious events. This is the most parsimonious possibility: all of the known valid predictions of quantum theory can be reproduced by limiting the actual events to brain events that correspond to experiential events.[144]

So, Stapp claims that the 'most parsimonious possibility' is to equate actual events with brain events that correspond to conscious experiential events. This may sound like a bold claim, but in fact it is concerned with a variation on von Neumann's measurement chain: the only known valid predictions from quantum theory are the ones that are actually observed. This observation combines steps 3 and 4 above. In the same text he refers to wave function collapse as a 'second process' in nature:

> This second process fixes the actual experiential aspect of nature, as contrasted to the potential aspect. It fixes what our experiences actually will be. And in the most

[144]Stapp 1996, *The Hard Problem: A Quantum Approach*, p.9.

> parsimonious of the available interpretations it consists of actualizations of precisely the functional states that we "feel" are being actualized by our intentional mood.[145]

It seems that Stapp from the standpoint of parsimony points at the actualizations, i.e. the Heisenberg events or the von Neumann collapse, as the natural place to put consciousness as the 'feel' of events in, thereby following step 5. From the perspective of the quantum dualistic ontology this may all seem reasonable: when consciousness is concerned with the subjective experience of determinate facts, then the actualization seems the only place to put it. However, the question remains why consciousness should be concerned *only with determinate facts*. That is to say, why should a non-reduced state not involve such an aspect? It must be noted that there seems to be a widely held implicit assumption that, because we do not *cognitively* experience states of superposition, there will be no consciousness involved in such a state. But this lack of cognitive experience is in itself no reason to allow consciousness exclusively in the realm of definite states. After all, when we have no reason to assume consciousness simply because we cannot explain its existence, then we cannot preclude it either. To avoid these difficulties Stapp appeals to parsimony: it is not relevant whether states of superposition have a conscious aspect, the essence is that *the claim that consciousness is real* is itself based on our experience of reduced states. Or to put it differently, when we hold that consciousness exists then we build this belief on the conscious experience of definite facts, i.e. our awareness of first personal data.

When we have a closer look at the role of the collapse postulate, we may ask ourselves whether we are justified in saying that first personal facts exactly coincide with state reduction. Recall that in my analysis of von Neumann's chain I concluded that the principle of psycho-physical parallelism implies that the representation of first personal data necessarily coincides with the actualization of the physical facts they represent. This implication rests on the assumption that collapse always involves a conscious aspect. But not all approaches to wave function collapse hold this position. Take for example the GRW version of quantum theory. In such a spontaneous collapse theory consciousness is not needed to induce state reduction. In fact, in GRW theory a non-linear component is added to the Schrödinger equation and collapse is considered as a process that can be described in physical terms. Clearly, Stapp's view of the quantum dualistic ontology does not apply to the GRW approach. In fact, within GRW theory determinate facts in the form of reduced states can exist as unobserved globally accessible facts. Clearly, such an interpretation is in disagreement with von Neumann's chain. But, even if there is

[145] Stapp 1996, p.21.

something like 'unconscious collapse' it still may seem that one could apply the parsimony argument: there is no reason to assign consciousness to spontaneous collapse, but we consciously observe definite facts, so that is the place to put consciousness. But a GRW supporter will claim that there is no reason to assume that the conscious perception of a determinate fact coincides with a collapse event, i.e. the state was already reduced on purely physical grounds before it was observed. Now, it is important to see how Stapp modifies von Neumann's principle of psycho-physical parallelism. For von Neumann state reduction and consciousness are on a par, i.e. consciousness gives evidence for the occurrence of collapse. That is to say, a conscious registration of a determinate fact must imply that reduction has occurred. Stapp goes further and claims that consciousness *causes* collapse:

> Why is consciousness so different from the other part of Nature, namely the objective aspect of reality? The objective part of reality has a different kind of beingness: it is mere 'potentia', whereas consciousness is a doer; it is a process of actualization.[146]

> An actualizing element that converts potentia to actuality is needed to complete quantum theory. A coherent role for experience is also needed. Quantum theory allows these two needs to be satisfied together.[147]

What can we conclude with respect to the hard problem from the foregoing observations? First, the existence of consciousness is not explained, but it is rather postulated as a real phenomenon in nature. Further, the zombie-argument eliminates the necessity to assign a conscious aspect to state reduction. However, under the assumption that the perception of facts from a first person's perspective coincides with state reduction, the actualization via Heisenberg events is the logical process to assign consciousness to. This is in a nutshell how the 'feel of experiences' enters Stapp's theory. I observe that Stapp does not 'solve' the hard problem by explaining consciousness, but he secures its place in nature. This place was in fact already identified by von Neumann. Indeed, the interesting thing to note is that the whole proposal *depends on a specific interpretation of orthodox quantum mechanics*. Without von Neumann's collapse postulate and the principle of psycho-physical parallelism there would be no basis for Stapp's picture. It must also be noted that the idea of the quantum dualistic ontology, which is based on the 'mind-versus-matter' distinction, will not necessarily hold for all approaches of quantum theory. But, as I will briefly show in the next section, it can be applied in an interpretation that does not rely on wave function collapse.

[146] Ibid., p.22.
[147] Ibid., p.23.

I end this section with two brief considerations about Stapp's proposal. First, one may ask how the theory should philosophically be classified. On many occasions Stapp distances himself from materialism. Moreover, he positions his theory as a form of *interactive dualism*. But, with regard to the agent that puts 'Yes-No questions' to nature he remarks "[...] the choice of which question will be put to nature, is not controlled by any rules that are known or understood within contemporary physics."[148] It sounds as he does not preclude a possible long-term physical explanation of the active agent involved in collapse. But such an explanation would demand a radical revision of physics. It seems that Stapp avoids Hempel's dilemma for now by taking the 'safe option' of dualism. However, an explanation of an active agent on a physical basis could leave open the option that this could be a 'zombie-agent'. So, the question remains whether Stapp believes that consciousness itself may eventually be understood in physical terms. It is difficult to decide whether he is a true dualist or rather a physicalist who endorses some form of strong emergence. This lack of clarity is perhaps most strongly expressed in the following fragment:

> Can consciousness be 'reduced' to matter? "Matter" is mere potentia for an event. But each conscious event is represented within matter (i.e., within the wave) as the collapse of the wave (function) to a form that embodies the actualized functional structure. The actualization cannot be expressed outside of the matter that embodies it, yet, by virtue of its being an actualization, it is not a mere potentia for such an actualization.[149]

We may infer from this statement that in the view of Stapp physical processes are at least *needed* to bring forth consciousness.

