Mind: A Property of Matter

by

Penelope Rowlatt

University College London

PhD

May 2017

Signed declaration

I, Penelope Rowlatt, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.[1]

[1] I am very grateful to those who have assisted me in this work. Paul Snowdon was very helpful and supportive in the early stages of the work, but retired before it was completed. Rory Madden made many useful comments in the following months. My greatest debt, however, is to Luke Fenton-Glynn, for the generosity with which he has given his time and energy in order to pick holes in my arguments and quarrel with my use of words, thereby improving the quality of this work beyond recognition; working with him has been a great pleasure.

Abstract

There are three broad possibilities regarding the basic ontology of mind. It could be a property of matter that reduces to the properties that are studied in physics. It could be a property of matter different from those that are studied in physics. It could be nothing to do with matter.

The second of these, known in the literature as non-reductive physicalism, is generally considered by philosophers in limited form with mental states, albeit non-reducible, fully determined by other properties of matter (taken to be 'emergent from', and 'supervening on', the properties of matter studied in physics). My thesis puts the case for the ontological status of mind being similar to that of the other properties of matter, those studied in physics. The approach lends itself to the proposition that mental states can be causally effective *per se*, since that is the case for the other properties of matter. This proposition runs counter to the usual assumption in the philosophy literature relating to mental causation known as "the completeness of physics", which requires that all physical events are fully caused by purely physical (non-mental) prior histories. However, theoretical physicists often propose new phenomena for a variety of reasons.

There is a lot in favour of this approach. None of the three anti-physicalist arguments (the knowledge argument, the conceivability argument and the hard problem) cause it difficulties. Effective mental causation means that the reason why creatures with consciousness abound in our world could be that consciousness enables effective decision-taking and so has been selected for by the pressures of survival. Effective mental causation would also explain why people feel as if they have freedom of the will: if mental states are causally effective there would be a sense in which people do have free will.

Contents

Chapter 1 Introduction

1.1 The purpose of the thesis

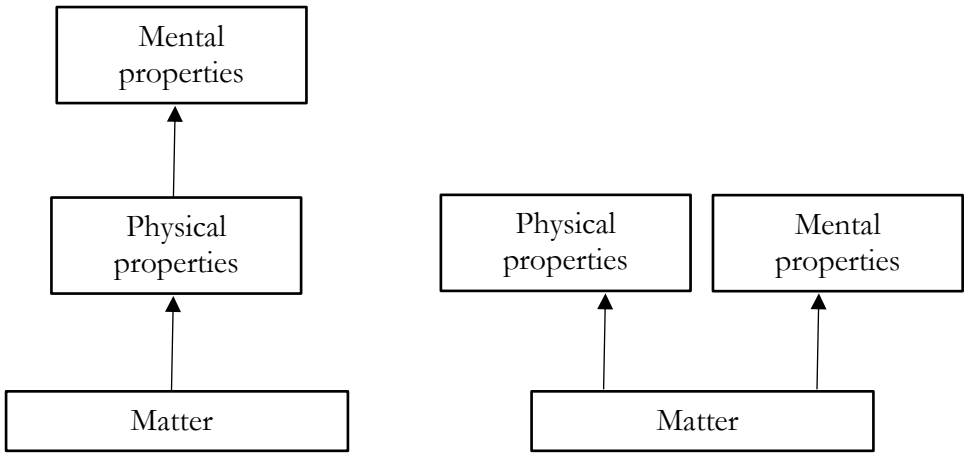
The question of the nature of mind is one of a handful of fundamental questions that relate to our existence which surely everyone must wonder about from time to time. An obvious issue that arises on these occasions is whether mind is something to do with matter, most probably with brains, or whether instead it is quite distinct from matter. If it is taken to be related to matter another question arises: are states of mind, 'mental states', part of the physical world, in the sense of being related to nothing more than the properties of matter that are studied in the physical sciences, the approach known as 'reductive physicalism', or is this not the case? And if it is not the case, there is yet another question: could there, then, be causal processes which are such that the mental state that we experience as the taking of a decision has a direct effect on our actions through its influence on the properties of matter studied in the physical sciences, or must such a mental state be 'epiphenomenal'?

There are three main possibilities concerning the relationship of mind and matter, and they can broadly be described as: dual substances (with mind a property of a different substance to matter), reductive physicalism (in which mental states are part of the physical world, related to nothing more than the properties of matter studied in the physical sciences) and non-reductive physicalism (where mind is a property of matter that is not related to those properties that are studied in physics).[2] My purpose in this thesis is to put the case for a particular form of non-reductive physicalism, one in which mental states (states of which a creature could be conscious) instead of being supervenient on properties of matter studied in physics, as is usually assumed, have an ontological status similar to those properties, as in Figure 1b.

Philosophers who study physicalism often confine themselves to the case in which mental states 'supervene' on the properties of matter that are studied in physics, with the implication that mind is something that could be called a 'higher level' property, ontologically different from the other properties of matter. Indeed, supervenience of this

[2] In Section 1.2 I discuss the meaning of the word 'substance' in this context and present definitions of 'physics', 'physical' and 'physicalism'.

sort is often viewed as being the minimal requirement for the approach that is generally known to philosophers as 'physicalism'.[3] I take a different approach in this thesis. I propose that mind could be a property of matter in its own right, just as is the case with the properties of matter that are studied in physics.[4]



a. Supervenience physicalism b. Property dualism

Figure 1: Supervenience physicalism and property dualism

In the figure the arrows imply that the properties at the sharp end only exist if the matter or the properties at the blunt end exist. The diagram illustrates the difference between the case in which mental properties are taken to 'supervene' on physical properties and the case in which mental properties are properties of matter in their own right, with an ontological status similar to that of the physical properties.

Given this 'property dualism' approach, there seems to be no reason to assume that the particular property of matter that we call 'mind' is epiphenomenal when all the other properties of matter are causally effective. Indeed, as many have observed, one might think that the mere fact that whenever you are aware of deciding to raise your arm it performs exactly the gesture you were aware of intending would be enough to convince

[3] We should note that, if 'physicalism' is defined as the proposition that *everything* is suitable for study in the physical sciences (see Section 1.2.1), 'non-reductive physicalism', which assumes that mental states are not suitable for study in the physical sciences, becomes a contradiction in terms (see Section 1.2.5). So with this definition of physicalism, we would need to refer to the proposal presented in this thesis, which does not involve 'two substances' (it assumes that there is only one substance and we call it 'matter', see Section 1.2.4), in some different way.

[4] The use of the term 'mental state' suggests that there must be a "thing" called 'mind'. But 'mind' may be nothing more than a convenient way of referring to the range of our possible mental states; the term comes in handy as a means of communication (as in the title of this thesis).

any reasonable person that there is a significant likelihood that your decision is, in itself, causally effective; and, as we shall see, there are reasons to suppose that feelings, and therefore consciousness, are a key factor in practical reasoning, the name philosophers use for the cognitive process underlying the taking of such decisions.

The proposal examined in this thesis, then, is a version of property dualism in which mental states can be causally effective, so I call it 'causally effective property dualism'.

Although the general approach will be familiar to philosophers, to the best of my knowledge there are no academic papers or books that spell out in any detail, or depth, the implications of assuming that causally effective property dualism is the correct ontological account of mind.[5] And this in spite of the fact that there is a lot that is supportive of it:

- The philosophers' three anti-physicalist points (encapsulated in the knowledge argument, the conceivability argument and the hard problem), which are generally thought to cause serious problems for reductive physicalism, have no power against causally effective property dualism (see Section 2.6), so there is no reason why this approach should not be accepted as a logically viable and plausible form of materialism. It enables materialism to be the case, even if reductive physicalism is deemed unsatisfactory (perhaps because of the anti-physicalist arguments), because it means that we don't need to postulate another substance besides matter in order to accommodate the existence of mental states (see Chapter 2).

- Effective mental causation means that the reason why creatures with consciousness abound in our world is likely to be that consciousness has been selected for by the pressures of survival. I argue that it is the ability to experience feelings that enables creatures to take decisions appropriate to their future welfare given the circumstances in which they find themselves (see Section 1.4.3), and that

[5] McGinn describes an approach somewhat similar to this in his 1982 book (*The Character of Mind*, Chapter 2), but he doesn't address the question of effective mental causation and he later suggests that: "the mind-body problem brings us bang up against the limits of our capacity to understand the world" (1989, p 354). Strawson (eg., 1994, 2008) is perhaps the only other modern philosopher who views "experiential phenomena" as relating to matter in a similar way to that of non-mental phenomena, but he does not explore the possibility of mental causation and its implications. McLaughlin refers to it (2010, p 268) but takes it for granted that mental states cannot be causally effective, referring to them as "nomological danglers" (see Feigl, 1958). However, dubbed "interactionism" (a term which usually describes the situation in which mind is a property of a different substance from matter and causal effects operate between mind and matter) this approach is included in Chalmers' (2002) classification of views on the metaphysics of consciousness under the heading "Type D Dualism".

this gives them an evolutionary advantage. If that were so, facilitating the taking of such decisions would be the 'role' played by consciousness (Chapter 4).

- If causally effective property dualism holds, this would explain why people feel as if they have freedom of the will; if mental states are causally effective there would be a sense in which people *do* have freedom of the will (Chapter 5).

However, the proposal also has the obvious implication that, if it were to hold, physics would not be 'causally complete' in the sense that: "all physical effects are fully caused by purely *physical* prior histories"; and this causal completeness of physics is an assumption that it seems every philosopher of mind signs up to these days, albeit, in some cases, not without question.[6] However, physics as we know it today is unfinished, and if there turns out to be another property of matter, and/or a small, localised, additional force, and/or some effect related to quantum mechanics, which only arises in the brains of creatures taking decisions, there is no reason to suppose that this would have any effect on the way physics operates in the situations in which it is normally studied. Although the influence of effective mental causation on the future course of the world can obviously be huge, the fact that mental states are causally effective, if that turned out to be so, seems unlikely to have any significant effect on the operation of physics outside the brains of conscious creatures. Further, it seems highly unlikely that neuroscientists would have happened on the existence of such a phenomenon, given the current stage of development of that specialty, particularly if they were not actually looking for it.

There seems to be no reason why these mental states and the causal processes relating to them should not be part of the combination of causal processes and events which, taken together, would then make up the discipline that we call 'science', so there would seem to be no reason why science, the combination of the physical sciences and the causal processes that involve mental states, should not be complete, in the sense of being the study of *all* the events that can be causes or effects along with *all* possible causal

[6] Questions have been raised about the causal completeness of physics by Baker (1993, p 77), who puts the case that *something* that is presently accepted by everyone has to go, perhaps the 'completeness of physics', and also by Burge (1993), see the discussion in Yoo (2016); but neither of these develops the possibility into a coherent proposal.

processes.[7] Although science would then be 'complete', in this sense, physics, of course, would not.

It soon becomes apparent, as one reads the words of philosophers and scientists, that many of those who consider this question of the nature of mind and also assume that there is only one substance do not share the above view that there could be something different from that which is studied in the science of physics.[8] Further, the proposition that mental states can be causally effective *per se* (that is, that it is *not* the case that: "all physical effects are fully caused by purely *physical* prior histories", see, for example, Papineau, 2002, p 17) is a philosophical position that receives little support nowadays from philosophers. Kim (2005, pp 15-17) rules out mental causation by assuming the completeness of physics in the above form. Papineau (2002) accepts that: "Conscious states involve awareness, feelings, the subjectivity of experience" (p 1) but he also signs up to the same formulation of the completeness of physics as Kim, nevertheless (*ibid*., pp 17-18, and see Section 3.4.4). Lewis (1988, p 90) finds it impossible to accept that phenomenal facts exist (what it is like to experience something, *eg*., what it feels like to see red, see Section 1.2.1) and continue to "uphold materialism", so he rejects the existence of phenomenal facts and mental causation *per se* goes too (see Section 3.4.2). Chalmers, however, while ruling out 'physicalism' in 'An Argument against Materialism' (Chapter 4 of his 1996 book, p 123) is also one of the few modern philosophers to acknowledge the possibility of effective mental causation, but only in the form of 'over-determination' (*ibid*., p 152 and see Section 3.4.5). These are just a few of a great many who have signed up to this causal completeness of physics principle in the recent philosophy literature.

In the rest of this chapter I first address, in Section 1.2, some fundamental assumptions about the nature of science and conventions regarding the meanings of words that I found I needed in the rest of the thesis. I refer back to this section whenever I use the material. Section 1.3 then addresses the question of what is meant by the term 'consciousness', and notes some facts that have been established by scientists about what we are conscious of and about the possible role of memory stores and the likely implications of memory stores for the structure of the mind. Section 1.4 contains three fundamental propositions about "mind" which I suggest are knowable from introspection;

[7] I take it here that all the other "sciences", such as chemistry, biology, economics *etc*. can be explained in terms of the physical sciences along with some additional causal processes that involve mental states.

[8] Which is not to deny that there are also plenty who support non-reductive physicalism.

clearly, others are not signed up to these propositions since if they were generally accepted there would be no need for me to write this thesis, so they should be considered to be controversial. Section 1.5 then spells out the justification for the approach to the ontology of mind that I present in this thesis, which derives from the material in Sections 1.2 and 1.4, while Section 1.6 summarises the structure of the thesis.

1.2 Fundamental assumptions

In questioning the nature of consciousness and the manner in which it relates, or does not relate, to the properties of matter studied in physics, I am raising questions that concern the foundations of science, and I found it necessary to spell out the following fundamental issues in order that I could be clear about the rationale underlying various propositions that arise later in the thesis.

First, what do we mean by 'physics', what is included in physics and what is not, and why? I address these questions in Section 1.2.1.

Then, in science, we assume that there are causal processes that operate "out there" and we aim to discover what they are; a distinction therefore needs to be drawn between the causal processes "out there" that we seek to understand and our current formulation of them, that is, the "laws of nature" that we propound. Importantly, it is a fact that some of these so-called "laws of nature" could, on further investigation, be found to be wrong. Related to that, there are some statements that we *know* are true, but some which, whereas we believe them to be true given our current knowledge, could need revision given further information. These issues are addressed in Sections 1.2.2 and 1.2.3.

Third, what do we mean by the assertion "we assume that there is only one 'substance'", particularly in the light of the developments in physics in recent decades? I discuss this issue in Section 1.2.4.

Finally, in Section 1.2.5, I note that the fact that the word 'physical' has two meanings can cause confusion, for example relating to the meaning of 'physicalism' and the meaning of 'materialism', and that this means that care is needed when using these terms.

1.2.1 Defining 'physics' and 'physical sciences'

It is often assumed by philosophers, either implicitly or explicitly, that *every* event can be explained in terms of facts relating to the sciences of physics, chemistry and neurophysiology *etc.*, known as the physical sciences (psychology, economics and some aspects of biology, which involve mental states, are not included here), and therefore in terms of physics, since it is generally assumed that all these so-called 'physical sciences' "reduce" to physics. But we can only take a view on whether *all* the facts that we can know are suitable for study in these sciences, or whether there is a type of fact which is not suitable, if we have a definition of 'physics'. To address this question we need first to be clear about what type of fact is studied in physics (or the physical sciences, since it is assumed that all the facts related to these can be explained in terms of physics); then we can address the further question of whether there may be some other type of fact that is not suitable for study in physics. If we think there could be another type of fact, not suitable for study in the physical sciences, we need to understand *why* it is not suitable for such study.

There currently seems to be no explicit agreement amongst philosophers as to what is necessary in order for a property of matter to be suitable for study in the sciences that we call 'the physical sciences'. Further, since philosophers often define the term '*physicalism'* in terms of physics (see Section 1.2.5), without clarity about the definition of 'physics' we cannot be clear about precisely what the important word 'physicalism' can mean.

A great many philosophers have asked the question "what is physical science?", see for example Crane and Mellor, 1990 (pp 186-87).[9] Beckermann is one of the few who have tried to spell out an answer to the question, saying of 'physicalism' (1992, p 2):

> "It [physicalism] was set forth as a theory of science and especially as a theory of the foundations of science. Its main claim was that 'physical language' is 'the universal language of science' and that therefore, even protocol sentences, i.e., those basic sentences of science which form the starting point of all hypotheses and the evidence by which to check their validity, should be couched not in phenomenalistic, but in physical language. The main reason for this claim was that only physical language is intersubjective and therefore apt to provide a suitable basis for scientific research".

[9] Amongst many others addressing this question are Papineau (2002, pp 40-44), Tye (2011, p 26), and Chalmers (1996, pp 118-119, 128-129).

Beckermann, however, then finds it unclear how the term 'physical language' is to be understood, noting (in a footnote to the above passage) that sometimes it is understood to be "thing language" and that on other occasions it includes only terms referring to physical quantities that can be ascribed to "space-time-points" (*ibid*., p 2).

A similar question relating the science of physics to language was raised by Davidson (1970, p 210) when he said:

> "What does it mean to say that an event is mental or physical? One natural answer is that an event is physical if it is describable in a purely physical vocabulary, mental if describable in mental terms".