Finally, it is interesting to imagine the application of Stapp's ideas to an artificial intelligent system. Suppose we have a system that is equipped with all sorts of devices for processing sensory input from the outside world, and which is capable of processing data exactly the way we do as humans. What should we hold then from such a system if it is programmed in a way that it probes nature by asking 'Yes-No questions', registering the results, and proceed this way by 'holding attention' via a simulated quantum Zeno effect? Supporters from strong AI will probably hold that this is all possible. And when they endorse the psycho-physical parallelism principle then they may claim that registering events by the machine induces state reduction, and therefore the system will be conscious. On the contrary, opposite views may hold that state reduction will only take place when a genuine conscious agent observes the recordings of the system.

[148] Stapp 2001, p.1483.
[149] Stapp 1996, p.22.

They may claim that the AI system will just play a role of a measurement apparatus somewhere within a von Neumann chain. It may be clear that such a discussion can give rise to interesting thought experiments, but the main issue is that such hypothetical considerations about the implications of Stapp's picture once more illustrate that it rests in the core on the idea that

all determinate facts are induced by consciousness.

This is the essence of Stapp's quantum dualistic ontology.

3.3 Consciousness without wave function collapse

In the previous section we observed how Stapp puts consciousness into what he refers to as the quantum dualistic ontology. Consciousness is then considered as an aspect of a Heisenberg event, i.e. the wave function reduction that is responsible for the actualization of an observed fact. The triggers for these actualizations are governed by von Neumann's process 1. The whole rationale behind the quantum dualistic ontology is in fact based on a single question that characterizes the special character of first personal data: "Why do we perceive physical facts, both elementary and compound, from a first person's perspective and why do we never seem to observe states of superposition?"[150] Or in other words, why does reality evolve in accordance with the Schrödinger equation, while at the same time it is only perceived as a set of determinate facts? The collapse postulate supports the idea that we consciously perceive determinate facts because upon observation these potential facts become real through wave function reduction. That is, in Stapp's picture consciousness forces nature into a state of definiteness. But is this necessary? Could we also conceive of a view in which no reduction takes place and in which we still can hold on to the perception of determinate facts?

The de Broglie-Bohm theory provides a picture in which collapse indeed is absent, i.e. the wave function is not reduced when we observe definite facts. But the theory introduces an additional hidden component, the pilot-wave equation, which governs the actual dynamics of particles. More convenient for a comparison with Stapp's view seems to be the many-minds view, which is based on Everett's *relative state interpretation* of standard quantum mechanics. In this view the Schrödinger equation does not collapse, there are no reduced states but rather 'branches' in which determinate facts are perceived by conscious observers. The quantum mechanical predictions are the same as in

[150] See also B. M. Loewer 2003, p.6

Stapp's picture, but the role for a conscious observer is totally different. A brief comparison between these two pictures dramatically reveals how a philosophical stand towards quantum theory can have its impact on the options left to put consciousness somewhere into nature.

3.3.1 The many-minds picture

In the previous section I already showed how Stapp's view on consciousness is intertwined with von Neumann's collapse postulate. That is to say, without wave function collapse the whole idea of actualization via consciousness vanishes. In Stapp's theory the wave function of the universe is constantly changed through von Neumann's process 1. In the many-minds picture this is not the case: there is only unitary evolution in accordance with the Schrödinger equation. I.e., a state of superposition is not reduced when an act of observation takes place and each of the terms in the state description are considered as equally real. A definite fact then that is perceived by an observer applies to one of these terms. But, there will be a 'separate observer' for all different terms in the state of superposition. This seems to imply that in accordance with the Schrödinger equation the 'observer must evolve into a superposition of belief states'.[151] I will briefly sketch how this is supposed to work, but for now it is important to emphasize that the many-minds approach is based on the assumption " [...] that all physical processes whatsoever are governed by the Schrödinger equation."[152]

Now, to understand the 'branching' into belief states, recall from 3.2.2 how equation (3.10) describes the entanglement of measurement apparatus and the observed system. When we put the conscious observer Alice into this entangled state we obtain:

$$\begin{gathered}(\alpha\,|1\rangle_S + \beta\,|0\rangle_S)\,|ready\rangle_M\,|ready\rangle_{Alice} \\ \xrightarrow{\text{observation}} \\ \alpha\,|1\rangle_S\,|\text{`1'}\rangle_M\,|\text{`1'}\rangle_{Alice} + \beta\,|0\rangle_S\,|\text{`0'}\rangle_M\,|\text{`0'}\rangle_{Alice}\end{gathered} \tag{3.17}$$

After the observation we are left with two often so-called *Everett branches*:[153]

$$|1\rangle_S\,|\text{`1'}\rangle_M\,|\text{`1'}\rangle_{Alice} \tag{3.18a}$$

$$|0\rangle_S\,|\text{`0'}\rangle_M\,|\text{`0'}\rangle_{Alice} \tag{3.18b}$$

[151] Albert and B. Loewer 1988, *Interpreting the Many Worlds Interpretation*, p.197.
[152] Ibid., p.203.
[153] Lockwood 1996, *'Many Minds' Interpretations of Quantum Mechanics*, p.166.

In Everett's own terminology, the *relative state* to Alice's mind state $|\text{`1'}\rangle_{Alice}$ is $|1\rangle_S\,|\text{`1'}\rangle_M$. Likewise, the *relative state* to Alice's mind state $|\text{`0'}\rangle_{Alice}$ is $|0\rangle_S\,|\text{`0'}\rangle_M$. The states described by (3.18) are the tensor product "of an observer state and the corresponding relative state of the remainder of the composite system of which the observer is a part."[154] Now, the essence of the many-minds view in the light of the present discussion is clearly expressed by Lockwood as follows:

> [...] if (in Nagel's felicitous but, by now, well-worn phrase) one asks *what it is like to be* Alice, when she is caught up in the entangled state (3.17), the answer is that it is like remembering seeing the dial read '0' and nothing else, *and* like remembering seeing it read '1' and nothing else. And unless Alice has been converted to the Everett point of view, these recollections will be accompanied with beliefs about the state of her world which respectively coincide with ((3.18)(a) and (3.18)(b)), Alice, we must conclude, is *literally* in two minds here![155]

This is remarkable: Lockwood seems to suggest that consciousness itself *splits* upon an act of observation, i.e. the 'what it is like-question' is distributed over a multitude of options. There is no 'Yes-No choice' involved in the process, nor is there a sense in which consciousness forces nature into some form of reduction. Deutsch points out that one should not misinterpret Lockwood by having the false impression that "it is *only* minds that are multiple".[156] In fact, " [...] it is of the essence of Lockwood's metaphysics that minds are physical systems, and have no preferred status under the universal laws of physics." The difference with Stapp's view in which the mind plays an explicit active role, becomes apparent when we look at Donald's version of a many-minds theory:

> My theory is dualistic in the sense that there are physical laws and there are observers, but there are no mental computations without observable physical structure. My theory is epiphenomenalistic in the sense that a mind does not direct a pattern of observed physical events, rather it has to make sense of such a pattern as it unfolds. Ultimately, however, my theory should probably be considered as idealistic because, in its final form, the central structures in the theory are mental structures.[157]

Donald and Stapp agree that consciousness is a "significant aspect of reality" and that its existence is connected with physical processes.[158] An essential difference with Stapp is

[154] Ibid., p.166.
[155] Ibid., p.166. The citation is almost literal, except for the fact that Lockwood refers to 'L' and 'R' options for the readings of the dial. The corresponding entangled state is of course slightly differently formulated. Lockwood's example is based on Alice doing a Stern-Gerlach type of experiment. Clearly, the essence of his words is not influenced.
[156] Deutsch 1996, *Comment on Lockwood*, p.224.
[157] Donald 2003, *On the Work of Henry P. Stapp*, p.7.
[158] Ibid., p.10.

that both Lockwood and Donald emphasize that there is no reason to presume that "one mind is far more likely to be present at some finite time than the others."[159] In fact, one must be willing to explore the possibility of "simultaneous presence". Another striking difference between Stapp and Donald is concerned with the idea of epiphenomalism: Stapps assigns a physically active role to the mind, whereas in Donald's view the different minds represent the mental structures of the universe, i.e. they do not direct patterns but rather make sense of them.

The relative state interpretation has its own technical difficulties. The *preferred basis problem* and the issue of giving a satisfactory account for the probabilities that arise in the Born rule are prominent examples. I will not go deeper into the technical aspects of the interpretation. Rather, as I announced earlier, my interest is with the question how Stapp's quantum dualistic ontology relates to a view of branching minds. This notion of 'branching' seems an unavoidable consequence from the combination of a) the denial of the collapse postulate, b) the assumption that the unaffected Schrödinger equation describes everything, c) the observation that conscious minds perceive definite facts. Lockwood describes it as " [...] an inescapable consequence of allowing superpositions of what classical physics would regard as mutually exclusive alternatives."[160] He notes that the fact that Alice seems to be in two minds may sound *remarkable*, but that it is "no more remarkable" than "the already utterly mysterious fact that, at a given time, there is even *one* 'what it is like to be' associated with my brain."[161] These words express the often felt uneasiness with the idea of collapse, i.e. the assumption that "when a measurement is carried out, *one* of the possible outcomes occurs *to the exclusion* of all the others.[162]" So, in consequently denying the collapse postulate one is left with the Schrödinger equation as the sole process underlying physical reality. But the observation that conscious minds perceive determinate facts seems to provide a natural reason to think about the quantum dualistic ontology. After all, even in the many-minds picture the only place to conceive of consciousness seems to be at places where definite data are perceived. However, we must be careful not to become sloppy when speaking about Stapp's idea of the ontology. Recall that his view was based on the two different kinds of 'beingness', the potential and the actualized. This idea is totally absent in the many-minds approach. But there is a way to mitigate the difference. Indeed, if Stapp's actualization of a first personal fact is identified with the 'actualization of a new mind through branching', i.e. to something like *coming into being*, then there seems to be,

[159] Donald 2003.
[160] Lockwood 1996, p.171.
[161] Ibid., p.166.
[162] Ibid., p.164.

at least partially, a strong analogy. One may object that this is all contingent and that the views are too different to be compared in such a way. However, this way of looking at the many-minds view reveals that we still might be able to ask again this very important question: '*If one wants to assign phenomenal consciousness a place in the many-minds theory, can one at least ask then whether the 'coming into being' points to the most reasonable option for doing so?*' I think that both Lockwood and Donald will agree with Stapp that this way of putting things reveals agreement with respect to the reality of consciousness in the sense that it has a role in the perception of a determinate reality. They disagree however about the exact contents of the ontology. For Stapp the mind is the active part in one side of the dualistic ontology, the ingredient that forces potentialities into real facts. For the Many-Mind theorists it is a physically passive aspect that has no effect whatsoever on the physical part of the ontology: it does not induce reduction, but it is rather passively split itself, thereby following the patterns that evolve according to the unitary Schrödinger equation.

I finish this brief discussion of the many-minds picture with noticing that this view and Stapp's theory are on a par with regard to the hard problem of consciousness. The conceivability argument and the explanatory gap apply equally forcefully to both, i.e. also in the many-minds view there is nowhere in the theory some indication that the zombie-brain could not be split as well. In fact, one can conceive of a 'zombie Many-Brains world'. This is of course what I tacitly assumed when I asked myself the question whether consciousness – if it is accepted as real – can be *assigned a place* in the world without further explaining its existence. In the next chapter I will argue that this is the only reasonable question to ask with respect to Chalmers's hard problem.