So how do we know what 'physical language' is; how do we know what can or cannot be spoken of as 'mental terms'?

A fundamental requirement of the physical sciences is that everything about experiments relating to some property of matter that are performed by one person can be perfectly reproduced by other people, including, in particular, the experimenters' observations regarding the outcome of the experiment. It then follows that these must all consist of 'public facts', public in the sense that the experimental design and the observations that result must all be capable of being examined and noted by any number of people on an *equal* basis. This means that it is necessary that these things can be communicated between people, perfectly, using language.

Therefore, we start from the fact that an essential feature of the physical sciences is the testing of hypotheses by experimentation and the reproducibility by another person of experimental results discovered by one person. This requirement, which relates to interpersonal communication, enables us to take a view about which properties of matter are and which are not suitable to be studied in the physical sciences. It is only if a fact relating to some property of matter is a public fact, one that is equally accessible to everyone (the weight of a stone in certain specified conditions, for example), that it can be referred to in words with no ambiguity ("this stone weighs 5 kilograms" combined with the definition known as the International Prototype Kilogram or IPK). If there are facts relating to a property of matter that are not public facts, that property is not suitable for study in the physical sciences, for the simple reason that such facts would not be equally available to everyone. Importantly, this means that for some property of matter to be suitable for study in physics it has to be the case that *all* the facts relating to that property

can, in theory at least, be communicated perfectly from one person to another linguistically.[10]

However, it is difficult to dispute the proposition that other facts exist, in addition to these public facts (see Section 1.3.2). This is the case for facts to which one person or creature has privileged access, such as "what it feels like for X to see red" (we can never be sure whether or not it feels the same for W to see red) or "what it feels like for Y to be sad about the death of his mother" or "what it feels like for Z to be aware that 'this stone weighs 5 kilograms'" (as opposed to knowing *that* it does; as well as "knowing that" something is the case, there is something it *feels* like to be aware that something is the case, see Section 1.3.2). These "what it feels like" facts are the phenomenal facts (or *qualia*). If a property of matter has facts relating to it that are not public facts, because they are not equally accessible to everyone, that property is considered to be unsuitable for study in the physical sciences. It follows from this that "what it is like"-type information about experiences relating to our senses and to our feelings, because it can only be directly observed by the person/creature who experiences the feelings, is not suitable to be studied in what we call "the physical sciences".

1.2.2 Causal processes, laws of nature and science

I assume in this thesis that either determinism or 'universal probabilistic causation' (the probabilistic version of determinism, see Section 5.2 for more details) holds. That is, I take it that every event arises as the result of a causal process, and may, itself, cause some other event to occur, given the causal processes (which may be probabilistic) and the situation that pertain. Following Davidson's comment "… laws are linguistic" (1970, p 215 in 2001 reprint) I take the 'causal process' to be something that exists "out there" and I take a 'law of nature' to be our linguistic/mathematical expression of some presumed causal process (so it can be the case that a law of nature is found wanting, given new information, see below). Where a causal process relates to events that can be described in public language, it can be captured precisely by a law of nature and can generally be expressed in the form (in the simplest possible terms): "if event A occurs then event B will occur given situation C" or perhaps "if A then B_1 with probability p_1, B_2 with probability

[10] Where some of the facts relating to a property can be treated in this way, like the surveys relating to how "happy" people "feel" about some experience, indicated on a scale of 1 to 10, for example, but others, like what it actually *feels* like to have the experience, cannot, the property is not suitable for study in the physical sciences.

p_2 …, given C" (but see Section 1.4.3 for a very different specification of a causal process). Where a mental characteristic is relevant to a causal process (for example, "the experience of pain causes avoidance behaviour") it will obviously be impossible to capture individual occurrences of it precisely in language/mathematics; but that should not be taken to imply that the causal process does not exist, or that true causal "law-like" statements cannot be made regarding it (such as "hunger causes a creature to seek food" or "X is likely to choose A in circumstances C").

I assume that causal closure pertains, that is, that *every* event is subject to a causal process and I take it that whereas the study of the causal processes that can be expressed precisely in language/mathematics constitutes what we call 'the physical sciences', the study of *all* the causal processes would constitute what we call 'science'.

I take the standard, necessitarian, approach to the causal processes: a causal process is a necessity relation in which the reason for the necessity is different from that which underlies logical necessity (as described by O'Connor, 1995; see also Johansson, 2006); I refer to this as 'nomological necessity'. However, the necessity relations that we humans postulate to hold at any point in time and therefore *assume* to pick out causal processes, having observed regularities in the way events occur in the world, are inevitably dependent on the possibility that some new observation (a black swan) suggests that we have got it wrong and that a reassessment is needed. Also, I do not rule out the possibility that, although we might be correct in assuming that the causal processes that we seek to discover exist, it could be impossible for us ever to discover the truth about some of them (see Section 3.4.5).

We generally refer to the type of relation described above as implying that: "A causes B given C", but we don't usually spell out exactly what we mean by the word 'cause'; and I have no need to take a view on this in the context of this thesis.

1.2.3 The truth of "statements of fact"

We saw in Section 1.2.1 above that two different types of fact exist: phenomenal facts (facts about what something feels like, facts that are "privileged" or "private" in that they cannot be communicated perfectly to another person linguistically), and public facts (facts that can be perfectly communicated using language). In this section we are concerned with

public facts.[11] Setting aside statements concerning historical events, there are three different situations that typically arise regarding the truth of statements about these facts. From time to time in the course of this thesis we will need to note which of these seems likely to apply (for example, in the cases of the three anti-physicalist arguments, see Section 2.6).

The reason why there are three such situations is the existence of two distinctions that can be made regarding the truth of statements of such facts. A distinction can be made between a statement of fact the truth of which can be discovered by introspection, and so can be ascertained "whilst seated in an armchair" (the term *a priori* is used for some of these facts), and a statement of fact the truth of which depends on the results of observations or investigations of some sort other than mere thought (the term '*a posteriori*' is sometimes used by philosophers in these circumstances).[12] A further distinction can then be made between two types of statements of fact which require observations or investigations; for some, once an observation has indicated support for them, we know that they are necessarily true, see below (in these cases the truth is necessary '*a posteriori*'), while for others, although investigations may indicate support for them, so we might at some point in time *believe* them to be true on the basis of these observations, it remains possible that they could be falsified by future observations. This third category can lead to confusion of exposition. The confusion arises because the word 'fact' is generally defined as "a thing that is known or proved to be true" (see, for example, https://en.oxforddictionaries.com). Further, in philosophy, as Mellor (1995, p 8) makes clear, 'facts' are taken to be: "actual states of affairs, corresponding to true statements". If this were correct it would mean that when I say to my friend: "you've got the facts wrong" that statement is incoherent. In the real world, the states of affairs that are presented to us as 'facts' all too often turn out to be untrue; they're often just plain wrong.

In more detail, the first case noted above concerns the possibility that a statement of fact can be known to be true "whilst seated in an armchair". There are two, different, ways in which this can be so. First, a statement of fact can be known to be true as a matter

[11] Some of the material covered in this section is similar to that which is treated in Kripke's famous book *Naming and Necessity* (1972), but I come at it here from a very different point of view. See also Johansson (2006).

[12] In philosophy and logic 'contingency' is the status of propositions that are: "neither true under every possible valuation (*i.e* tautologies) nor false under every possible valuation (*i.e* contradictions)" (see https://en.wikipedia.org/wiki/Contingency (philosophy)).

of logic; this will be the case if the statement derives from a statement that is true by definition (such as "no unmarried man is married") as the result of the replacement of one of its terms by a synonym (as in the case of: "no bachelor is married", see Quine (1951, p 23). In this case, the statement of fact can be said to be true *a priori*.[13] Second, a statement of fact can be known to be true as the result of introspection, as in "I think" (*cf.*, Descartes': "I think, therefore I am"), which has the implication that each of us knows, from introspection, that mental states exist (see Section 1.4.1). One might say such a statement can be deemed to be true through introspection.

Next, we turn to the type of statement of fact that encapsulates the results of observations or experimentation, and consider the situation in which, once some observation or experimentation suggests that some proposition may be the case, we know that the statement that it is the case is indubitably true; it cannot be found in the future to be false (that is, we know that the "black swan" situation cannot arise). Simple observations such as the fact that "you can't have one side of a coin without the other", and the fact that "the world is round" come into this category,[14] as do identities relating to composition such as "water is H_2O" (where the meaning of 'water' is "that which, in humans, results in the experience of waterish *qualia*") and the proposition that Hesperus is the same object as Phosphorus (see Kripke, 1972, p 28). Some of these are referred to as 'laws of nature' or 'natural laws' (see Section 1.2.2) in which case the truths they convey are deemed to be 'nomologically necessary', but some are identities of the "two names for one thing" variety (numerical identities) so their truth is logically necessary.

The other type of statement that encapsulates the results of observations or experimentation, consists of those which, although thought to be true as the result of some experimentation, might be found to be false on further investigation. This is the category which is not generally spelt out clearly in the philosophy literature. Although we can *never* be certain that statements of this sort are true, we often treat them as being true (perhaps with a general understanding that they may not be), so the language used in relation to them can be misleading. Like some of the statements in the previous category, these may be thought of as 'laws of nature' of some sort and, when it is taken that they are a true

[13] The term *a priori* is used when the truth of a statement can be justified in a manner that is independent of experience, see Russell, Bruce, "A Priori Justification and Knowledge", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.).

[14] Apparently, it was thought by many in past centuries that the earth was flat and that you could sail to the edge of the world and then you might fall off, but now we have been out there and seen what the world looks like from space.

representation of some causal process that operates, their validity would be viewed as being 'nomologically' necessary. But we know that they could turn out to be false. For example, various laws relating to classical physics were found wanting when quantum mechanics was developed; and laws such as: "all ravens are black" are always vulnerable to being proved untrue. These laws are statements of fact that appear to be true given observation or experimentation relating to the world around us, but sometimes a statement of this sort turns out to be false on further investigation (as in "all swans are white").

Finally, from time to time in this thesis statements of 'existence' arise (in Section 2.6.3 concerning the hard problem, for example, and implicitly in Figure 1), such as the proposition that some constituents of the substratum that we call 'matter' instantiate an electric charge. Obviously, these statements are not true as a matter of logic. However, some of them are, indubitably, true. I therefore refer to them in this thesis as laws of nature; they are statements of truths of the "you can't have one without the other" variety that we have discovered as a result of our observations of the circumstances in which certain things arise. Thus if we were to discover that certain configurations of neurons are associated with consciousness I would refer to this fact as a law of nature and state that it is nomologically true; it should then be borne in mind that further evidence might be found in the future that suggested otherwise.

1.2.4 The meaning of "there is only one 'substance' …"

An element of the background to the proposal regarding the nature of mind that is presented in this thesis is the assumption that there is only one 'substance', which we call 'matter'. We all understand broadly what is meant, in this context, by the words 'substance' and 'matter', which I use from time to time throughout this thesis (one of which occurs in its title). But both these terms raise questions.[15]

First, consider the case in which a 'substance' (or 'substratum') is that which is the bearer of properties (see Crane, 2001, p 35) in the sense that: "A substance is what *has* properties, what is the *bearer* of properties. Properties are *possessed* by a substance, they

[15] The word 'substance' can have two meanings (see Robinson, Howard, "Substance", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.)). Sometimes it is used to pick out something that could be said to 'stand under' or 'ground' things, a 'substratum' perhaps. Sometimes it is used to pick out a particular kind of entity that may consist of a bundle of properties.

belong to a substance or they *inhere* in a substance." In this thesis I refer to a substance, when the word is used in this sense, as 'instantiating' properties.

Second, we should note that the way in which the word 'matter' (*cf.*, "there is only one 'substance' and we call it 'matter'") used to be used in physics encapsulates a view of the universe that is now somewhat out of date: that view can be characterised as assuming that matter consists of particles similar to billiard balls or marbles in that they occupy space and have mass (along with other properties). We now know that some so-called particles, photons, have no mass and so cannot really be said to be "made of matter" in the old-fashioned understanding of that term, and we also know that energy is as fundamental to our understanding of the universe as mass and that the two can convert, one into the other. We need a more sophisticated statement of what we mean when we refer to there being "only one 'substance' and we call it 'matter'".[16]

The aspect of a 'substance' or of 'matter' that is important in this study of the ontology of mind is the existence of what we call "properties of matter". This raises two questions. The first concerns the use of the word 'substance' in this context. Should we be replacing the proposition that there is something that could be called a 'substance', a *substratum* that is the bearer of properties (see Armstrong, 1989, and Schneider 2013), by the proposition that there is nothing other than a simple bundle of properties, an approach known as 'bundle theory'? But there is no necessity to take a view on this issue in this thesis; what matters here is the completeness (or otherwise) of physics/science, and therefore the existence of causal processes relating to the properties we associate with this hypothetical substance we call matter, and to those properties' interactions. The second question is the more interesting: could the proposition that there is only one 'substance' be replaced by a different type of proposition, such as the proposition that science is causally closed?

As explained in Section 1.2.2, in this thesis I assume that events are involved in causal processes; that is, I assume that every event that occurs has some sort of a cause or collection of causes (that is, that every event can be viewed as being the end-point of a forking chain of causes going back to the Big Bang, see Section 5.4.1), and also that every event may have a causal role with respect to some subsequent event. Where possible, we encapsulate these causal processes in statements that we call 'laws of nature' or 'natural laws', which may take a probabilistic form (see Sections 1.2.2 and 5.2). These laws of

[16] Crane and Mellor make this point in their 1990 paper (see p 186).

nature are part of the systematised body of knowledge about the natural world that we call 'science'. And we assume in this thesis that a finished science would be complete, in the sense that all the causes and all the effects along with all the causal processes would be included within it.

This suggests that we could replace the requirement that "there is only one 'substance'" with the requirement that "a finished 'science' would be complete" ('complete' in the above sense).

The implications of this become clearer if we consider counter-examples. If supervenience physicalism (see Figure 1) were to hold it would not be the case that physics (to which a subset of the natural laws in science would belong, see Section 1.2.1) could be described as being complete in this sense. This is because if mental states supervene on the properties of matter studied in physics, but do not have the characteristics that make it appropriate for them to be included in the science of physics (because there are facts relating to them that cannot be communicated perfectly by one person to another in language) there would be natural laws that specify the way the physical properties of matter must be for some particular mental state to arise, and these natural laws would have an effect that is outside physics; so physics would not be complete in this sense (although, of course, it would still be causally complete in the sense that: "all physical effects are fully caused by purely *physical* prior histories"). Another example of a lack of completeness of this sort arises in the case of libertarianism as it is described in Section 5.2.3 (where a mental event could arise uncaused). If such libertarianism pertained, various physical properties could be affected by something that was not part of science, and this would mean that some events that affect the properties of matter studied in science would have no cause. In this case, therefore, science would not be complete in this sense.

Before leaving this discussion of the meanings of 'substance' and 'matter' we should note that in general conversation the word 'matter' is often implicitly taken to imply "lifeless", and therefore "non-experiential"; "that which occupies space and is distinct from mind and spirit". In this case, 'matter' would be lifeless and lack consciousness *by definition*. But there is no particular reason to suppose that matter is lifeless or lacks consciousness; indeed, one could argue that there is every reason to suppose the opposite: casual observation suggests that matter is closely associated with both life and consciousness (see, for example, Strawson, 2008). It may be that philosophers sometimes commit themselves to this rather restricted conventional meaning associated with the word 'matter' without being aware of what they are doing, and that that has led to some of the

problems that arise in this area of philosophy, such as the well-known "hard problem" (how can it be that consciousness arises in the context of matter?) and the lack of belief in panpsychism and emergence (see Levine, 1983 and Chalmers, 2007 and Section 2.6.3).

1.2.5 The meaning of 'physical' and 'physicalism'

The term 'physical' has two distinct meanings in common parlance: it can mean "pertaining to objects such as the body, as opposed to the mind" and it can mean "pertaining to matter" (see Section 1.2.4).[17] In modern philosophy, however, the term 'physical' almost invariably nowadays has the meaning: "pertaining to that which is suitable for study in the physical sciences" (see Section 1.2.1) which I take to be the sciences that reduce to physics, such as chemistry and neuroscience, as well as physics itself.

This distinction between the two meanings is touched upon by Searle when, in his 1983 paper entitled 'Why I am not a property dualist', he complains of: "the inadequacy of the traditional terminology" saying:

> "The property dualist wants to say that consciousness is a mental and therefore not physical feature of the brain. I want to say consciousness is a mental and therefore biological and therefore physical feature of the brain" (*ibid.*, p 61).

Horgan (1984), also, notes the ambiguities that can arise regarding the use of the term 'physical' (see Section 2.6.1). Strawson (2006, p 4) refers to the physicalism that assumes that everything can be explained by the physical sciences as 'physicSalism', while calling the physicalism that requires that everything pertains to matter, so could be said to be "fully realist about consciousness", 'real physicalism'.