Chapter 4

The hard problem reconsidered

> *On the most common conception of nature, the natural world is the physical world. But on the most common conception of consciousness, it is not easy to see how it could be part of the physical world. So it seems that to find a place for consciousness within the natural order, we must either revise our conception of consciousness, or revise our conception of nature.*
>
> - David Chalmers, *Consciousness and its Place in Nature, 2003*

Recall that at the end of the section about the difficulties with respect to first personal data I concluded that 'the trouble with the very nature of first personal data is that it sabotages every experimental effort to explain *why it is like this to be me.*' (39) My view is based on two observations. First, the simple thought experiment, the one I brought in to elucidate what it actually means when one hopes to identify the presence of phenomenal first personal data, reveals that the very nature of observation makes it impossible to switch between a first– and a third person's perspective. It seems that the nature of first personal data is that it forbids a 'view from the outside', even with regard to its existence.[163] Secondly, considering the matter of consciousness from the perspectives on measurement in physics shows that, whatever one's position with respect to the interpretation of quantum physics, the explanatory gap and the zombie-argument will easily continue to cast doubts on every *explanation* of conscious phenomena. Even without exploring all physical theories about the mind I dare to say that this is easily

[163] Filk and von Müller use the notion of 'self-reference' for systems like the autocerebroscope, i.e. systems that have an influence onto themselves due to an act of "self-observation". They show that this form of self-reference is closely related to "non-separability of observer and observed." For details, see Filk et al. 2009, *Quantum Physics and Consciousness: The Quest for a Common, Modified Conceptual Foundation*, p.10

to be seen: physical theories are by their very nature about views from the outside, and therefore all physical processes that are described could be processes in a zombie-world as well. This is the essence of the hard problem of consciousness and the burden for physics that comes with the explanatory gap. Therefore, a reasonable formulation of the hard problem in terms that can be digested in physics should not be concerned with the provision of an *explanation of a physical role for consciousness*, but rather with the observation that *the physical view from the outside may leave metaphysical gaps to put consciousness into nature*. Once we acknowledge that the 'hardness' of the problem is concerned with the acceptance of this view by science, then one definitely must embrace the view that philosophy and physics are indeed on a par with regard to the issue.[164] It is interesting to have a look at what Chalmers himself has to say with respect to a solution of the hard problem:

> I suggest that a theory of consciousness should take experience as fundamental. We know that a theory of consciousness requires the addition of something fundamental to our ontology, as everything in physical theory is compatible with the absence of consciousness. [...] If we take experience as fundamental, then we can go about the business of constructing a theory of experience.[165]

I agree with Chalmers that 'everything in physical theory is compatible with the absence of consciousness', this is why the zombie-argument is so persistent. But then he proceeds with:

> [...] a nonreductive theory of experience will add new principles to the furniture of the basic laws of nature. These basic principles will ultimately carry the explanatory burden in a theory of consciousness. [...] These psychophysical principles will not interfere with physical laws, as it seems that physical laws already form a closed system. Rather, they will be a supplement to a physical theory. A physical theory gives a theory of physical processes, and a psychophysical theory tells us how those processes give rise to experience. We know that experience depends on physical processes, but we also know that this dependence cannot be derived from physical laws alone. The new basic principles postulated by a nonreductive theory give us the extra ingredient that we need to build an explanatory bridge.[165]

[164] It is interesting to notice how Deutsch and Donald differ with respect to a possibly decisive role for metaphysics in this sense. In a comparison of Stapp's theory with his own many-minds approach, Donald does not think that metaphysics "by itself can provide a convincing refutation [of a physical theory of consciousness]." I think he is mistaken when the issue is truly about the hard problem, i.e. about the 'what it is like to be me' aspect that accompanies perception of facts. Deutsch on the contrary, holds that Lockwood has convincingly put metaphysics back in the arena. For a comparison of the positions regarding the role of metaphysics of both Deutsch and Donald, see Deutsch 1996, p.228 and Donald 2003, p.11

[165] David J Chalmers 1995, *Facing Up to the Problem of Consciousness*, p.14

What I infer from these words is that all statements about consciousness must be foreign to physics. This aligns perfectly well with the idea that the possible worlds described by physics entail zombie-worlds as well. But then there is the curious claim that a psychophysical theory will explain how physical processes give rise to consciousness. At the same time, Chalmers holds that we 'know that experience depends on physical processes', but that we cannot explain this dependence from physics itself. So, this explanation must be left to a theory that is in a certain sense foreign to the theories of physics. With this statement he seems to adhere to a form of strong emergence. However, how should such a psychophysical theory be able to free us from the zombie-argument without giving consciousness a role in physics? After all, how can we conceive of an explanation that tells us how consciousness arises from physics without any consequence for the facts of physics? It feels that we have to accept that the psychophysical theory Chalmers has in mind is in its core based on metaphysical assumptions. To see this it is good to ask the following question:

Q1 *To what extent do Stapp and Lockwood provide reasonable candidates for such a non-reductive theory?*

And related to this question:

Q2 *To what extent does Chalmers's hard problem coincide with philosophical issues about the interpretation of a physical theory, e.g. quantum mechanics?*

It is not difficult to recognize that the earlier discussed collapse-based approaches (von Neumann, Stapp) and the non-collapse many-minds view can be assessed against the background of Chalmers's claims. In all proposals there is a strong connection between consciousness and physical processes without the first entailed by the latter. However, none of the theories *explains* how physical processes give rise to consciousness, i.e. they all merely postulate the existence of consciousness. Obviously, a theory that postulates phenomenal consciousness to give rise to the aspect of 'what it is like to be me' does not entail zombies because these creatures will not have that postulated 'feel'. It must however be noted that this postulate concerns a purely metaphysical claim. In fact, the physics will 'run' independently from the consciousness. This aligns with Chalmers's observation that 'physical laws already form a closed system' and it applies to both Stapp's view and the many-minds picture. But there is also a significant difference between Stapp's theory and the many-minds view. In Stapp's picture the structure of physical reality arises from active minds. In fact, consciousness modifies the physical world. When the mind is left out in his theory then the processes that give rise to the

physical reality must be either spontaneous, or the behaviour of a zombie should also rely on the quantum Zeno effect in his brain, obviously not accompanied with any 'feel'. Although Stapp proposes a 'new brain-physics', it still can run independently without consciousness. On the other hand, in the many-minds theory the mind arises from the patterns dictated by physical laws.

Now the first of our two questions is, should these two theories about the mind be considered as candidates for Chalmers's psychophysical theory? Both theories postulate the existence of consciousness on their own reasonable grounds. But, they do not suggest whether or how consciousness arises from physical facts. This is the reason why in both theories consciousness can be left out without disturbing the physics, although leaving it out obviously conflicts with their metaphysical claims. So, these will not be the kind of theories that Chalmers has in mind. However, I observe that they can be reasonable candidates when we mitigate the hard problem. That is, when we accept that the psychophysical candidate theory will combine a metaphysical stand with regard to the existence of consciousness, thereby eliminating the zombie-option, with a metaphysical position regarding where physics allows us to put consciousness into nature. The price we pay for such a position is that we have to drop Chalmers's requirement to explain how physical processes give rise to consciousness. As an answer to the second of our questions I hold that along such lines the hard problem indeed becomes strongly related to the interpretation issues surrounding physical theories. For example, we earlier saw how von Neumann's collapse postulate provides on parsimonious grounds a reasonable location in physics to put consciousness into the theory. On the contrary, spontaneous collapse theories eliminate this option because reduction is assumed to be a process that can be explained entirely in physical terms. And also in the many-minds picture there is a metaphysically reasonable candidate for such a location, in this case for putting in a multitude of conscious minds. But now the obvious question is: What makes these locations for putting consciousness into physics reasonable from the perspectives of a specific interpretation? The answer lies in the understanding of what first personal data are about. Physics *has* something to say about first personal facts, namely that they are determined under the constraints that physics describes, i.e physical theories tell us what these facts can be. It could even be that future physical theories will explain how the definiteness of these facts comes about and why our brains perceive them. Physical theories per se will not tell us whether zombies could perceive these facts as well. But a physical understanding of the process of observation, i.e. of the perception of first personal data, will point at locations of which we reasonably may assume that physical processes are accompanied with conscious experience, that is to say, as long as we hold

that consciousness is real. This is in fact what von Neumann, Stapp, and Lockwood in essence propose. To end this brief discussion about the hard problem I summarize my most essential observation as follows:

> The essence of the hard problem is not to *explain* the existence of consciousness, but rather to identify metaphysically reasonable grounds to decide 'where to put it into nature'. This is the real burden of the problem for physics. Moreover, this is exactly the reason why philosophy is on a par with physics with respect to the problem: to claim a natural place for putting consciousness into nature the physicist is forced to commit him-/herself to a metaphysical stand regarding the existence and observation of first personal data. In fact, the question where to put consciousness is strongly related to the 'observation-determinate facts' relation.

Summary and conclusion

> *It may be premature to believe that the present philosophy of quantum mechanics will remain a permanent feature of future physical theories; it will remain remarkable, in whatever way our future concepts may develop, that the very study of the external world led to the conclusion that the content of the consciousness is an ultimate universal reality*
>
> - Eugene Wigner, *Remarks on the Mind-Body Question,1961*

The initial 'working title' of this text was *Theories of Everything in a Mindful World: In Search of a Role for Consciousness in Physical Theories*. In fact, it was clear from the start that, paraphrasing Abner Shimony, physical theories of everything should be aimed at 'closing the circle' by the inclusion of the aspect of the world we are most familiar with, our own consciousness. The hard problem of consciousness is concerned with the difficult question of how to explain in physical terms a phenomenon that is not entailed by physics. It seems to me that an obvious starting point for an approach of the problem is the question whether physical theories provide room for a *role* for the phenomenon. As a consequence, I decided that this inquiry should at least touch on the following issues: an understanding of the hard problem of consciousness in relation to the question of what physics is actually about, an analysis of the role of consciousness within some major proposals for solution coming from the side of physics, and aspects of the hard problem in relation to issues in the philosophical foundations of physics. And as we have seen, a treatment of these issues will obviously address topics from the philosophy of mind, from physics, and from the foundations of physics.

An appreciation of the hard problem demands an understanding of why a role for consciousness seems to be absent in physics, i.e. absent in the sense that physics describes worlds that could exist both with and without consciousness. As I discussed throughout the text, this apparent absence of consciousness in physics is dramatically demonstrated by the zombie-argument and the explanatory gap. I have discussed how these arguments

figure in the philosophy of mind, but I have also shown by a thought experiment how they figure 'in practice', e.g. how they put into perspective our implicit interpretations of what can be seen on a MRI-scan of the brain of a conscious person. In short, the arguments tell us that we cannot explain consciousness in the way we explain other phenomena in physics.

The major obstacle for a physical theory of consciousness is concerned with the necessity of a shift in perspective. If we want to study the contents of our conscious experiences we are restricted to an investigation of our own mental phenomena. I have explained that for such inquiries we have to look at the phenomena from a first person's perspective because we need to have immediate epistemic access to it. Moreover, I have also argued that one needs this perspective for even deciding whether consciousness or the 'feel' of things is present at all. As an implication of this claim I inferred that it makes no sense to make the conscious person, or in the physical context the 'conscious observer', the subject of observation in an experiment in which consciousness is the explanandum.