If physical is used in the "pertaining to matter" sense, then mind, if it is an irreducible property of matter, could still be termed a 'physical' phenomenon in spite of the fact that things related to mental states are excluded from the science that we call physics. To take 'physical' as having both meanings at the same time is implicitly to assume that everything that pertains to matter is non-mental, or generates facts of the type studied in the physical sciences, when there may be no reason why that should be the case. One might think that the answer is always to interpret the word 'physical' in modern philosophy as meaning "pertaining to physics"; but if the 'physicalism' in 'non-reductive

[17] See for example *http://www.dictionary.com/browse/physical.*

physicalism' is interpreted as implying that everything relating to matter is of the sort that can be studied in physics then the term, which refers to the possibility that mental states *cannot* be reduced to things that are studied in physics, becomes self-contradictory.[18]

In terms of our investigation of the nature of mind the fact that the word 'physical' has these two different meanings matters because it causes confusion. But in addition, simply to assist general communication about things that matter to us, we could do with a term that encapsulates the assumption that there is only one substance, matter, without carrying with it the automatic implication that the substance called matter has no properties other than those that are studied in physics. If mind were independent of matter it would remain possible that minds could survive the death of the body, while if mind is a property of matter (albeit not a property suitable for study in the science of physics) this is not the case. Nowadays, in philosophy, *both* of the terms 'materialism' and 'physicalism' are generally taken to imply that all the properties associated with matter are suitable for study in the science of physics. Using the terms in this way means that the arguments presented by philosophers against reductive physicalism appear as arguments against 'physicalism' or 'materialism' in general even when they do *not* carry the implication that if something exists that cannot be explained in terms of the physical sciences that thing has to be 'immaterial' (see Section 2.6).

1.3 About consciousness

In this section I summarise the facts about consciousness that have been proposed by philosophers and psychologists and are relevant to the material in the rest of this thesis. I start by defining consciousness; then I note the different sorts of things of which we can be conscious and some of the different types of things which we can 'know'. This is followed by a discussion of the relation between consciousness and memory and a formalised hypothetical structure of the mind, showing some of the relations psychologists

[18] In modern philosophy 'physicalism' is sometimes defined as the situation in which: "any possible world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world" (see, for example, Block and Stalnaker, 1999, p 10). I have avoided that definition in this thesis because of this ambiguity concerning the meaning of the word 'physical'. For if 'physical' is taken to mean "pertaining to physics" the definition implies that the situation that I am exploring in this thesis, the case in which mental states are instantiations of a property of matter that is different from those studied in physics, is *not* a case of 'physicalism' (for it would not then be the case that when the matter in two worlds is such that the physical properties are the same, the mental properties are necessarily always the same too). But if 'physical' is taken to mean "pertaining to matter", property dualism *can* be classed as a type of 'physicalism'.

have suggested exist between different types of consciousness, different types of memory store and attention. The section finishes by noting that one day we might be able to form a better view, albeit one that is likely to be considered controversial, concerning which creatures besides ourselves are conscious and which are not.

1.3.1 Defining consciousness

Some mental states are deemed to be 'conscious', so it seems sensible to take a look at what is meant by that term before proceeding further.[19] What do we mean when we use the term 'consciousness', or when we assert that a creature "is conscious" or appears to be "conscious of" something?[20] People have often remarked that it can be difficult to define consciousness, since most of the initial thoughts people have as they make the attempt turn out to be near-synonyms for consciousness (awareness, wakefulness, cognizance, and so on). I therefore start by examining what a definition is required to achieve.

Suppe (2000, p 76) says that "[i]n the most fundamental scientific sense, to define is to delimit." This suggests that a formal definition of a term Y, might assert that a thing is a Y if it is a thing of type X and has the property Z, thereby enabling us to decide, definitively, whether or not some particular thing is a Y. The definition of being conscious that I use in this thesis takes that broad form, but it has to be slightly adapted in order to cater for the fact that, at any rate given the current state of science, we cannot be absolutely certain whether another creature is capable of consciousness (as in "the problem of other minds"). It is derived from the definition proposed by Nagel, in 1974. He said: "an organism has conscious mental states if and only if there is something it is like to *be* that organism – something it is like *for* the organism" (1974, p 436, his italics). Of course, Nagel's definition is concerned with organisms; but it is easily generalised to cover robots as well. The result can be expressed, broadly in Suppe's type of formulation, as: "a thing is considered to have conscious mental states if we take the view that it has the property of there being something it is like to *be* that thing – something it is like *for* the thing".

---

[19] Psychologists tell us that most of what goes on in our minds is not conscious, it is hidden from our view, with many of our routine decisions taken on automatic pilot. Kahneman says (2011, p 52): "The notion that we have limited access to the working of our minds is difficult to accept because, naturally, it is alien to our experience, but it is true: you know far less about yourself than you feel you do." I take a mental state to be something of which a person/creature could be conscious, if they were attending to it, but at any point in time may not be.

[20] To suggest that a creature "is conscious" or may be "conscious of" something seems to be a meaningful suggestion. Because that is so, we have coined a noun, "consciousness"; but it is not clear to what, exactly, it refers.

We can check this definition for plausibility by looking at examples. Given this definition of being conscious we would probably reckon that a table, a plant, a stone or a mountain is not conscious. A person in a deep sleep or a person in a coma might be viewed as not conscious, although judged to be capable of being conscious in other circumstances. A wakeful person would be judged to be conscious by everyone (barring only solipsists, if such exist). We might be less certain when it comes to other animals or robots. Is a dog capable of being conscious, or a horse, a mouse or an ant, or a robot? Importantly, it seems to be impossible for us to be 100 per cent certain about the answers to some of these questions (see Sections 1.4.2 and 2.5.3), although that situation might change if neuroscience, as it becomes more highly developed, were able to identify, convincingly, the neurological counterparts of being conscious (see Section 1.3.6).

1.3.2 What we are conscious of

It is clear that conscious mental states consist of what are generally known as feelings and thoughts. In this section we identify three different types of thoughts and feelings of which we can be conscious.

First, there is the type of consciousness known as 'phenomenal consciousness'. This is the term used for the wide variety of sensory experience and emotions of which we can be aware, it picks out the "what it feels like" element of the things we can be conscious of; it also picks out the what it feels like element of things that are "viewed with the mind's eye" after being recalled from the long-term memory. 'Phenomenal facts' is the term we use for the facts regarding what something *feels* like. Chalmers, in his 1995 paper 'Facing up to the Problem of Consciousness', cites a number of examples of phenomenal experiences in his list of things we might be conscious of:

> "the felt quality of redness, the experience of dark and light, the quality of depth in a visual field ... the sound of a clarinet, the smell of mothballs ... bodily sensations, from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion …" (*ibid.*, p 201).

As mentioned above, the quality of these experiences has been given the name '*qualia*' by philosophers.[21] David Hume was among the first to construct a catalogue of the types of

[21] Some philosophers deny the existence of *qualia*, which seems strange given that the term refers to the only things we can be certain exist, namely the experiences that make up our own feelings.

things of which we are conscious, in 1739/40, and later attempts have done little to improve on his efforts. He described the nature of phenomenal consciousness, stressing the fact that what something feels like cannot be communicated in words, saying: "To give a child an idea of scarlet or orange, of sweet or bitter, I present the objects, or in other words, convey to him these impressions; but proceed not so absurdly as to endeavour to produce the impressions by exciting the ideas" (1739/40, p 52), and: "We cannot form to ourselves a just idea of the taste of a pineapple, without having actually tasted it" (*ibid.*, p 53; see also Locke, 1671).

A person/creature has privileged access to this type of "what it feels like" fact, known as phenomenal facts; we can never be certain that the same sensory input – seeing some particular shade of red, for example, or the characteristic pain of a stubbed toe – feels exactly the same for different people. Indeed, it clearly can be the case that it does differ if only in the case in synaesthesia, where a person is "wired up" in such a manner that stimulation of one sense or part of the body produces a sense impression relating to another sense or part of the body. Information about these phenomenal facts, or *qualia*, can be stored in our long term memories and recalled, as in Chalmers' "mental images that are conjured up internally" (see above), for use in cognitive analysis. Many phenomenal facts relating to the current phenomenal experience of your external and internal senses are available to your consciousness at any moment (the pressure of the chair on your back, the sound of a passing car outside the window) and some of this information may be used in cognitive processing – in thinking, reasoning, taking decisions, *etc.* Recalled phenomenal facts often form an element of what we call 'concepts' (see footnote in Section 3.3.5).

We can make a distinction between two kinds of phenomenal information: first, there is something it is like to receive information about things that are external to our bodies by our external senses such as sight, hearing, taste and so on, and to receive information about the state of our bodies through the directly body-related internal senses such as proprioception and pain receptors *etc.* Second, information about the emotions we experience, such as anger, love, sadness or joy generally consists of a bundle of sensations received through internal senses containing information about the physical aspects of emotions such as the effects on the body of adrenalin and the like, along with the effects we observe emotions to have on the patterns of our thoughts and desires.

A second type of mental activity of which we can be conscious relates to certain aspects of our thoughts. Block has suggested (1995), that although much of the information consciously used in cognitive processing is of the phenomenal type, there is

another type of information of which we are conscious, and which we use in cognitive processing, alongside these phenomenal facts. He stresses that these other things of which we are conscious are available: "for use in reasoning and rationally guiding speech and action" (*ibid.*, p 227). He calls this 'access consciousness', and stipulates that a state is access conscious if "a representation of its content is (1) … poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech" (*ibid.*, p 231). Here, I take this second type of information of which we can be conscious to consist of information that can, in principle, be communicated perfectly from one person to another using language; and, because information of this sort is available equally to everyone, I call facts of this type 'public facts'. Consciousness of public facts, facts that can be expressed in language and perfectly communicated from one person to another linguistically, may not be precisely what Block had in mind for what he called "access consciousness". But in this thesis a lot rests on the question of whether information can or cannot be perfectly communicated linguistically, given our definition of physics and therefore physicalism (Sections 1.2.1 and 1.2.5), so I hope that I can be forgiven for adapting Block's classification to take account of this distinction between different types of fact.

It is clear that as well as being conscious of knowing what it feels like to experience something, we can be conscious of knowing that the statement of some supposed fact about the external world may be true (the question of whether there is "something it is like" to know a public fact is the question of whether there is such a thing as 'cognitive phenomenology'). Such a statement might be that "Paris is in France" or that "elephants cannot jump" or that "this stone weighs 5 kilograms". These facts differ from the phenomenal facts described above in that they are statements relating to 'public' facts, facts that are equally available to everyone and can be readily communicated from one person to another linguistically. Like the phenomenal facts, these public facts also can be fed into and retrieved from our long-term memories and, in addition to the memories of the types of thing on Chalmers' list of phenomenally conscious experiences, memories of these facts are also available 'for use' in cognitive analysis. And, again like the phenomenal facts, these facts are often included in some bundle of facts that is generally accepted as being a 'concept' (the concept named 'dog' generally involves having four legs and being hairy, as well as having the capacity to learn to obey human instructions). Where phenomenal consciousness corresponds to "knowing what it feels like" to experience something, access

consciousness (defined in this manner) can be termed as "knowing that" something is the case.

As mentioned above, Hume (in his *Treatise*, 1739/1740), along with other philosophers at that time, also made distinctions of this sort regarding the types of information of which we are conscious. Hume (*ibid*., p 49) called all the information that appears in our minds 'perceptions', distinguishing between two different varieties of perception, one that seems to correspond to what we call 'feelings', which he called 'impressions', and the other, which he described as being "those that appear in thinking and in reasoning", and so appears to correspond to what we call 'thoughts', which include public facts, and which Block (perhaps) considers to be things of which we can be 'access conscious'; Hume called these 'ideas'. Hume further divided phenomenal consciousness into two 'kinds', what it is like to receive information from the external senses and the directly bodily-related internal senses (that is, from sight, touch *etc*., and pain, itching or proprioception *etc*.), and what it is like to receive information from the other internal senses, those that relate to emotions such as desires, hopes, fears and moral feelings *etc*. Hume gives as an example of his 'ideas': "all the perceptions excited by the present discourse" *excluding* those that arise from the senses, such as sight and touch, and also excluding (interestingly) any emotional reaction, such as "the immediate pleasure or uneasiness it may occasion", because emotions are classified as 'impressions' (*ibid*., p 49). So this leaves only the information about external facts (our public facts) which "the present discourse" contains, or draws from the long-term memory to be included in his category 'ideas'. It is interesting and impressive that Hume, in 1739/40, along with many other thinkers, such as Locke, Descartes and William James, reached a conclusion on the types of thing of which we are conscious that so closely resembles the one psychologists use today.

In addition to these two different types of fact of which we are conscious, there is another, third, type of mental activity of which we are aware and it is cognitive processing, that is, thinking (imagining, reasoning, taking decisions, *etc*.) along with the implementation of decisions (the triggering of speech, muscular activity and so on), making use of the two types of fact. This type of consciousness is less well represented in the philosophical and psychological literature than phenomenal consciousness and Block's access consciousness (but see Peacocke, 2007 and Prinz, 2007). Chalmers refers to it in his 1995 paper (p 201) citing: "… the experience of a stream of conscious thought …" as an example of 'experience'. Block (2001) calls consciousness of cognition 'reflexive consciousness'.

1.3.3 Varieties of 'knowing'

As noted above, it is often said that there are different types of 'knowing'; Ryle (1949, Chapter 2) famously argued that "knowing that" is different from "knowing how". And as Snowdon (2003, footnote, p 1) notes, there is also the use, in the English language, of 'know' as in "acquaintance", that is, know followed by a noun (knowing a place, or a poem, or a person).

But this can be controversial. Some philosophers deny the existence of phenomenal facts, the "knowing what it feels like" type of knowledge, saying that all that is learned when a new experience takes place is an ability, something of the "knowing how" variety (see Lewis 1988 and Section 3.2.4). Some people have raised questions regarding Ryle's proposal, suggesting that the "knowing how" variety of knowing could be just a collection of "knowing that's" (see Stanley and Williamson, 2001 and Snowdon, 2003). These issues matter because they arise in the context of Jackson's knowledge argument (see Section 2.6.1) and Lewis' case against the existence of phenomenal information (see Section 3.4.2), both of which raise important questions about the ontology of consciousness. However, it should be noted that the phrase "knowing how" can be used both in a "knowing what it feels like" situation (knowing how to identify the smell of bacon) and in a "knowing that" situation (knowing how to get to British Museum); the linguistic constructions are not unique to just one variety of knowing.[22] If I "know how" to get myself from London to Paris all I need is a bunch of "knowing that"-type facts (apart from routine "knowing how"-type facts relating to the use of my body); but "knowing how" when used of the ability to play a sonata or ride a bicycle is very different from this. Knowing what it's like to taste truffle because I have had the experience certainly results in my "knowing how" to imagine the taste of truffle, but only because I then "know what it feels like" to experience it.

[22] Recent research by neuroscientists into the learning of "the Knowledge" (how to get around London) by London cabbies demonstrates that this has a similar effect on the neural connections to practicing the piano. The cabbies' hippocampus, the bit of the brain in which the new connections supporting such an activity are developed, was found to be significantly larger, on average, than those of the rest of the population (see Woollett and Maguire, 2011). This fact is also supportive of our proposition that there is a close relation between mental states and the material structure of the brain.

1.3.4 Relation between consciousness and memory

Consciousness has been linked with memory in academic writing for well over a century. Richet said, in 1886: "Sans **mémoire** pas de conscience" (1886, p 570). William James, writing in 1890, quotes Richet on the possibility of a direct connection between phenomenal consciousness, which he calls "conscious sensation", and memory (see James, 1890, Chapter XVI, footnotes 2 & 6). He says Richet stresses the fact that a sensation needs to persist for a short time to enable phenomenal consciousness to occur: " … for a conscious sensation … to occur, there must be a present of a certain duration, of a few seconds at least"; and: "to suffer for only a hundredth of a second is not to suffer at all; and for my part I would readily agree to undergo a pain, however acute and intense it might be, provided it should last only a hundredth of a second, and leave after it neither reverberation nor recall"(Richet, 1884, p. 32). More recently, Ray Jackendoff writes of "awareness" as being "supported by the contents of short-term memory" (1987, p 280). Francis Crick in his book *The Astonishing Hypothesis* (1994, p 14-15), notes that: "as long as one hundred years ago" an idea that was "already current" was that "consciousness involves some form of memory, probably a very short term one". Cowan (1997, p 77) says he sees: "Short-term memory taken as a whole, including both sensory and nonsensory aspects" as representing "the subject's present mind".

Psychologists, neuroscientists and philosophers have all speculated about the whereabouts in the brain of some neural correlate of consciousness and the possible nature of the mechanisms involved. Is it localised or dispersed? Could there be a specialised type of neuron or a specific neuronal mechanism underlying it (such as, for example, oscillatory or synchronized discharges, see Rees *et al*., 2002, p 266)? Is a satisfactory answer to such questions possible (Noë and Thomson, 2004)?

There are a number of suggestions for the neural correlates of consciousness in the literature, but strangely, none of them relates to short term memory stores.[23]

[23] Crick, in his 1993 book *The Astonishing Hypothesis*, suggests (p 251) that the neural activity associated with consciousness might be largely in the lower cortical layer (layers 5 and 6), which he says: "expresses the local (transient) results of computations taking place mainly in other cortical layers". Baars, in his 1988 book *A Cognitive Theory of Consciousness*, presents a possible structure of brain activity which he suggests may be relevant, calling it Global Workspace Theory. This proposes that there is a 'module', in a person's brain and that the material that is selected for inclusion in this module at any point in time is what the person will be "conscious of", or "attending to". The key aspect of Dennett's "explanation" of consciousness, in his 1991 book, is that it is a property that arises whenever there is a particular sort of physical structure. Dennett suggests that

Another relevant concept, of which people have been aware for many years, is 'attention'. It seems that we are presented with a situation in which there are many things that are immediately available to consciousness and we select from these what to focus on, what to pay attention to, at any point in time. William James (1890) famously refers to attention as "… the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought" (Chapter XI, p. 403). Sir Francis Galton (1883) describes it in the following way: "The ideas that lie at any moment within my full consciousness seem to attract of their own accord the most appropriate out of a number of other ideas that are lying close at hand, but imperfectly within the range of my consciousness" (p 203).

Baars (1997) has suggested that attention is akin to the spotlight that picks out some activity on a theatre stage leaving the rest of what is on the stage in the dark. In his words:

> "In the working theatre, focal consciousness acts as a 'bright spot' on the stage, directed there by the selective 'spotlight' of attention. The bright spot is further surrounded by a 'fringe,' of vital but vaguely conscious events …. The entire stage of the theatre corresponds to 'working memory', the immediate memory system in which we talk to ourselves, visualize places and people, and plan actions" (p 292).

If this is the case, and introspection certainly suggests that it may be, the sensory memory store must have what the psychologists call a "buffer" containing phenomenal information that is available to consciousness but is not attracting attention.

---

the structure in question must be a "virtual machine", and that the key feature it must have is what he calls a "Joycean" nature. This Joycean property refers to the "stream of consciousness" type of monologue used by James Joyce in his novels. In conjunction with Hameroff (see Hameroff and Penrose, 1996) Penrose has suggested that the microtubules within neurons might provide the brain with the hardware needed for quantum effects to take place, and therefore, in their view, for consciousness. These suggestions regarding quantum mechanics have been criticised by Tegmark *et al.* (2000); they argue that the time scale of neuron firing, and of excitations in microtubules, is out by more than one order of magnitude from what it would need to be in order for Penrose's theory to be plausible.

1.3.5 Structure of the mind[24]

'Memory' is defined in the *Stanford Encyclopaedia of Philosophy* by the statement: "'Memory' labels a diverse set of cognitive capacities by which we retain information and reconstruct past experiences, usually for present purposes" (Sutton, 2016). Psychologists have found that we have a range of memory stores, which vary in their properties and functions. These memory stores bear a close relation to the different types of thing of which we can be conscious, suggesting that if memories are stored in the material of the brain consciousness may be property of that matter.

In this section I sketch a structure for the mind in terms of a "central executive" postulated by psychologists to facilitate cognitive activities along with the different types of memory store they have identified and the type of consciousness to which each of these is related, see Figure 2 below.

Short term memory stores

The memory stores that we are mainly concerned with in this section are the short term memory stores where information is typically held, available to consciousness, for a period of less than a second, but once gone it cannot be recalled from this source.

It seems that we humans probably possess three different versions of short term memory stores of this type. I make a clear distinction in this thesis between things that can and things that cannot be communicated perfectly using language (Section 1.2.1). Given this distinction, which is not always stressed by psychologists, it seems likely that we have short term sensory memory stores which contain phenomenal information corresponding to our senses, and short term memory stores that contains public information; it seems likely that these will differ significantly in their material structure. All sensory experiences continue for a short period after the original stimulus has gone, being experienced precisely as if the stimulus were still operating.[25] There must therefore be

[24] The material in this section has been published in Rowlatt, 'Consciousness and Memory', 2009.
[25] The visual form of sensory memory store is called the 'iconic memory', the auditory form is called the 'echoic memory', and the olfactory form the 'odor memory'. The duration of sensations in the iconic and echoic memories has been studied in considerable depth in recent decades and it seems to be well-established that information remains in them for a minimum of a few hundred milliseconds. This was demonstrated originally by Segner in 1740. He attached a red hot coal to a cartwheel and rotated the cartwheel at a speed just fast enough for the light to form a complete circle to the eye in order to estimate the duration for which a signal remained in the iconic memory. Later, more complex, experiments have confirmed that duration (see Neath and Surprenant, 2003).

short term memory stores related to all sensory experiences; these would be memory stores that contain phenomenal information, the "short term sensory memory stores" in Figure 2, and they must store the information in phenomenal form, the 'what it feels like' aspects of that sensory experience.

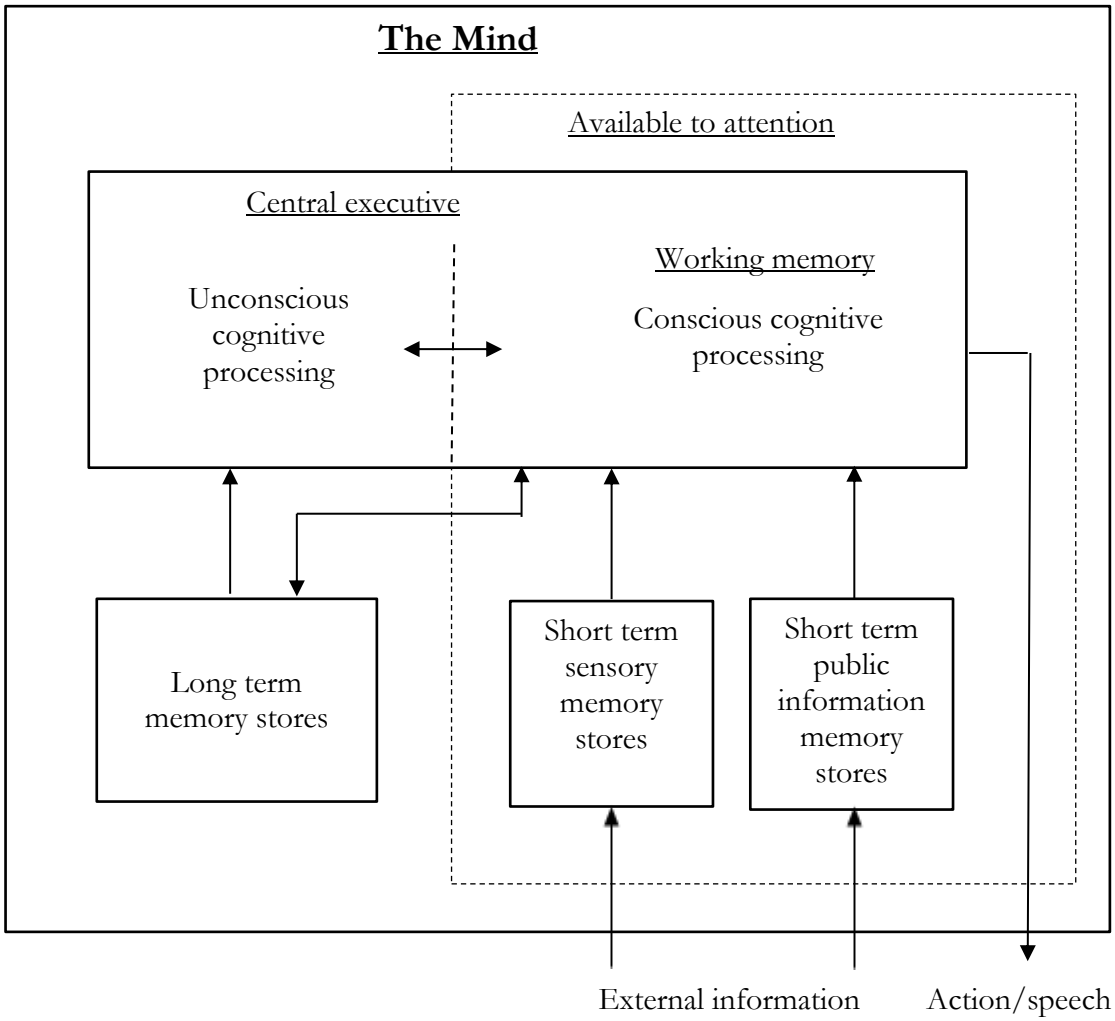


Figure 2: Consciousness, attention and memory stores

The figure shows a stylised arrangement of some of the postulated relations between consciousness, attention, and the memory stores referred to in the text.

The persistence of sensory perception has been confirmed a great many times and its characteristics have been studied in some depth for both visual and auditory experiences; it is not so easy to experiment with this phenomenon for other senses. Psychologists call memory stores of this type the "short term sensory memory stores" (I abbreviate this to 'sensory memory stores').

It is difficult to see how we could be aware of pain or itching, emotions, or proprioception if there were no sensory memory store associated with these, too. As William Burnham (1888, p 575) said: "Each sense … of the body may be said to have its memory …", but not surprisingly, no-one seems to have attempted the investigation of the likelihood of similar types of memory store for the internal senses.

As well as these sensory memory stores, associated with phenomenal consciousness, there must be short term memory stores that can contain public facts, the type of fact that can be expressed in words, that is, the type of fact associated with access consciousness (see Figure 2). Information of the type that can be expressed in words, and may be needed for cognitive or creative purposes, is likely to turn up from three sources:

- it may be recalled from the long-term semantic memory (see later); this will consist of public facts long known, such as "grass is green", personal factual memories such as "I had trout for lunch yesterday" or facts that have been learned from others, such as the likely causes of the French revolution;

- it may arise from recent experiences that have been processed and stored in the form of facts that can be communicated in words to others, for example, the words on a road-sign recently passed;

- it may be the result of cognitive processing, such as "it seems likely that it will rain soon".

These short term public information memory stores would be likely to have a different material structure from that pertaining to the sensory memory stores. Where information of this sort is available to consciousness it is said by psychologists to be in the 'buffers' of the conscious part of what psychologists call "the central executive" section of the mind (see Baars, 1997, and see Figure 2).

In addition, there must be another sort of short term memory store, known as the 'working memory' (see Figure 2). This is the name psychologists use for the short term memory where both public information and phenomenal information are used in thinking, cognitive analysis, deciding and such-like. It can therefore hold the beginning of a train of thought, or a sentence, or the early parts of the imagining of some incident, while the person's thoughts are moving on towards the end. We therefore assume that this memory store uses the information available for the other memory stores in order to arrive at decisions concerning speech and actions (see Section 1.3.2). The 'working memory' is a

term much used by psychologists, but with varying nuances of meaning depending on the particular author (see Baddeley and Hitch, 1974; in this field the names used for the various memory stores differ considerably according to the author involved, so care is needed). Again, the capacity of the working memory, the part of the central executive that is in the spotlight of consciousness, and the duration for which information remains within it, has been much researched.[26]

It would not be possible to create something in one's imagination, or follow a line of reasoning, or take a decision and implement it if one were unable to remember the early parts of some train of thought. As explained in Section 1.3.2, Block (2001) has suggested that there may a form of consciousness that could be termed 'reflexive consciousness', so this may be what he had in mind. If this is to be treated as a separate concept from phenomenal consciousness and access consciousness, there is a question as to where the information about these trains of thought are stored. Many psychologists take the view that this is a task that is performed by the working memory and that this is part of what they call "the central executive", see below.

The short term sensory memory stores are associated with feelings, that is, with phenomenal consciousness. The short term public information memory stores are associated with awareness of things, thoughts, that can be expressed linguistically. The working memory is generally associated with more complex thoughts of which some elements can be expressed perfectly linguistically; but it must also be able to contain feelings since it is where psychologists take it that the conscious parts of practical reasoning is performed and where creative thinking takes place.

Long term memory stores

There are other types of memory store besides these short term memory stores (see Atkinson and Shiffrin, 1968, for a sketch of a possible system of human memory stores).

[26] The capacity of the working memory that is in the focus of attention is surprisingly limited. Psychologists have found that it can hold only about seven items of information (plus or minus two) at any point in time (in an experimental setting these may be numerals, "chunked" numerals – 4-digit dates, for example – or short syllables). In order to hold information in this memory for any length of time you need to repeat it verbally, either aloud or silently; this is illustrated by the effort needed to remember the directions someone gives you to find a location, or the difficulty of remembering a telephone number as you cross the room. Without such reinforcement an item of information will remain in this memory store for less than a minute at most (see the classic paper by Miller, 1956). It is not yet clear whether the information in this memory store is replaced by new information or whether it simply decays over time.

Short term memories that are converted into long term memories are thought to end up stored in what the psychologists call the 'declarative memory'. Memories in this store may remain for a lifetime, although retrieval can be an issue. They are 'available to' consciousness in that they can, in principle, be recalled into the working memory store if needed. This long term memory store is therefore fundamentally different in nature from the short term memory stores described above: it is a 'recall' type of store.

It has been suggested that the declarative memory has two main subdivisions: the 'episodic memory store', which holds events involving a person's memories of external and internal sensory experiences and the temporal relations between them, that is, phenomenal facts (this is also known as the 'autobiographical memory'), and the 'semantic memory store' which stores knowledge about the meanings of words and the relations between them, that is, public facts (that Paris is in France, for example, or that driving a car too fast could result in a fine). This classification, due to Tulving (1972), is still not universally accepted.

There is another type of long term memory store where motor skills are stored. This has been called the 'procedural' memory by psychologists and is considered to be an 'implicit' memory store because its contents cannot be brought into consciousness. This memory store is well illustrated by the patient known as HM, whose case it was that led to the realisation that this memory store must be distinct from the declarative memory (Scoville and Milner, 1957). HM had crippling epilepsy. Following the removal of his hippocampus and the inner surface of his temporal lobes he lost the ability to transfer information from his working memory store into his long term (declarative) memory: he could no longer remember what had happened an hour ago. However, he was still able to learn a new skill and maintain that knowledge over the long term (Milner *et al.*, 1998). Hence there must be a second kind of long-term memory, independent of the hippocampus, where learning associated with skills is located. Neuroscientific investigation suggests that the cerebellum may be involved. Psychologists take the view that this (so-called) procedural memory contains knowledge of the "knowing how" variety, in the sense of knowing how to ride a bicycle (but as we have seen in Section 1.3.3, there are other ways in which the words "knowing how" can be used, and in some of these cases one *is* aware of how one is doing something).

The three types of long term memory store described above, that have been identified by psychologists and neuroscientists, therefore correspond to three of the types

of 'knowing' noted in Section 1.3.3: "knowing that" something is the case, "knowing what it is like" to experience something, and "knowing how" to do something.

The central executive

The hypothetical element of the mind that psychologists call "the central executive" (see Figure 2) is envisaged as a system that "facilitates a range of cognitive activities, such as reasoning, learning and comprehension" (Baddeley, 2003, p 829). But it also has to provide interfaces between information received by the senses (perception), information available from the long-term memory and the initiation of speech and action (Baddeley & Hitch, 1974). Indeed, Alan Baddeley remarks that: "In some ways the central executive functions more like an attentional system than a memory store" (1997, p 85).

Baddeley has identified a number of functions that this 'central executive' needs to fulfil (Baddeley, 2003, p 835). One role is to take the decision about what, out of the multitude of information that turns up in the sensory and working memory stores, should be stored in the long-term memory store. Then, although all the information in the sensory memory stores is viewed as being available to consciousness, what one is actually conscious of at any moment in time is determined by 'attention'. In Baddeley's view another role of the central executive is to decide which features of the scene before one, the sounds one can hear, the sights one can see and the complicated range of emotions one feels at any particular moment, should receive this conscious attention (see for example Baddeley, 1996, p 8). A third task that needs to be done is the selection of appropriate items from the long-term memory stores for availability to the working memory, given what is going on. If the central executive is to play the key role Baddeley envisages for it in cognitive activities, then it must be able call on information from both internal and external senses, both in immediate form and in 'visualised' form (drawn from the long term memory). It needs to be able to call on proprioceptive information (so that it can initiate action and speech) and on emotional information (so that it can assess whether some course of action is likely to increase or decrease welfare) – interestingly, Alan Baddeley says: "… it is useful to conceptualise action as ultimately steered by emotion" (2003, p 837), a statement that is consistent with the approach to decision-taking described in Section 1.4.3.

But of course, the processes involved in thinking extend far beyond these limited, but necessary, functions. Effective logical thought requires that a person can both identify logical relationships between the items that are focused upon, and use logical processes to

draw conclusions from a collection of related items; and effective creative thought must have another set of requirements involving aesthetic judgment.