At this point in my analysis I concluded that consciousness is fundamental in the sense that physical theories will not predict its presence. Because I presume consciousness is real, I concluded that both first personal data and public facts seem to exist at the same time. Physics provides a 'view from the outside', i.e. a third person's perspective on public facts. The mind is concerned with an internal view on first personal data, a view from the first person's perspective. The question remains whether the presumed existence of first personal data could play a role in physical theories. I rephrased this question by asking myself whether the presumed consciousness of the observer could be physically relevant in an act of observation. We have seen that this question was already raised by von Neumann, London, Bauer, and Wigner in the context of orthodox quantum mechanics. The answer to the question must be that *one can assign a role to consciousness, but this will be a metaphysical claim since the experiments could be performed in a zombie-world as well.*

Both the psychophysical theory of consciousness of Stapp and the version of the many-minds view by Lockwood show that the physics therein does not demand a role for consciousness. Although the mind figures in both theories, it could be left out, leaving us with a zombie-world. In Stapp's theory this would imply that the causes of the physical processes would be different, i.e. processes initiated by an act of the mind would become spontaneous. In Lockwood's picture there would be no consequences for the causes of physical facts. So, there is no physical need for mental phenomena to keep the physics 'running'. However, there *is* consciousness in both theories, and not without

metaphysical reason. Both Stapp and Lockwood regard consciousness as *real*, something which explains why both point to a reasonable place for putting it into the physical world. Their metaphysical views on quantum mechanics, one collapse-based and the other based on a Many-Worlds picture, allow them to think about where consciousness manifests itself in the physical world. Again, it must be noted that their observations presume that this manifestation is not merely apparent, i.e. consciousness exists. Based on what is known from physics both decide that the place where we observe the presence of conscious experiences is where we consciously perceive determinate facts about nature. So, the rationale is: *consciousness is real and we can point at where it at least exposes itself.*

I want to add two important remarks to this observation. First, I hold that this is a convincing strategy for approaching the problem of consciousness from the perspective of physics. As I explained, the essence of first personal data implies that the zombie-argument and the explanatory gap will remain unbridgeable for physics. But I also hold that if one presupposes the existence of consciousness then it is natural to think about where it is at least observed. An obvious place is there where the conscious observer perceives a determinate fact. From thereon the interpretation of physical theories will do the job: collapse-based views on consciousness *could* hold that it coincides with state-reduction (e.g. Stapp), views without collapse *could* point to other explanations of the definiteness of facts (e.g. Lockwood). Obviously, both positions do not provide a physical explanation of consciousness, rather they point, *from the perspective of their interpretation of quantum mechanics, to the most reasonable option for putting consciousness into nature.*

The second point I want to make is that Stapp puts forward an additional metaphysical claim: he assigns consciousness an *active* role in the physics he proposes. I.e., consciousness modifies the physical world. This claim is somewhat more speculative than the assumptions about where to put consciousness in the world. In fact, it reveals an important difference in the interpretations of von Neumann and Stapp: in von Neumann's picture the role is restricted to the registration of a determinate fact, in Stapp's picture the mind enforces the fact. As we have seen, in his view it is the role of the quantum Zeno effect which makes consciousness more than a mere epiphenomenon.

My analysis of the hard problem from the perspective of physics, combined with the exploration of two example physical theories of consciousness, bring me to the premature conclusion that Chalmers's hard problem of consciousness will not be solved in the way he has in mind: a psychophysical theory is not going to *explain* how consciousness arises from physical facts. That is to say, unless one holds that such an explanation will be

purely metaphysical, immune to every form of falsification. But I expect that this is not the kind of theory Chalmers has in mind. But a mitigation of the problem definitely makes sense for the philosophy of physics. That is to say, the interpretations of physical theories provide different metaphysical frameworks to think about the most reasonable place to put consciousness in. In that respect, I think it is quite natural that questions about the ontology of quantum mechanics can coincide with issues about the ontology of consciousness. To underline my view I recall what I said at the end of the final chapter in slightly different words:

> *The goal of physics can not be to* explain *the existence of consciousness, but it can be about the identification of metaphysically reasonable grounds for where to put it into nature. This must be the real burden of the hard problem for physics.*

Of course, both the philosophy of mind and the foundations of quantum mechanics are concerned with a huge amount of different topics. Many of these are related to specific aspects of consciousness. It was not possible to touch on all of these. Rather I had to make a small selection of issues and articles that seemed relevant for the discussion. As an example, much more can be said about the role of the relatively new notion of phenomenal concepts in relation to immediate epistemic access. Future discussions in this area may shed some new light on the relation between public facts and first personal data, although I do not expect that this will relieve physics from its burden. As a suggestion for further inquiry I must point at the other prominent (physical) theories of consciousness that are currently available. Some examples to be mentioned are the proposals by Hameroff and Penrose, Hiley and Pylkkänen, Tegmark, and Chalmers and McQueen. Although I focused my research on two specific examples (i.e. Stapp and Lockwood), similar points as the ones I made about these can be made about the others as well.

References

Albert, David and Barry Loewer (1988). 'Interpreting the Many Worlds Interpretation'. In: *Synthese* 77.2, pp. 195–213.

Bacciagaluppi, Guido (2016). 'The Role of Decoherence in Quantum Mechanics'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Balog, Katalin (1999). 'Conceivability, Possibility, and the Mind-Body Problem'. In: *The Philosophical Review* 108.4, p. 497.

— (2009). 'Phenomenal Concepts'. In: *The Oxford Handbook of Philosophy of Mind*. Ed. by Brian McLaughlin, Ansgar Beckermann, and Sven Walter. Oxford University Press.

Bayne, Timothy J. and David J. Chalmers (2003). 'What is the Unity of Consciousness?' In: *The Unity of Consciousness*. Ed. by Axel Cleeremans. Oxford University Press.

Blackmore, Susan (2002). 'There Is No Stream of Consciousness.' In: *Journal of Consciousness Studies* 9.5-6, pp. 17–28.

Bokulich, Peter (2011). 'Hempel's Dilemma and domains of physics'. In: *Analysis* 71.4, pp. 646–651.

Bourget, David (2004). 'Quantum Leaps in Philosophy of Mind: A Critique of Stapp's Theory'. In: *Journal of Consciousness Studies* 11.12, pp. 17–42.

Brockman, J (1998). 'Consciousness is a Big Suitcase: A Talk with Marvin Minsky'. In: *Edge .org*.