1.3.6 Identifying conscious creatures

It seems that if a creature had no short term sensory memory stores, its eyes and ears (if it had such) would receive information relating to the surroundings and send this information in a message to the brain, but each bit of information would be gone a short moment after it arrived, so no record would be available for the creature to access and connect to the information received the following instant. Professor Richet says (1886, see Section 1.3.4): "Without memory no conscious sensation, without memory no consciousness". Nicholas Humphrey refers to the "Self" as being "phenomenally thick and substantial" and says: "a temporally-thick Self is something to build a rich subjective life on" (2006, p 131), perhaps making the same point. Without some form of memory for sensory information, it seems, the creature could not be 'aware' of the information it receives through its senses. Some of the experiments referred to in the footnote to Section 1.3.5 support this intuition; they demonstrate that interrupting the formation of the record of a signal in the sensory memory store can disrupt phenomenal consciousness; in these circumstances the subjects' reports reveal that they are unaware of the signal. This suggests that some form of memory associated with the receipt of sensory information is *necessary* for the experience of *qualia*, and therefore, for phenomenal consciousness; it also seems that some sort of memory associated with information that can be expressed in words is necessary for a person to have thoughts that involve words.

Given this, a question arises as to whether, if we knew how to identify the existence of such memory stores in some particular type of creature's brain, and we discovered that that particular type of creature had no memory stores of the relevant type, we would conclude that there could not be "something it is like" to be that type of creature.[27] Indeed, could we take it further and propose the obverse of this, that is, if we found that a creature *did* have memory stores of this sort, should we conclude that that creature *does* have consciousness?

The definition of consciousness that I have used for the purpose of this thesis (which derives from that of Nagel, 1974, and is: "a thing is considered to have conscious

[27] Obviously, if the only way of being conscious was to have this particular sort of a sensory memory store, this would rule out the possibility that there could be a computer that is conscious.

mental states if we take the view that it has the property of there being something it is like to *be* that thing – something it is like *for* the thing", see Section 1.3.1) is highly subjective in nature, in that it leaves room for dispute about which creatures are and which are not conscious. The thoughts presented in this section suggest that it might be possible for science, at some point in the future, to give us a more satisfactory, more objective (although likely to be controversial) indication of which things are, in fact, capable of consciousness.

1.4 Propositions relating to mental states

In developing the approach to the theory of mind presented in this thesis I found that I needed to make three assumptions. These support the approach I want to explore; if these propositions are valid the approach looks plausible (see Section 1.5). In this section I present the following three propositions as facts relating to mental states and, whilst they are clearly not true as the result of logic (definitions of the terms used), I suggest that their validity can be ascertained through introspection (see Section 1.2.3).[28] However, many people are not signed up to these propositions, and in Chapter 3 I examine the cases for and against other proposals regarding the ontology of consciousness.

1.4.1 Conscious mental states exist

Proposition 1: If we can think/feel we are able to experience conscious mental states, so we know that conscious mental states exist.

This proposition is a variant on Descartes' profound insight (1641), usually presented as "cogito, ergo sum" or "I think, therefore I am".[29]

If we can think, we know that we have experiences of the type that can be expressed in statements such as "this is what it is like to see red" or "this is what it is like to feel hungry", so we know that mental states exist. An implication of this is that we know that phenomenal facts, facts of the type "*this* is what it is like to experience X", exist

[28] Of course, what we think we introspect is what we are aware of introspecting (*cf.*, what we think we see is what we are aware of seeing) but it must be possible to misinterpret the significance of what we introspect (just as what we think we see may not correspond to the reality out there of what we are looking at), so care is needed.

[29] Apparently, Descartes first wrote this phrase, "I think, therefore I am", in French ("je pense, donc je suis"), in his *Discourse on the Method* (1637).

(see Section 3.4.2 for Lewis' argument against this proposition). The argument in support of this proposition is spelt out in more detail in Section 3.2.1.

1.4.2 There is an epistemic gap

Proposition 2: We cannot communicate to another person, in language, exactly what it is like to experience something.

So long as we don't sign up to solipsism we accept that there are phenomenal facts that relate to other people's/creatures' experiences.[30] So the question here is whether it could be possible for phenomenal facts (facts to which one person or creature has privileged access and which may depend on particular aspects of that person/creature) to be converted in some way into public facts (facts that can be communicated in words without any loss of information). If that is not possible there can be said to be an 'epistemic gap'.[31]

If it were possible to explain a phenomenal experience in terms of the causal processes suitable for study in the physical sciences it would be possible to express in words, precisely, what it feels like to experience something, for "what it feels like" would then be a concept that is capable of expression in terms of the properties of matter that are studied in physics, and all the concepts studied in physics can be expressed precisely in words/mathematics given our definition of physics (see Section 1.2.1). We would then, therefore, be able to express what it is like to experience something in the vocabulary which we use for concepts relating to the physical sciences.

But we know that we cannot do that. If someone were to try to tell you what it is like to taste truffle when you haven't previously tasted it, you know that they will not succeed. It seems that it is impossible to explain phenomenal experience in terms of the concepts of the physical sciences.

The argument in support of the existence of an epistemic gap is spelt out in more detail in Section 2.4. The case for the existence of an epistemic gap along with ontological reduction of the mental to the physical is made in Section 3.3.4.

---

30 Taking solipsism to be the metaphysical position that other minds do not exist.

31 The 'epistemic gap' is related to the 'explanatory gap' (see Levine, 1983). 'Epistemic gap' refers to the fact that there is a class of facts, phenomenal facts relating to other people or creatures, that we can never know. 'Explanatory gap' refers to a problem that arises in certain physicalist approaches to the theory of mind: the possibility that it is impossible to give a wholly satisfactory explanation of mental states in terms of the physics of the brain. The relation between the two is addressed in Section 3.3.4.

1.4.3 Feelings are necessary for the taking of certain types of decision

Proposition 3: When we take a decision between alternative possible actions we pick the one we judge would lead to the most desirable outcome, all things considered.

When faced with certain types of decision regarding actions, decisions that involve consequences that cannot be directly compared, the chooser examines the way he or she *feels* about the likely consequences of choosing each of the alternative options available in order to identify the action that is likely to result in the most desirable outcome, all things considered.[32] This process is described by philosophers as 'practical reasoning' and the feelings involved in practical reasoning, which often involve events that affect other people and have consequences which are risky in different ways and that take place at widely varying points in the future, are generally viewed as being related to the person's 'values' (the dispositions people have to experience certain feelings in certain particular circumstances).[33] What a person needs in order to take decisions of this sort effectively is a mechanism that can enable the "desirability" of different bundles of non-comparable properties to be gauged one against another.[34]

For the complex decisions that involve practical reasoning a type of process that is altogether different from the causal processes described in Section 1.2.2 is required. In this situation, what is needed is a process that, in the simplest possible terms, can be described by:

$$\text{choose } A_i \text{ to maximise (or satisfice) } F_i(E_i^1, E_i^2, \ldots, E_i^m)$$

where the set of available options, given the situation that pertains, is $\{A_1, A_2, \ldots, A_n\}$, the set of effects perceived as likely to follow the choice of option A_i is $\{E_i^1, E_i^2, \ldots, E_i^m\}$ and $F_i(..)$ can be described an indicator of the desirability of the perceived set of effects.

The effects of the A_i *cannot* be directly compared because they can be quite different; perhaps one involves the remote possibility of something very beneficial for

---

[32] The fact that certain rather trivial timing decisions have been shown by Libet and others (see Libet *et al.*, 1982, for example) to have been taken before the person concerned is aware of having taken them does not necessarily have the implication that every decision is taken in that way (see Section 2.3.2).

[33] As Watson says (1975, p 346): "The valuation system of an agent is that set of considerations which, when combined with his factual beliefs (and probability estimates), yields judgements of the form: the thing for me to do in these circumstances, all things considered, is a".

[34] There is some empirical support for the theory outlined in this section, see Section 4.3.

survival or procreation and another the certainty of something totally different which has some small positive value to the chooser (a pleasant taste, say).[35] The F_i are viewed as indicating the 'desirability' of the i'th bundle and it is assumed that the F_i *can* be directly compared (psychologists refer to the feelings associated with the 'desirability' of the perceived outcomes as 'motivational affect', see Sections 4.3 and 5.3.2). In practice, however, the end result of practical reasoning is often viewed by the chooser as being a "gut feeling" that one particular action is "probably the best one to choose".[36]

This issue is discussed in more depth in Section 2.4.3 and in Chapters 4 and 5. The implication of the fact that it is a person's *feelings* that determine what action they choose to implement in a choice situation is that the ability to experience feelings is *necessary* for the taking of this sort of decision; somehow, the information contained in the feelings has to be transmitted to the physics that underlies the action, see Section 2.5.[37]

1.5 Justification for 'causally effective property dualism'

In this section I spell out the essential steps in the reasoning that has led me to the view that causally effective property dualism provides a reasonably plausible explanation of the phenomenon of consciousness, and that it should therefore be taken seriously.

My starting point is the question of the definition of physics. I could find no clear definition in the literature (but see Section 1.2.1). The definition of physics, and therefore of the physical sciences, that I use in this thesis (see Section 1.2.1) follows from the requirement that in order to be acceptable to physics an experiment must be reproducible: it must be possible to communicate in language all the details of the manner in which an experiment is performed, along with its findings, so that somebody else, perhaps on the

[35] Of course, the desirability of some bundle, as indicated by the way one individual feels about it, may be quite different from its desirability as perceived by some other individual.

[36] This 'desirability' or 'motivational affect' is generally viewed as depending on the probability distribution of the 'value' (or 'utility') associated with the bundle. It depends on the discounted expected value, of course, but it is also affected by the risk of the actual value turning out to be different from that discounted expected value. Markovitz (1952, p 77) explains why we should reject: "the rule that people maximise discounted expected, or anticipated returns. … both as a hypothesis to explain, and as a maximum to guide investment behaviour", and says that we should take the variance into account. Kahneman and Tversky (1979, p 263) present: "… a critique of expected utility theory as a descriptive model of decision making under risk, and develop an alternative model, called prospect theory"; this can be viewed as taking account, also, of the third moment of the probability distribution, the skew.

[37] There remains a question as to whether it would be possible for a decision such as this to be taken by the type of process studied in the physical sciences, see Section 2.4.3.

other side of the world, can reproduce that experiment and check that their findings are the same as those of the originator. It is therefore necessary that all the facts relating to a property of matter that is suitable for study in the physical sciences can be communicated perfectly from one person to another in language.

Next, we know from introspection that we are unable to communicate in words exactly what it is like to experience a feeling (see Section 1.4.2); that there is an "epistemic gap". Given our definition of physics, this appears to suggest that mental states are not instantiations of some collection of the properties of matter that are studied in physics; it suggests, therefore, that ontological reduction may not hold. If this were the case, and if mental states are instantiations of a property of matter, there would then have to be another property of matter besides those studied in the physical sciences. However, many philosophers take the view that ontological reduction combined with the existence of an epistemic gap is a viable proposition (see Section 3.3.4 for a summary of this view), so reductive physicalism remains a possible explanation of the existence of mental states. In this thesis, however, I look at the case for there being another property of matter, different from those studied in the physical sciences, that instantiates mental states.

The next question, then, is whether this property of matter, mind (the existence of which I am postulating in this thesis), instantiations of which are known as mental states, could be causally effective *in its own right*, as can the other properties of matter (those suitable for study in the physical sciences). For if mind is a property of matter it could be the case that whenever a particular mental state arises some characteristic state pertains amongst the *matter* involved (the arrangement and type of the atoms involved *etc.*; presumably the matter involved is the neurons and such-like in the brain).[38] If that were so, some particular set of physical properties that is also associated with that particular state of the *matter* would also arise whenever the particular mental state arose, so it would be impossible to have the mental state without also having the relevant physical state and *vice versa*; it would be the case that you couldn't have one without the other. But how could we then, ever, discover which of the mental and the physical is actually "doing the causing" when a person takes a decision and implements it? If you speak to a neuroscientist or a physicist, or even a philosopher, about mental causation, and agree to take it that the mental is a property of matter and that you can't have some specific mental state without

[38] See Figure 1b; 'matter' is viewed as being a substratum that instantiates two sorts of properties, mental properties and properties suitable for study in physics. Here we are taking it, for the state of argument, that mental states are not multiply realised.

the associated, specific, physical (non-mental) state and *vice versa*, he or she will immediately assert that: "it is the physical state, not the mental state, that does the causing". But is it?

I claim here, see Section 1.4.3, that we know, from introspection, that when we choose to go for a walk instead of going to the library, or decide to climb Everest instead of remaining safe at home, we choose the action we do because of the way we *feel* about the situations that we think may arise as the future unfolds given the various options available regarding our actions. I claim in this thesis that we know, from introspection, that *feelings* (*e.g.*, Watson's 'valuation system', see footnote in Section 1.4.3) are necessary for practical reasoning, and therefore, since consciousness is necessary for experiencing feelings, we know that it is the person's ability to experience mental states that enables the identification of the action which he or she *feels* is most desirable, all things considered. Somehow, the way the person *feels* about the expected outcomes of the available actions must affect the choice he or she makes, and implements, regarding which action to perform; the mental state must be able to influence the physics of the action. This assertion gains support from the observation that if the part of the brain thought to be involved in assessing the relative desirability of the available actions (the basal ganglia) is damaged in an accident, or following surgical intervention, there appears to be no effect on the cognitive functions, but the decisions the person takes thereafter appear random, rather than reflecting the value a normal person would put on the consequences of the alternatives available.

It therefore seems that feelings are necessary for effective decision-taking. Consciousness, with the ability that it bestows to experience feelings about alternative bundles of expected outcomes and to compare them, appears to be the "gadget" that evolution has found that enables decisions of this type to be taken effectively; and because of this, consciousness is a great asset for enhancing the likelihood of survival. It is worth noting that the form of the law of nature that describes practical reasoning, spelt out in Section 1.4.3, is quite different from the form usually taken by the laws of physics, see Section 1.2.2, and maybe this adds an element of plausibility to the judgment that it is necessary that there exists something different from the properties of matter studied in physics, something along the lines of a mental state, if it is going to be possible for decisions of this sort to be taken.[39]

[39] Even neuroscientists and physicists have been known to balk at the proposition that the type of process studied in the physical sciences can perform something that amounts to a form of cost-benefit analysis.

At present, we can only make guesses regarding the mechanism by which the taking of a decision might affect the neurons so that the appropriate action is initiated (see Section 2.5), but if the above justification is accepted, it would be the case that it does.

So, I conclude that, since feelings are necessary for practical reasoning, it must be the mental state that is responsible for the action chosen, *not* the physical (non-mental) properties that are *also* instantiated by the specific situation regarding the neurons; it seems that the 'role' of consciousness is to facilitate survival by enabling effective and efficient decision-taking (see Chapter 4). And finally, with mental states causally effective we do, of course, have free will of a type, and there is therefore a sense in which a 'person' can be said to be morally responsible for the effects of his or her actions (see Chapter 5).

1.6 Structure of the thesis

With this as background, I start by looking (in Chapter 2) at the implications for physics and for science generally of the proposition that mind is a non-reducible, causally effective, fundamental property of matter, with an ontological status similar to that of the properties of matter studied in physics as the result of a law of nature. Then I spell out what this assumption implies for the philosophers' three anti-physicalism arguments: the knowledge argument, the conceivability argument and the hard problem. It turns out that whereas each of these is thought to cause problems for reductive physicalism, none of them need cause any problems for a materialism in which a law of nature requires that consciousness is a property of matter different from those studied in physics, each for quite simple and straightforward reasons.

In Chapter 3 I examine the alternative approaches to the nature of mind, summarising the *pros* and *cons* for the three main alternatives, the dual substance and dual "aspect" approaches, the identity thesis/reductive physicalism approach and the approach known as non-reductive physicalism, including supervenience physicalism and the form of property dualism explored in this thesis. This clarifies the reasons why it seems sensible to examine our causally effective property dualism assumption regarding the nature of mind in some depth. We also investigate how this proposition regarding mind fits with some of the multitude of issues, such as "multiple realisation" or the rejection of so-called "over-determination", which have arisen and been discussed during the last few decades in this section of the philosophy literature.

Then we can examine the implications this simple assumption has for other important issues. We can see whether it explains why conscious creatures have taken the world by storm in the way they have (Chapter 4), addressing the question of the role of consciousness. We can ponder on the question of whether, with this assumption regarding the nature of mind, we can justify the proposition that humans, and maybe other creatures, possess something that could be called 'free will', so that we can consider ourselves to be morally responsible, in some sense, for what we decide to do (Chapter 5).