Brown, Laurie, B Pippard, and Abraham Pais (1995). *Twentieth Century Physics*. CRC Press.

Busch, Paul, Pekka J Lahti, and Peter Mittelstaedt (1996). *The Quantum Theory of Measurement*. Springer.

Butterfield, J and G N Fleming (1995). 'Quantum Theory and the Mind'. In: *Proceedings of the Aristotelian Society*.

Carruthers, Peter and Bénédicte Veillet (2007). 'The Phenomenal Concept Strategy'. In: *Journal of Consciousness Studies* 14.9-10, pp. 212–236.

Chalmers, David J (1995). 'Facing Up to the Problem of Consciousness'. In: *Journal of Consciousness Studies*.

— (1996). *The Conscious Mind*. English. In Search of a Fundamental Theory. Oxford Paperbacks.

— (2006). 'Strong and Weak Emergence'. In: *The Re-Emergence of Emergence. The Emergentist Hypothesis from Science to Religion*. Ed. by P. Clayton and P. Davies. Oxford University Press, Oxford.

Clark, Andy and David Chalmers (1998). 'The Extended Mind'. In: *Analysis* 58.1, pp. 7–19.

Crane, Tim and D Hugh Mellor (1990). 'There is No Question of Physicalism'. In: *Mind* 99.394, pp. 185–206.

Demertzi, A. et al. (2019). 'Human consciousness is supported by dynamic complex patterns of brain signal coordination'. In: *Science Advances* 5.2.

Demircioglu, Erhan (2013). 'Physicalism and Phenomenal Concepts'. In: *Philosophical Studies* 165.1, pp. 257–277.

Dennett, Daniel C (1991). *Consciousness Explained*. English. Boston : Little, Brown and Co.

— (2016). 'Illusionism as the Obvious Default Theory of Consciousness'. In: *Journal of Consciousness Studies* 23.11-12, pp. 65–72.

d'Espagnat, Bernard (2013). *On Physics and Philosophy*. Princeton University Press.

Deutsch, David (1996). 'Comment on Lockwood'. In: *The British Journal for the Philosophy of Science* 47.2, pp. 222–228.

Dirac, Paul Adrien Maurice (1981). *The Principles of Quantum Mechanics*. 27. Oxford University Press.

Donald, Matthew J (2003). 'On the Work of Henry P. Stapp'. In: *arXiv preprint quant-ph/0311158*.

Dowell, J.L. (2006). 'Formulating the Thesis of Physicalism: An Introduction'. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 131.1, pp. 1–23.

Elitzur, Avshalom C (1989). 'Consciousness and the Incompleteness of the Physical Explanation of Behavior'. In: *The Journal of Mind and Behavior*, pp. 1–19.

— (2009). 'Consciousness Makes a Difference: A Reluctant Dualist's Confession'. In: *Batthyány, A., & Elitzur, AC (Editors) Irreducibly Conscious: Selected Papers on Consciousness. pp. 43-72. Heidelberg: Universitätsverlag Winter.*

Faye, Jan (2014). 'Copenhagen Interpretation of Quantum Mechanics'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2014. Metaphysics Research Lab, Stanford University.

Feynman, Richard P, Robert B Leighton, and Matthew Sands (1965). *The Feynman Lectures on Physics, vol. 3*. Addison-Wesley Reading, Massachusetts.

Filk, T et al. (2009). 'Quantum Physics and Consciousness: The Quest for a Common Conceptual Foundation'. In: *Mind and Matter* 7.1, pp. 59–79.

Frankish, Keith (2016). 'Illusionism as a Theory of Consciousness'. In: *Journal of Consciousness Studies* 23.11-12, pp. 11–39.

Franklin, Allan and Slobodan Perovic (2016). 'Experiment in Physics'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Georgiev, Danko (2012). 'Mind Efforts, Quantum Zeno Effect and Environmental Decoherence'. In: *NeuroQuantology* 10.3.

Hameroff, Stuart and Roger Penrose (2014). 'Consciousness in the universe. A review of the 'Orch OR' theory'. English. In: *Physics of Life Reviews* 11.1, pp. 39–78.

Heisenberg, Werner (1958). *Physics and Philosophy*. New York: Harper.

Hempel, Carl G (1980). 'Comments on Goodman's Ways of Worldmaking'. In: *Synthese* 45.2, pp. 193–199.

Howard, Don (2004). 'Who Invented the "Copenhagen Interpretation"? A Study in Mythology'. In: *Philosophy of Science* 71.5, pp. 669–682.

Hut, Piet and Roger N Shepard (1996). 'Turning 'The Hard Problem' Upside Down & Sideways'. In: *Journal of Consciousness Studies* 3.4, pp. 313–329.

Isham, Chris J (2001). *Lectures on Quantum Theory. Mathematical and Structural Foundations*. Allied Publishers.

Jackson, Frank (1982). 'Epiphenomenal Qualia'. In: *The Philosophical Quarterly (1950-)* 32.127, pp. 127–136.

— (1986). 'What Mary Didn't Know'. In: *The Journal of Philosophy* 83.5, pp. 291–295.

James, W (1890). *The Principles of Psychology, NY, US. Henry Holt and Company*.

Jammer, Max (1974). *Philosophy of Quantum Mechanics: The Interpretations of QM in historical perspective*. John Wiley and Sons.

Kirk, Robert (1999). 'Why There Couldn't Be Zombies'. In: *Aristotelian Society Supplementary Volume*. Vol. 73. 1. Oxford University Press Oxford, UK, pp. 1–16.

— (2015). 'Zombies'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2015. Metaphysics Research Lab, Stanford University.

Kripke, Saul A (1972). 'Naming and Necessity'. In: *Semantics of Natural Language*. Springer, pp. 253–355.

Levine, Joseph (1983). 'Materialism and Quanta: the Explanatory Gap'. English. In: *Pacific Philosophical Quarterly* 64.4, pp. 354–361.

— (2006). 'Phenomenal Concepts and the Materialist Constraint'. In: *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, p. 145.

Lewis, David (1983). 'New Work for a Theory of Universals'. In: *Australasian Journal of Philosophy* 61.4, pp. 343–377.

Libet, Benjamin (2006). 'Reflections on the Interaction of the Mind and Brain'. In: *Progress in Neurobiology* 78.3, pp. 322–326.

Libet, Benjamin et al. (1983). 'Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act'. In: *Brain* 106.3, pp. 623–642.

Loar, Brian (1990). 'Phenomenal States'. In: *Philosophical Perspectives* 4, pp. 81–108.

Lockwood, Michael (1996). ''Many Minds'. Interpretations of Quantum Mechanics'. In: *The British Journal for the Philosophy of Science* 47.2, pp. 159–188.

Loewer, Barry M (2003). 'Consciousness and Quantum Theory: Strange Bedfellows'. In:

London, Fritz and Edmond Bauer (1983). 'The Theory of Observation in Quantum Mechanics'. In: *Quantum Theory and Measurement*. Ed. by J.A. Wheeler and W.H. Zurek. Princeton University Press. Chap. II.1, pp. 217–259.

Lowe, E Jonathan (2000). 'Causal Closure Principles and Emergentism'. In: *Philosophy* 75.4, pp. 571–585.

Meehl, Paul E (1966). 'The Compleat Autocerebroscopist: A Thought-Experiment on Professor Feigl's Mind-Body Identity Thesis'. In: *Mind, Matter and Method, eds. Feyerabend PK & Maxwell G.. University of Minnesota Press.[aJAG]*.

Montero, Barbara (1999). 'The Body Problem'. In: *Nous* 33.2, pp. 183–200.

— (2003). 'Varieties of Causal Closure'. In: *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, pp. 173–187.

— (2009). 'What is the Physical?' In: *The Oxford Handbook of Philosophy of Mind*.

Nagel, Thomas (1974). 'What is it Like to be a Bat?' English. In: *The Philosophical Review*, pp. 435–450.

— (1989). *The View from Nowhere*. Oxford University Press.

Nida-Rümelin, Martine (2015). 'Qualia: The Knowledge Argument'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2015. Metaphysics Research Lab, Stanford University.

Penrose, Roger (1987). 'Quantum physics and conscious thought'. In: *Quantum implications: Essays in honour of David Bohm*, pp. 105–120.

Pippard, A. Brian (1988). '"The Invincible Ignorance of Science", Eddington Memorial Lecture'. In: *Contemporary Physics* 29.4, pp. 393–405.

— (1992). 'Counsel of Despair'. In: *Nature* 357, 29 EP -.

Putnam, Hilary (1992). 'The Nature of Mental States'. In: *The Philosophy of Mind: Classical Problems/Contemporary Issues*, pp. 51–58.

Redhead, Michael (1987). *Incompleteness, Nonlocality, and Realism: a Prolegomenon to the Philosophy of Quantum Mechanics*. Oxford, Clarendon Press.

Robinson, Howard (2017). 'Dualism'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University.

Schwartz, Jeffrey M, Henry P Stapp, and Mario Beauregard (2005). 'Quantum physics in neuroscience and psychology: a neurophysical model of mind–brain interaction'. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 360.1458, pp. 1309–1327.

Sellars, Wilfrid (1963). 'Philosophy and the Scientific Image of Man'. In: *Science, Perception and Reality* 2, pp. 35–78.

Shimony, Abner (1963). 'Role of the Observer in Quantum Theory'. English. In: *American Journal of Physics* 31.10, pp. 755–773.

Smart, J. J. C. (2017). 'The Mind/Brain Identity Theory'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University.

Stapp, Henry P (1995). 'The Integration of Mind into Physics'. In: *Annals of the New York Academy of Sciences* 755.1, pp. 822–833.

— (1996). 'The Hard Problem: A Quantum Approach'. In: *Journal of Consciousness Studies* 3.3, pp. 194–210.

— (2000). 'Decoherence, Quantum Zeno Effect, and the Efficacy of Mental Effort'. In: *arXiv preprint quant-ph/0003065*.

— (2001). 'Quantum Theory and the Role of Mind in Nature'. In: *Foundations of Physics* 31.10, pp. 1465–1499.

— (2004a). 'A Quantum Theory of the Mind-Brain Interface'. In: *Mind, Matter and Quantum Mechanics*. Springer, pp. 147–174.

— (2004b). 'Quantum Leaps in Philosophy of Mind: Reply to Bourget's Critique'. In: *Journal of Consciousness Studies* 11.LBNL-55887.

— (2009). *Mind, Matter and Quantum Mechanics*. The Frontiers Collection. Berlin, Heidelberg: Springer Berlin Heidelberg.

Stapp, Henry P (2011). *Mindful Universe: Quantum Mechanics and the Participating Observer*. Springer Science & Business Media.

Stoljar, Daniel (2005). 'Physicalism and Phenomenal Concepts'. In: *Mind & Language* 20.5, pp. 469–494.

Tegmark, Max (2000). 'Importance of quantum decoherence in brain processes'. English. In: *Physical review E* 61.4, pp. 4194–4206.

Tye, Michael (2018). 'Qualia'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University.

Von Neumann, J (1955). 'Mathematische Grundlagen der Quantenmechanik (Berlin, 1932)'. In: *English Edition Princeton, NJ*.

Wallace, David (2016). 'What is orthodox quantum mechanics?' In: *arXiv preprint arXiv:1604.05973*.

Weinberg, Steven (1994). *Dreams of a Final Theory. The Scientists's Search for the Ultimate Laws of Nature*. Vintage.

Wigner, Eugene P (1995). 'Remarks on the Mind-Body Question'. In: *Philosophical Reflections and Syntheses*. Springer, pp. 247–260.

Wilson, Jessica (2006). 'On Characterizing the Physical'. In: *Philosophical Studies* 131.1, pp. 61–99.

Zurek, W H (2007). 'Decoherence and the Transition from Quantum to Classical–Revisited'. In: *Progress in Mathematical Physics*.