Chapter 2 Mind as a property of matter

2.1. Introduction

In this chapter I present the case for causally effective property dualism. The chapter starts with a description of this approach to the ontology of mind along with a brief description of the most similar approaches in the literature (Section 2.2). The arguments for and against the key issue of the causal completeness of physics are spelt out in Section 2.3 and 2.4; if feelings are necessary for the taking of certain types of decision, as implied by Section 1.4.3, it has to be case that physics is not causally complete. In Section 2.5 I note some possible (although rather fanciful) approaches to the process by which mental causation might operate in practice. This is followed, in Section 2.6, by a discussion of the three anti-physicalist arguments, the knowledge argument, the conceivability argument and the hard problem, in the light of our assumption concerning the nature of mind.

2.2 Causally effective property dualism

2.2.1 What is property dualism?

The approach to the nature of mind that is explored in this thesis assumes that there is just one kind of 'substance', which we call 'matter' (see Section 1.2.4), and that mind is a property of this substance called matter, a property that is different from the properties of matter that are studied in the physical sciences but has a similar ontological status to those properties (see Figure 1b in Section 1.1). Further, I suggest (Section 1.4.3) that there are reasons, based on introspection, for taking seriously the possibility that mental states can be causally effective *per se*, and I assume that the interactions mental states have with the properties of matter studied in physics are governed by causal processes so either determinism or universal probabilistic causation (see Section 1.2.2 and Chapter 5) holds. This approach, which I call 'causally effective property dualism', is consistent with science being 'complete' in the sense that it would encompass all the causes and all the effects, mental as well as physical, of all the causal processes including, therefore, all mental/physical causal interactions.

The fact that this approach has mental states not reducible to the properties of matter studied in the physical sciences has the implication that, just as there are facts relating to the properties of matter that are studied in the physical sciences such as the direction of motion and the momentum of some body, or its mass or charge, so there are facts of the "what it feels like" type (what it feels like for a person to see red or to be hungry) known as phenomenal facts or *qualia* that relate to this other property of matter (see Section 1.3.2). And just as there are causal processes relating to the properties of matter that are studied in the physical sciences, so one might expect there to be causal processes regarding the effect the properties of matter studied in physics have upon mental states and regarding the effects mental states have on the properties of matter studied in physics (see Sections 2.4 and 2.5).

2.2.2 Property dualism versus supervenience physicalism

The ontological difference between property dualism and supervenience physicalism is small, and probably of little real significance. In both supervenience physicalism and property dualism, mental properties are instantiated by some particular configuration of matter (some particular arrangement of a number of different types of particles *etc.*), and this configuration of matter will also instantiate some particular collection of physical properties, see Figure 1. In both cases we would therefore expect to see a correlation between some mental property and a set of physical (non-mental) properties.[40]

With supervenience physicalism, however, a change in a mental property *requires* that there is also a change in the set of physical properties. But in the case of property dualism it is possible that there could be a change in the mental property, and therefore a change in the matter, but no change in the physical properties because it is logically possible that two configurations of matter that instantiate different mental properties, could both instantiate the precise same collection of physical properties[41] This difference between supervenience physicalism and property dualism is more obvious in the case of the bundle

[40] Of course, much of the matter associated with a brain state, along with its physical properties, will be irrelevant to the mental state. Also, it may be that two or more mental states that arise in association with different configurations of matter could be indistinguishable to the person concerned.

[41] This might arise if there were two varieties of some element of the collection of matter that instantiated a mental state which were the same in every way except regarding the aspect relating to the mental state to which they, in combination with the rest of the matter, gave rise; just as the charge, or the spin of a fundamental particle can differ. So two bits of matter might differ in their mental properties but not in their physical (non-mental) properties.

theory approach (see Section 1.2.4). If supervenience physicalism were the case, the same mental property would always arise in the context of some particular bundle (or bundles) of physical properties; in property dualism, however, there could be two bundles of properties that were identical in their physical properties but differed in their mental properties.

The more important difference between supervenience physicalism and property dualism is simply that if mental properties have the same ontological status as physical properties we would expect mental causation to hold, other things equal, because all the other properties of matter are causally effective.

2.2.3 Evidence in support of some form of physicalism

Increasingly neuroscience has been discovering that when a person has a subjective experience of some particular mental state, there is some particular state of activity in their brain, a development that is clearly supportive of the proposition that mental states are instantiations of a property of the matter in the brain. A person (A) using an electroencephalograph (EEG) can observe on a screen the brain state of another person (B) and through these observations may be able to identify certain aspects of their mental state. For example, it is already well-established that if A observes that the fusiform face area of the brain (FFA) is active, then person B will be considering the features of someone's face, and a great many similar correlations have been established between brain areas and conscious experiences.

Indeed, we now know enough about the structure of the human brain to enable direct brain-to-brain communication to take place; a very simple message has been sent from the brain of a person in Kerala (extracting the information by electroencephalogram, EEG) to the brain of a person in Strasburg (inducing conscious reception by robotised transcranial magnetic stimulation, TMS) with email wired to carry the information, automatically (using the Morse Code!), across the globe (see Grau *et al.*, 2014). Further, recent experiments at MIT (Louie & Wilson, 2001), found that patterns of brain activity identified when rats explored a maze during the day were exactly replicated when the rats were sleeping at night; the experimenters reckoned that they knew what the rats were dreaming about.

The proposition that there is a relationship between a person's brain activity and the mental state he or she is experiencing is thus pretty well established now as a fact that

can be observed; and this is just what one would expect if mental states were, indeed, a property of the material in the brain.

2.2.4 References to property dualism in the literature

In many ways causally effective property dualism seems the simplest and most obvious assumption to make regarding the nature of mental states (states of which it is possible for a creature to be conscious), that is, regarding the nature of mind, although few philosophers have supported this proposal in recent years. Smart, in 1959, must have perceived the possibility that consciousness might be 'merely' another property of matter and he rejected it in favour of the identity thesis when, speaking of "states of consciousness", he wrote:

> "I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles) all put together for all the world as though their main purpose in life was to be a negative feedback mechanism of a complicated sort" (p 143, and see Section 3.3.3).

I suggest in this thesis that the approach that Smart is rejecting here, that mind is a property of matter that relates to neurons and that it is a form of negative feedback system (a form of homeostasis, see Section 4.4), deserves serious consideration. Galen Strawson (2006) also must have had something along the lines of this approach in mind when he rejected the approach Smart was advocating back in 1959, writing: "… real physicalism can have nothing to do with *physicSalism*, the view – the faith – that the nature or essence of all concrete reality can in principle be fully captured in the terms of *physics*" (p 4, his italics).

There are three approaches in the literature that come close to the causally effective property dualism presented here. First, there is Chalmers (1996). Chalmers calls his approach 'naturalistic dualism' (Section 3.4.6 contains more details regarding this approach). He takes it that there must be something other than that which is studied in physics: "The physical facts incompletely constrain the way the world is; the facts about consciousness constrain it further", he says (1996, p 123), and he finds it plausible that if, for example, his own physical [material?] structure was precisely replicated, his conscious experience would be replicated also, as would be the case if mental states were instantiations of a property of matter that is different from those studied in physics. But Chalmers says, on page 124: "… it remains plausible that consciousness supervenes

naturally on the physical. It is this view – natural supervenience without logical supervenience – that I will develop". The difference between 'supervenience physicalism' and the type of 'property dualism' that I am exploring in this thesis is illustrated in Figure 1 (in Section 1.1, see also Section 3.4.7). The reason why this difference between my approach and Chalmers' naturalistic dualism matters is because my assumption, that mind is a property of matter different from those studied in physics but of similar ontological status, lends itself naturally to a causal role for mental states (all the other properties of matter are causally effective so why not this one), while Chalmers' approach, assuming that mind supervenes 'naturally' on the physical, is usually interpreted as lending itself to the assumption that, as Chalmers says: "the physical domain is causally closed" (1996, see p 161). Chalmers therefore takes it that epiphenomenalism pertains: that there is no causal role whatsoever for mental states regarding the physical world.

The second approach with similarities to causally effective property dualism is functionalism; the only approach in the recent literature that does not deny that something that could be called "effective mental causation *per se*" exists (see Section 3.4.8). But the heart of functionalism is that the key feature of something that can be called a 'mental state' is the *function* it plays; so effective mental causation *per se* is a given with functionalism. However, in functionalism a mental state is, *by definition*, that which performs a certain function (rather than being the feelings and thoughts which are normally associated with mental states) and this leads to a range of somewhat fanciful consequences (see Section 3.4.8).

Finally, McGinn presents an excellent description of the case in which mental states are instantiations of a property of matter that differs from, and is not reducible to, the properties of matter studied in physics (1982, pp 27-33). But he makes no reference to mental causation, and does not identify this approach as having any advantage over the other approaches he considers or as being the one that should be pursued.

2.3 Arguments for the causal completeness of physics

We now take a look at the case for the causal completeness of physics. As we have seen, it is often asserted that mental states *per se* can play no causal role in the physical world, that is, that physics as we see it today, excluding mental states, is causally complete (in the sense that all physical effects are fully caused by purely *physical* prior histories), so everything that happens is subject to causal processes that are suitable for study in the

physical sciences and therefore exclude mental states. I suggest in this thesis that there is a certain class of physical events (the physical consequences of complex decisions taken by sophisticated creatures, see Section 1.4.3 and Chapter 5) that depend on *feelings* and that therefore these can only take place if conscious mental states can have an effect on physical (in the sense non-mental) events; I suggest that it cannot be the case that the physical correlates of those feelings are doing the causing. If this is correct, then if the world did not include whatever it is that enables creatures to experience conscious mental states, these creatures would have been unable to take these complex decisions in the way they have and a great number of the events that have taken place in the world would never have happened.

The causal completeness of physics is taken, in philosophy, to be the thesis that "all physical effects are fully caused by purely *physical* prior histories" where the term 'physical' is interpreted here as meaning "pertaining to the science of physics" (rather than "pertaining to matter"). It therefore does not rule out the possibility that things exist that are outside physics (*e.g.*, mental states, as in non-reductive physicalism); so although it ensures that physics contains all the causes of physical events, it allows physical properties to interact causally on a one-way basis with things that are outside physics, in the sense of being not suitable for study in the physical sciences (if such things exist).

Also, although "the causal completeness of physics" thus defined appears at first sight to rule out the possibility of mental causation, and it often seems to be thought that this is the case, it does not actually do so. A physical (non-mental) event could be caused by an event that has mental characteristics, as well as having physical (pertaining to physics) characteristics (see Figure 5 in Section 3.4.5), and we humans could take the view that there are two causal processes, one with a mental cause and the other with a physical cause (and a purely physical prior history), each of which could fully explain the occurrence of the physical event (on a regular basis, not as a one-off coincidence), just so long as there is also a law of nature ensuring that if one of these causal processes operates the other one does too. This situation is often ruled out by philosophers as being something called 'over-determination', which is deemed unacceptable, see Section 3.4.5; but dual causation of this sort is not obviously unacceptable. Laws of nature encapsulate our understanding of the causal processes that operate in the world, and there is no reason in logic why a causal process should not operate in this way.

2.3.1 The success of physics

Probably the most convincing reason for signing up to the proposition that physics is causally complete in the above sense is the incredible success that science, and physics in particular, has had during the last couple of centuries in explaining so many things that have puzzled people during previous aeons. As bastion after bastion of unknowns and questions fall to the approach of the physical sciences it seems that this process is unstoppable; surely a day must come when the physical sciences explain the occurrence of every event in terms of physical (non-mental) properties of matter and the laws of nature relating to them?

However, the scientists who have developed physics as we know it today have made no attempt to incorporate decision-taking by creatures into their discipline, as an effective causal phenomenon, in spite of the fact that it meets all the criteria of good science: every time I decide to tap my finger on the table my finger makes the movement I intended. In any other branch of science one would expect the decision to be taken seriously as a possible cause of the repeatable event. However, as I understand it, although the Standard Model, the structure the physicists have developed in recent years that classifies fundamental particles and forces, has no obvious gap that suggests that something might be missing, there are still a number of unexplained phenomena which lead physicists to postulate new particles or forces; in addition, there could be a role in quantum mechanics as a result of which consciousness could influence outcomes (see Section 2.5.2). So there is still room for more developments when neuroscientists, or physicists, apply their scientific expertise to the minutiae that arise within the brain and actually seek to identify such an effect.

2.3.2 The brain decides before we are aware

Evidence that at first sight looks to be supportive of the causal completeness of physics comes from Libet's famous experiment on the timing of decisions. Experiments performed by Libet and his colleagues (see Libet, Wright & Gleason, 1982), and repeated a great number of times since, famously including the work of Soon *et al.* (2008), show that people can take certain types of rather unimportant decisions concerning the timing of actions a significant time before they are consciously aware of having taken them, suggesting that consciousness, and therefore conscious mental states, might have had no role in initiating the action.

In these experiments subjects are seated; they are asked to make a movement with their finger at a time of their own choosing and to note the position of a rotating dot on a clock-face at the time when they are first aware of having taken the decision to make the movement. They are often, but not always, instructed "to let the urge appear on its own at any time without any pre-planning or concentration on when to act" (Libet *et al.*, *ibid.*, p 324). The time of the actual movement is determined by means of an electromyogram connected to the subjects' wrists. The start of neuronal activity associated with the movement, the initial rise of the so-called 'readiness potential', is obtained by means of an EEG (electroencephalogram). The experiment shows that the readiness potential begins to build up approximately 350 milliseconds (a third of a second) *before* the subject is aware of the intention to move his or her finger. The implication often drawn from this is that if these findings regarding the gap between the build-up of the readiness potential and the awareness of the decision apply to this one type of decision they must apply to all other decisions, and that therefore consciousness must be epiphenomenal and free will must be a mere fantasy.

But it does not follow that Libet's result, although clearly true of this particular type of decision in these circumstances, is therefore necessarily true of all other types of decision. We already know that there are decisions, triggering actions, which are taken without being initiated by conscious thought, such as those that arise from activity in the procedural memory (like riding a bicycle, or playing a familiar piece of music on the piano, see Section 1.3.3); there is also the multitude of decisions that are regularly taken on automatic pilot (see footnote in Section 1.3.1). Libet seems to have identified another class of such decisions; or perhaps it is just a particular case of a decision taken on automatic pilot. In any case, this does not rule out the possibility that conscious mental states *are* necessary for the causation of certain classes of action; for a start, it may that the decision prior to the start of Libet's experiment, that one will *enable* decisions relating to the timing of actions to be taken without conscious intervention, may itself need to be taken consciously, and it seems likely that conscious thought is needed to initiate actions when one is beginning to learn to ride a bicycle or play the piano. These experimental findings do not amount to a convincing case for the lack of mental causation, and therefore free will, nor do they establish the causal completeness of physics.

It is salutary to note that the substance of Libet's result was already known to William James (and probably to many other people; certainly, I discovered this useful trick

myself as a child). Long before Libet appeared on the scene, James wrote the following enjoyable piece in his book entitled *The Principles of Psychology* (see Chapter 26, 1890, p 524):

> "We know what it is to get out of bed on a freezing morning in a room without a fire, and how the very vital principle within us protests against the ordeal. Probably most persons have lain on certain mornings for an hour at a time unable to brace themselves to the resolve. We think how late we shall be, how the duties of the day will suffer; we say, "I must get up, this is ignominious," etc.; but still the warm couch feels too delicious, the cold outside too cruel, and resolution faints away and postpones itself again and again just as it seemed on the verge of bursting the resistance and passing over into the decisive act. Now how do we ever get up under such circumstances? If I may generalize from my own experience, we more often than not get up without any struggle or decision at all. We suddenly find that we have got up. A fortunate lapse of consciousness occurs; we forget both the warmth and the cold; we fall into some revery connected with the day's life, in the course of which the idea flashes across us, "Hollo! I must lie here no longer" – an idea which at that lucky instant awakens no contradictory or paralyzing suggestions, and consequently produces immediately its appropriate motor effects. It was our acute consciousness of both the warmth and the cold during the period of struggle, which paralyzed our activity then and kept our idea of rising in the condition of wish and not of will. The moment these inhibitory ideas ceased, the original idea exerted its effects."

Indeed, the mere fact that one is aware that these decisions are being taken without conscious intervention, and that this is worth remarking on (*eg.*, by James), suggests that this is not the case with many of the other decisions that we take.

2.3.3 Conservation of energy

Papineau has a good account of the historical development of the idea that physics is complete in his book *Thinking about Consciousness* (Papineau 2002, Appendix, see also Papineau, 1990). In it he presents a line of thinking, additional to that of the inexorable march of physics, which might have convinced people that physics is, indeed, causally complete; although it does not come close to establishing beyond doubt that this is so. It concerns the principle of conservation of energy.

The law regarding the conservation of energy is the generally accepted principle that energy cannot be created or destroyed through physical processes (processes subject to the laws of physics) although the form in which energy is instantiated can change:

kinetic energy can become potential energy and *vice versa*, for example, and mass can become some form of kinetic or potential energy and *vice versa*. Papineau points out that if this principle applied *only* to activities relating to physics, then if mental states can cause physical (pertaining to physics) events in the brain to take place, leading to actions, this would mean that mental states must be able to affect the motion of some electron or the outcome of some chemical reaction in the brain; this would inevitably involve a transfer of energy and therefore it would contradict the principle of conservation of energy. Does the principle of conservation of energy therefore rule out the possibility of mental states being causally efficacious in this way?

First, suppose that mental states are part of the material world, instantiations of a property of matter that is not studied in the physical sciences, as in the case examined in this thesis. In this case it would seem reasonable to expand the principle of conservation of energy to include *all* the properties of matter, those not suitable for study in physics as well as those that are suitable for study in physics. Then, so long as all the causal processes relating to mental states were consistent with the conservation of energy, the changes in potential or kinetic energy relating to mental causation at any point in time would simply be part of the totality of changes in the flows of energy taking place in the material world at that point in time; and all of these, taken together or separately, would be consistent with the conservation of energy; activating a decision would involve a transfer of energy from the mental state to the electron or chemical reaction that triggered the action. It would not have to be the case that the conservation of energy applied only to activities relating to physics. It is difficult to see any reason why conservation of energy should rule out the possible existence of a property of matter, not presently included in science, with causal processes relating to the way it operates (but for which there is not as yet adequate empirical support). So long as any force that it is proposed exists is part of science (subject to causal processes that may be deterministic or probabilistic), as in the case considered here, there is no reason to suppose that it would break the appropriate conservation of energy principle – a point acknowledged by Papineau (see p 249).[42]

[42] In this context we should note this statement in Henry Marsh's excellent book on neurosurgery: "One quarter of the blood pumped every minute by the heart … goes to the brain. Thought is an energy-intensive process" (*Do No Harm*, 2014, p 41). Kahneman (2011) also is impressed by the amount of energy thinking requires, saying: "… effortful mental activity appears to be especially expensive in the currency of glucose" (see p 43). While Ashcroft (2012) says: "The brain alone uses about 10 per cent of the oxygen you breathe to drive the sodium pump and keep your nerve

Next, suppose that substance dualism holds (see Section 3.2), and that mental states are outside the material world. Then, if mental causation were to hold it would be as if some cause outside our material world is making an electron move in a way that it would not otherwise have moved, or making a chemical reaction reach a conclusion that it would not otherwise have reached. Indeed, given the current state of knowledge, we cannot rule out the possibility that there is something mysterious and beyond our understanding that causes physical (pertaining to physics) events to take place in our material world (see Section 5.2.3, on libertarianism). If some inexplicable intervention caused a mass to accelerate (see the approach to laws of nature in Nancy Cartwright, 1983) we could calculate how large the force would have had to be, had physics as we know it applied, using the law of physics "force equals mass times acceleration":

$$F = ma$$

although the force in question could in this case be outside the science of physics. What would be happening in such a situation is that the total energy in the physical universe would be increasing (unless there were sinks occurring somewhere else). It is not the case that such a situation would run counter to the principle that energy cannot be created or destroyed through processes that are subject to the laws of physics. It seems highly unlikely that it will prove to be the case that creatures that are conscious emit more energy than they absorb, as would be the implication of mental states being a source of energy in this way, but I doubt if anyone has explicitly tested such a hypothesis with sufficient accuracy for the point to be established.[43]

As yet, there is no direct empirical evidence of a force that is related to mental causation. But it is not beyond all credibility that it might turn out to be the case that someday there will be convincing evidence for the existence of forces not currently studied in physics (see Section 2.5.2), or even mysterious interventions beyond our current understanding, implausible though that might seem.

---

cell batteries charged. Perhaps somewhat surprisingly, it seems that merely thinking is energetically expensive." (p 40). None of which carries the implication that the energy is spent on anything other than the phenomena already included in the science of physics, of course.

[43] Papineau (2002, p 252) recounts an experiment, performed in 1889, in which the energy flows relating to a small dog were measured and it was found that the energy the dog expended was *exactly* equal to the energy contained in the food it consumed (given the accuracy of such experiments at that time). If such an experiment could be properly validated it would clearly support any case someone might wish to make against the dual substance approach regarding mental causation.

2.3.4 Conclusion

Of course, the fact that the causal completeness of physics, defined in the manner described in this section, is as yet unproven is no reason to reject the proposition that it is the case, and it should be noted that among philosophers there is, currently, a broad consensus concerning its validity. I now move on to present the case *against* this proposition that almost everyone seems to think is correct.

2.4 The case against the completeness of physics

It can only be the case that physics is not complete in the sense noted above ("all physical effects are fully caused by purely *physical* prior histories", where 'physical' means pertaining to the physical sciences) if something that is not ontologically reducible to the properties of matter that are studied in the physical sciences exists and, in addition, if that something is causally effective; that is, if events involving facts that are studied in the physical sciences can be caused by events involving facts that are *not* included in the physical sciences. So the questions to be addressed in this section are the following. First, do mental states exist? Second, how convincing is the proposition that there could be an ontological gap, that is, that mental states could be a property of matter different from those studied in physics, and therefore how convincing is the proposition that there is an epistemic gap (clearly, if there were no epistemic gap there would not be an ontological gap)? And third, how convincing is the case for mental states being causally effective *per se*? I argue in Section 1.4 that each of these three propositions is knowable from introspection. Here I spell out the case for each of these propositions in more detail.

2.4.1 The existence of mental states

The fact that mental states exist is implied by Descartes' famous statement (1637): "cogito, ergo sum"; the mere fact that one can be aware of having a thought or a feeling is enough to establish for each of us privately, through introspection, that such a thing as a mental state does exist (see Sections 1.4.1 and 3.2.1); there is no need for anything more to be said.

2.4.2 There is an epistemic gap

The important point to make about the possibility that there is an epistemic gap is the one made in Section 1.4.2. If it is not possible for phenomenal facts, facts to which one person or creature has privileged access and which may depend on particular aspects of that person/creature, to be communicated in words without any loss of information, it can be said that there is an 'epistemic gap'.

In the rest of this section we look at three related reasons for accepting this proposition regarding the existence of an epistemic gap. The first is actually enough to make the point; it is simply the fact stated above: that we know from introspection that there are two different classes of fact and that one of them, the phenomenal facts, cannot be communicated perfectly in terms of language while the other, the public facts, can. But the point that it is knowable from introspection that mental states cannot be explained in terms of physics is important to the ideas propounded in this thesis, so we need to look carefully at all the evidence related to it. So next we make the same point in a different way; we note that creatures exist that can access information about the world that is not available to our senses (*eg.*, magneto-reception), and that we cannot imagine what it might feel like to have direct access to that type of information. The third reason concerns the prevalence of views amongst philosophers that are supportive of the existence of an epistemic gap, not a surprising phenomenon if we are correct in asserting that its existence can be known from introspection.

First, barring neurological stimulation (or magic, see Lewis, 1988, p 263) we find that it is impossible to acquire new "what it feels like" type facts (what it feels like to experience something) through verbal communication from others; it is necessary to have the experience oneself (for an opposing view see Lewis, 1988, summarised and discussed in Section 3.4.2). For example, if someone tells you about some totally novel sensory experience that they have had you will not be able to acquire from them the new piece of "what it feels like" type information; perhaps they took a new drug and were able to 'see' ultraviolet light and they report the resulting "seeing experience" as being totally unlike any experience of seeing a colour or colour combination they had ever had before. Indeed, it is generally accepted that someone who has no eyes, or has always been colour-blind or who has never seen colours for some reason cannot discover what it is like to see red through being told what it is like by others, they have to experience it themselves (see Nordby 2007 and Jackson 1982). Of course, this is not to say that feelings cannot be

instigated in a person by the use of words. Obviously, a story recounted by a person, or a poem, can induce powerful emotions in another person.

The second piece of evidence for an epistemic gap is a variation on the first. It arises from the fact that it is now well established that various creatures have phenomenal experiences that are quite different from ours. We know this because, for example, experiments have revealed that certain birds, dolphins, fish, rats and so on can experience magnetism (magneto-reception) in a manner that enables them to use information relating to the earth's magnetic field, both the direction and the dip (the angle between the horizontal and the magnetic field lines), in their daily lives (see Kremers *et al.*, 2014, and many more). If the information content of sensations such as these consisted of public facts, that can be communicated verbally from one person to another, we would in principle be able to know what it feels like to experience magneto-reception.

The third type of evidence supporting the existence of an epistemic gap is simply the fact that so many philosophers (and others) have signed up to this as a fact about phenomenal experience: they all believe that facts about phenomenal experience cannot be made available to others through verbal communication. Could so many philosophers all be wrong? Nagel (1974) describes the difficulties associated with imagining what it is like to be a bat, explaining that insofar as he can imagine it, it tells him: "only what it is like for *me* to behave as a bat behaves" (p 439); he wants to know what it is like for a bat to be a bat. He concludes that: "We cannot form more than a schematic conception of what it *is* like"; the wiring of the neurons in the bat's brain will be different from ours because their senses are different from ours, so even neurological stimulation will not suffice. Searle (1983) says: "… consciousness has a first person ontology; that is, it only exists *as experienced* by some human or animal, and therefore, it cannot be reduced to something that has third person ontology, something that exists independently of experience. It is as simple as that" (*ibid.*, p 60, his emphasis). William James writes (1890, *Principles of Psychology*, volume 1, p 226):

> "… The only states of consciousness we naturally deal with are found in personal consciousnesses, minds, selves, concrete particular I's and you's.
>
> "Each of these minds keeps its own thoughts to itself. There is no giving or bartering between them. No thought even comes within direct *sight* of a personal consciousness other than its own. Absolute insulation, irreducible pluralism, is the law. It seems as if the elementary psychic fact were not *thought*, or *this thought* or *that thought* but *my thought*, every thought being *owned*. Neither contemporaneity, nor proximity in space, nor similarity of quality and content

> are able to fuse thoughts together which are sundered by this barrier of belonging to different personal minds. The breaches between such thoughts are the most absolute breaches in nature."

Levine (1993, p 125) says: "After all, in order to know what it's like to occupy a state one has actually to occupy it." Horgan (1984, p 149) says, of Jackson's fictional character Fred (1982), who could see two totally different colours for what the rest of us call red: "… I am quite prepared to concede that we do not know what Fred's red_1 and red_2 experiences are like, no matter how adequate a physical account we have of Fred's visual processes …". There are a great many more such examples in the philosophy literature (see, for example, Ball, 2009).

2.4.3 Are mental states causally effective?

The case for effective mental causation arises when one considers a choice between two or more courses of action with very different consequences (see Section 1.4.3). For example, what type of mechanism would one put in place, if one were designing the universe, in order to enable a choice to be made by some creature between two actions, one of which looks likely to result in something very beneficial to survival or procreation but also carries with it a substantial risk of physical damage or death, while the other involves the certainty of some rather minor increase in comfort? If the species is going to survive it is of the utmost importance that a choice such as this is made appropriately, on balance. Somehow, the characteristics of these two bundles have to be transformed into something that allows a direct comparison to be made of the likely consequences of the choice for the welfare of the creature and of its species, and this has to take into account the possibility of learning from experience, learning from others' experience, changing technology, climate change, and so on.

As we have seen (footnote in Section 1.4.3), Watson views the feelings that influence the person's choice as constituting the person's 'valuation system', saying:

> "The valuation system of an agent is that set of considerations which, when combined with his factual beliefs (and probability estimates), yields judgements of the form: the thing for me to do in these circumstances, all things considered, is a" (1975, p 346).

Economists, however, model this choice process by defining the thing people find desirable as something called 'utility'. Neuroscientists have also found it necessary to postulate something that performs this function and they talk of 'motivational salience' in this context. Psychologists refer to 'motivational intensity' and identify this with 'affective feelings', or 'motivational affect'.

'Motivational affect' is thus the name given to the element of an emotion that moves one to do or not do something (as opposed to the physical sensations that the emotion brings with it). 'Motivational affect' has valence (positive or negative) and intensity, so a comparison of the 'motivational affect' experienced by a person considering the perceived consequences of each available action will indicate the one that should be chosen, the one that seems most desirable. The analysis requires that there is some way in which a direct comparison between non-comparable, non-commensurable, items can be enabled and, however it is done, it has to end up with picking the options that are the most beneficial, or the least detrimental, to the survival of the chooser or to his or her species. If this is not the case, the individual or the species in question will not thrive. How can this be done (see Section 4.3 for a fuller discussion of this issue)?

The way it *is* done for us and for other creatures, it would seem, is through the existence of and the experience of 'feelings'. And the ability to experience feelings requires consciousness, and therefore mental states. And importantly for our purposes here, if this is the role of consciousness, it has to be the case that mental states can have an effect on physical (non-mental) things (see also Chapter 4).

The obvious question that arises then is: could there be another way of achieving this outcome? Could there be another mechanism, not involving being conscious, which would enable a choice between two bundles of characteristics that are not directly comparable to ensure, insofar as it is possible, the survival of the chooser or of its species, and that in circumstances in which unexpected, even unprecedented, events can occur? And if so, could it be as effective and efficient as the one we have identified here which, one might say, has been designed by nature (by evolution)? It may be that a process akin to the one described above could be programmed into the physics of some creature's brain or into a robot (silicon-based?) along with the necessary one-dimensional valuation function (in order to make the choice it is necessary that a better/worse pairwise comparison can be made of all possible options and this requires that they be valued on a single criterion). If there were another way this type of choice could be made it may be that that is the way it is actually made, and if that were the case, consciousness could be

epiphenomenal after all. Whereas we cannot rule out this possibility on the basis of either logic or current knowledge of the laws of nature, the most likely outcome must surely be that one day it will be found to be nomologically impossible, that is, ruled out by the laws of nature.

In this thesis I take it that the 'role' of conscious mental states is to facilitate survival by enabling decisions of this sort to be taken effectively (see Section 1.4.3 and Chapter 4).

2.4.4 Physics would not be complete but science could be

Finally, if it is the case that mental states are instantiations of a property of matter that is not reducible to physics, so materialism (physicalism in the broad sense) holds, and if everything is determined by causal processes, so science can be complete (see Section 1.2.4), there has to be a causal process that is, in part, outside physics; this would be so even if mental states were unable to cause events relating to physics to take place. For there would be a property of matter, mind, that would be outside physics, and there must be a law of nature that requires that that property exists (see Section 1.2.3) so there must be a law of nature in which something physical, or 'material', has an effect that is outside physics. Although it still could be that causal completeness of physics holds: "all physical effects are fully caused by purely *physical* prior histories", it would not, then, be the case that "all the effects that are fully caused by purely *physical* past histories are physical", which might be required if a more general form of causal completeness of physics were to hold.

The fact that mind, or aspects of it, are outside physics does not lead to any problems with our assumption that materialism (physicalism in the broad sense) might hold, but it does raise the question as to why mental states couldn't be causally effective, having an effect on some of the things that belong to physics. Why rule out that possibility when we are allowing the possibility that there are effects that are outside physics?

There is nothing incoherent about the proposition that a physical event (an event of the sort that the science of physics is concerned with) amongst the neurons can be caused by a mental state (an instantiation of a property of matter not incorporated into the science of physics) as the result of some causal process of which we are currently unaware, but this would run counter to the assumption that physics is causally complete. With this particular type of causal process amongst the neurons operating the science of physics as currently construed, even as it would be if it were 'finished', could not be considered to be 'complete' in this broader sense.

With our proposal, then, a particular type of mental state, an 'intentional' mental state, might cause some physical (pertaining to the science of physics) event to take place amongst the neurons as the result of a causal process which is such that the cause is not included in the science of physics; 'science' would therefore include other causal processes in addition to those incorporated into the physical sciences, and science would be complete.

2.5 How mental causation might operate

2.5.1 Introduction

There is a problem with accommodating mental causation in our understanding of the universe. Lynne Rudder Baker (1993, p 77) voices the concerns of many modern philosophers when she asks: "How can mental events, in virtue of having mental properties, make a difference to behaviour?" After spelling out the metaphysical assumptions that lead to this question arising, she then remarks: "we must either give up (part of) the metaphysical background picture or give up almost all explanations that have ever been offered for anything". In the same vein, Crane notes at the beginning of his 1995 paper 'The Mental Causation Debate' that abandoning the dual substance approach in favour of some type of physicalism does not solve the problem regarding mental causation. If we continue to take it that physics is causally complete in the sense that all physical effects are fully caused by purely physical (that is, non-mental) prior histories, and we reject dual causation ("over-determination"), then inevitably, there is no causal role for mental states to fulfil. The one substance, property dualism type of approach opens up once again the issues relating to mental causation that Leibniz (1695) was concerned about (see Section 3.2.3) unless we abandon the assumption that physics is 'causally complete' in this sense.

Giving up part of the metaphysical background that Baker, Crane and others, assume to prevail is precisely what I am suggesting in this thesis that we should do. If the role of feelings is to facilitate the taking of complex decisions, as I suggest here (Section 1.4.3), it *has* to be the case that the feelings have an influence on the outcome of the decision-making process; it has to be the case that mental events*, in virtue of having mental properties*, affect behaviour and therefore actions. In this section I note some of the possible practical implications of postulating that there must be effective mental causation *per se*.

The next section therefore carries a health warning: much of the material in it could probably be viewed as being fantasy.

2.5.2 Hypothetical mechanisms for mental causation

Arguably the most obvious possibility for a process underlying mental causation is that the effect of the mental state associated with the taking of a decision is transmitted directly to the action neurons which implement the decision by one of the force fields already familiar to fundamental particle physicists. This would entail this particular "feeling", the feeling that accompanies the taking of a decision, which we take in this thesis to be an instantiation of the property of matter we call 'mind', interacting directly with another property of matter such as electric charge (perhaps because it is associated with an electric field), or directly instigating a chemical reaction through either the weak or the strong nuclear force. The explanation of effective mental causation would then lie simply in the fact that the taking of a decision, following the stimulation of the feelings (the 'motivational affect', see Sections 2.4.3 and 4.3.2) associated with the different options available, itself causes the emission of some neurotransmitter at one point amongst the neurons rather than at another, or influences the path of an electron through its effect on the electric or magnetic field. Feelings, as mental states, postulated here to be instantiations of a property of matter, could then affect other properties of matter through the type of causal processes that we express as laws of nature when considering the manner in which physics operates.

However, nowadays, in physics, causes are generally taken to operate at a distance through fields. These fields arise in the context of some entity, a fundamental particle or some-such (the 'cause'), and affect the behaviour of another such entity (the 'effect') some distance away. According to the present understanding in physics, taking into account the Standard Model, there are four fundamental forces: electromagnetism, the weak nuclear force, the strong nuclear force and gravitation (which operates on the large scale). If mental causation turns out to be effective in initiating actions it could be the case that the process involves a fifth force, a highly localised 'field', which one might call a 'decision field', which could be different from the forces familiar to physics today. It would be produced by the instantiation of the particular type of mental state that is associated with the taking of a decision and would therefore be related to what the psychologists call 'motivational affect'. It would then have to be able to influence some aspect of the activity amongst the neurons that initiates actions; perhaps it influences the speed or direction of travel of an electron in a neuron, or perhaps the outcome of a chemical process. If such a

field were to exist, we would consider it to be one of the fundamental properties of matter. However, although theoretical physicists apparently postulate new forces fairly regularly to explain astronomical phenomena not already explained in a satisfactory way by physics as currently formulated, I am told that they are reasonably confident about the existence of no more than these four forces on the micro level.[44]

A third possibility arises in the context of quantum mechanics. Quantum mechanics describes a system in terms of its "wave function"; from this wave function the probabilities of the results of measurements made on the system can be derived. According to the Copenhagen interpretation of quantum mechanics, physical systems do not have definite properties prior to some measurement being made; quantum mechanics merely indicates the probability that a measurement will produce some particular result (see von Neumann, 1932, translated 1955). The act of measurement causes the set of probabilities to reduce to only one of the possible values ("wave function collapse"). Thus the position (or momentum) of some particle only becomes "certain" when it is measured by "an observer". On the Copenhagen interpretation of quantum mechanics, therefore, one could argue that there is a role just waiting to be filled by consciousness; because there is a role for something that "performs a measurement", something that "is an observer" (the Copenhagen interpretation is, of course, highly controversial).[45] So now, suppose that a person is taking a decision about what to do next. He or she is considering all the options for action, weighing up what the consequences of each are likely to be, examining the perceived benefits, the disadvantages, the extent of the uncertainties and the type of risks to which they are subject, and trying to assess which action would be most desirable (see Section 1.4.3); so he or she is comparing the probability distribution of 'value' associated with the different available actions in order to decide which of those probability distributions is most desirable, and so should be chosen (as noted in Section 1.4.3, various moments of the probability distribution are relevant to its desirability, not just its expected value). The state of the relevant bit of the brain (presumably part of the basal ganglia, see Chapter 4) at this time could presumably be represented by a wave function. When the

[44] However, when I contacted Tom Kibble, an ex-colleague, just before his death ("one of the world's foremost theoretical physicists", *The Guardian*, 9 June 2016) explaining my proposal and asking his view: "on the question of whether there is room in physics, as you know it and understand it, for an extra force or property that is not currently thought to exist", he responded saying: "… there is always room for the introduction of a new force …".

[45] See London and Bauer (1939), von Neumann (1955), Wigner (1961), Shimony (1963), Jammer (1974), and Hodgson (1991).

person actually identifies and decides upon the option that appears to be the most desirable, when he or she makes that measurement regarding the intensity of the 'motivational affect' that is associated with each of the different options, the wave function collapses, picking out the chosen action with 100 per cent probability. If this were the correct option then we wouldn't be suggesting that there may be something missing in physics as it is today, because it doesn't take account of mental causation, we would instead be saying that there is already known to be a role in physics for an observer, a mind; that the measurement of the intensity of the 'motivational affect' associated with each of the available options, by the person (or creature), triggers the wave function to activate one action rather than another. We would be saying that this is how mental causation operates.

Obviously all these possibilities are highly speculative. But as Baker says (above), we must be getting something wrong.

Since it is important in science that hypotheses can be tested, we should note that the hypothesis that the taking of a decision can have a direct effect on an action neuron could, in principle, be explored experimentally. For example, this could be done by investigating whether there are circumstances in which one creature's decision could directly initiate an action by another creature through triggering a relevant change in the second creature's action neurons instead of in its own. This might be done by placing the decision neurons of the first creature (perhaps the relevant part of the first creature's basal ganglia, since that is where it is currently thought such decisions take place, see Section 4.3.1) in the correct position relevant to the action neurons in the second creature's brain. If some type of stimulation of the external senses of the first creature, perhaps suggesting imminent danger of damage or pain, were found to induce an avoidance movement by the second creature, this might enable us to discover the precise mechanism by which the communication of the message came about. The way this might be done could be similar to the way in which Kandel and his colleagues (see Kandel, 2006) were able to confirm their hypotheses regarding the way in which the short term memory in sea snails (Aplysia) operates. The scientists tested their hypotheses about the manner in which a stimulus to the tail of the sea snail results in the sea snail's gill withdrawal reflex by applying the chemicals they believed act as neurotransmitters in the appropriate places relative to the sea snail's sensory neuron instead of applying the external stimulus. This confirmed their hypothesis that these neurotransmitters were sufficient to trigger the response.

David Lewis, in his influential 1988 paper 'What Experience Teaches', said: "… if something nonphysical sometimes makes a difference to the motions of physical particles, then physics as we know it is wrong. Not just silent, not just incomplete – wrong" (p 95, and see Section 3.4.2). But actually, if one of these proposals regarding mental causation proved correct it would simply mean that physics as we know it today is incomplete, because something which is "not physical" (in the pertaining to the science of physics sense) would have been shown to be capable of making a small difference to the motions of some particles in the brain (and therefore, potentially, an enormous difference to the future path of the world). In all other circumstances, that is in any place other than inside the brains of certain types of creatures, the working of physics would be likely to be totally unaffected. Further, *science*, which would include laws of nature relating to mental states as well as those that relate to physics, could be complete. There is nothing in physics as we know it today that suggests that no further developments of this sort can take place; indeed, theoretical physicists regularly postulate new phenomena to explain physical observations that do not fit easily into the current understanding.

2.6 The three (so-called) anti-physicalist arguments

There is no need for any of the three physicalist conundrums that have for some years been much debated in philosophy of mind, the knowledge argument, the conceivability argument and the hard problem (the existence of an explanatory gap), to provide difficulties for the causally effective property dualism proposition ("physicalism in the broad sense", "materialism") under examination in this thesis. Arguably, this is one of the more attractive features of this proposal from the viewpoint of modern philosophy.

2.6.1 The knowledge argument

Consider Jackson's 'knowledge argument' (1982 and 1986). Mary knows everything physical (non-mental) that can be known about seeing and about colours (in Jackson's words: "[s]he knows all the physical facts", 1986, p 392), but she has grown up without having ever actually seen colours herself. The knowledge argument proposes that Mary learns something new when she first sees colours, and therefore there must be something over and above the "physical facts" and this means that "physicalism", as envisaged by Jackson at that time, is false.

But Jackson's "physical facts" consist of "information about the world we live in and about ourselves" that has been provided by "the physical, chemical and biological sciences" and the "information that automatically comes along with it" (1982, p 127; we should note that Jackson says here that he does not mean these "sketchy remarks" to constitute a definition of 'physical information'). His knowledge argument suggests that, if Mary does learn something new, there must be other facts, facts relating to what are known as '*qualia*' (Jackson believed in *qualia* when he wrote this paper, claiming that he was a "*qualia* freak"), which are different in some way from the so-called physical facts that Mary already knew. So the implication is that if these other facts exist, they must either relate to the properties of a different substance from matter or, and this is the possibility that I am examining in this thesis, there must be another property of matter over and above those properties that Jackson calls "physical". Jackson concludes the presentation of his knowledge argument saying:

> "It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false" (*ibid.*, p 128).

Whereas Jackson's thesis that if Mary learns something new then "Physicalism" is false is clearly true of the type of physicalism Jackson was interested in at that time, reductive physicalism, it does not hold for the non-reductive type of physicalism in which mind is a property of matter in its own right. If mind is a property of matter, with facts about conscious mental states (*qualia*, or phenomenal facts) analogous to, but different in nature from, the facts physicists describe about non-mental states, it is obviously possible for Mary to learn new facts about seeing and about colour, even though she already knew all the facts the 'finished' physical sciences can produce about them, without physicalism in this broad sense being false. These phenomenal facts relate to mental states – to the property of matter we know as 'mind'; and if mental states are instantiations of a non-reducible property of matter these phenomenal facts cannot be reduced to facts relating to the properties of matter studied in the physical sciences. So there is no reason to assume that physicalism in the broad sense ("everything can be explained in terms of the properties of matter") is false just because when Mary first experiences what seeing red feels like she learns something she didn't know before.

In the world of Jackson (1982 and 1986) "physicalism" is reductive; it is the thesis that there are no facts other than those that are of the type studied in the physical sciences, no information other than that which has been provided by "the physical, chemical and biological sciences". In this thesis I propose that we know *from introspection* that "what it feels like" type facts, so-called 'phenomenal' facts, exist, and that we also know *from introspection* that these are different from the facts of physics because they cannot be perfectly communicated in words (see Section 1.4.2). In this context, Jackson's conclusion is quite correct, and just what we would expect.

The really strange thing about the Mary story, and about the existence of the enormous number of academic papers written on the subject, is that this simple, totally obvious explanation of the knowledge argument (of the fact that our intuitions are at variance with its implications for what Jackson terms 'Physicalism') is rarely proposed and that no-one explains why this is so. The only person I have found who gives this explanation for the knowledge argument is Terence Horgan (1984, see p 147), who argues that: "his [Jackson's] attack on Physicalism is fallacious, being an equivocation on two different senses of the phrase 'physical information'" (although, of course, on Jackson's meaning for term "Physicalism" it is not fallacious, it is correct). He says (pp 149-150):

> "What I want to question is Jackson's supposition that a completely adequate physical account of a creature's visual processes gives us complete *physical information* about those processes. In one sense of 'physical information', this supposition is virtually a tautology: for, physical information is just the information that would be provided by a theoretically adequate physical account. But in another sense – the sense really required by the knowledge argument – the supposition is one that Physicalists can and should reject."

Horgan then distinguishes two relevant senses of the term 'physical information'. One he calls 'explicitly physical information'; he says this is information expressed "in overtly physicalistic language". In the framework that I develop in this thesis Horgan's explicitly physical information consists of information about facts relating to the material world that can be communicated perfectly using language. The other he calls 'ontologically physical information'; he says this is information expressed: "by other sorts of language – for instance, mentalistic language" (*ibid*., p 150). In my framework, Horgan's ontologically physical information, which is expressed (presumably imperfectly) by mentalistic language, is information about facts of the "what it feels like" variety; it could be referred to as some

sort of 'physical information' if the word 'physical' is interpreted in the "pertaining to matter" sense and if mind is a property of matter.

If we accept that there *are* facts of the "what it feels like" type (*qualia*) and that these cannot be communicated perfectly from one person to another using language, then there are facts that Mary could not have got to know in her black and white room. If we don't want to accept that the implication of Jackson's knowledge argument ("Physicalism is false") is that there must be "immaterial minds", there is another, perfectly obvious and plausible possibility: there could be another property of matter besides those studied in the physical sciences. It is difficult to see why this possibility is so universally ignored.

Perhaps the work of Lewis is relevant here. Lewis, in his 1988 paper 'What Experience Teaches', rules out the possibility of phenomenal information being associated with a property of matter different from those studied in physics by defining 'phenomenal information' as being *independent* of physical (pertaining to physics) information in the sense that: "The two are independent. Two possible cases might be exactly alike physically, yet differ phenomenally" (see Section 3.4.2). But this definition of phenomenal information is unduly restrictive; it rules out the possibility of there being another property of matter besides those studied in physics; for if this were the case one would expect phenomenal information to be *correlated* with physical (non-mental) information. Perhaps this has not always been apparent to people considering the knowledge problem.

It is obvious that materialism, which can also be called "physicalism in the broad sense", is consistent with the knowledge argument; the knowledge argument clearly does not have the implication that *all* versions of physicalism are false (although it does, of course, have the implication that the type of physicalism Jackson referred to as Physicalism in his 1982 and 1986 papers is false).

Finally, besides that of Horgan with which I agree, there are three other main explanations of the knowledge argument that arise in the philosophical literature; all three are predicated on the proposition that there is no such thing as a phenomenal fact, a fact of a different type from those that are studied in physics.

- First, there is the view that the knowledge argument is correct, but the conclusion that we draw from it, using our intuition, is just plain wrong. If Mary actually could know *all* the so-called 'physical' facts, and if there are *no* facts other than physical facts, she would indeed have known all along what it would be like to see red. But, "… in any realistic, readily imaginable version she might know a lot, but she would not know everything physical" (Dennett, 1991, p 400); so our intuition is leading

us astray when it tells us that Mary must learn something new when she leaves her room; it may be that, in fact, if she actually did know all the 'physical' facts she would not learn anything new, after all.

- Second, there is the proposition that Mary does learn something new, but what she learns is not a new 'fact'; she acquires a new skill. This is Lewis' "ability hypothesis" (1988). The idea, in Nemirow's words, is: "some modes of understanding consist, not in the grasping of facts, but in the acquisition of abilities…. I understand the experience of seeing red if I can at will visualize red" (1980, p 475). But acquiring a skill, in the riding-a-bicycle or playing-a-piano-sonata sense, takes time, practice and the reprogramming of the neurons in the cerebellum; learning to visualize red doesn't take practice, so it is not analogous to learning a skill in this sense (the ability thesis is discussed in greater depth in Section 3.4.2).

- Third, there is the proposition that what Mary learns is not a new 'fact', and nor is it a new ability; she has merely learned to think about seeing red in a new, different, way. In the words of Papineau (2002, p 51), she has acquired "a new *concept* of seeing something red" (his italics), a so-called phenomenal concept, a different type of concept from the ones used in physics. This is the approach to consciousness known as "concept dualism". If reductive physicalism can hold along with the existence of an epistemic gap, Mary will be unable to acquire this concept of seeing red from the knowledge she has acquired through learning and concept dualism may hold. I spell out the case for concept dualism in Section 3.3.5.

2.6.2 The conceivability argument

Now consider the conceivability argument. This concerns the possible existence of zombies: creatures that are physically (that is, in all things that pertain to physics) identical to some human and behave in exactly the way that human behaves in spite of being incapable of conscious experience (see Chalmers, 1996, pp 94-99). The proposition is that if zombies are conceivable it must be possible that the particular arrangement of matter that is associated with the physical and the mental characteristics of some person *could* be associated with only the physical characteristics; and if that is possible, physicalism must be false.[46]

The conceivability argument is often summarised thus:

[46] In philosophy, the word 'conceivable' is generally used to imply only that that the statement is not incoherent; in casual conversation, it usually means that a proposition is nomologically conceivable as well as not being incoherent.

i. P_t & ~Q_t is conceivable
ii. If P_t & ~Q_t is conceivable, P_t & ~Q_t is possible
iii. If P_t & ~Q_t is possible, physicalism is false.

where P_t is the conjunction of all the facts about the way matter is distributed and therefore the facts about all the physical properties of matter that obtain. Q_t is some truth about a conscious mental state.[47]

First, we should emphasize that zombies are defined to be exactly like some human being in every physical (pertaining to physics) respect. They look exactly like some human being and behave in exactly the way that human behaves. But they lack consciousness; "There is nothing it is like to be a zombie" (Chalmers, *ibid.*, p 95). So the proposition that they exist looks highly improbable given our third fundamental proposition (see Section 1.4.3). For if the existence of a zombie were to be possible it would have to be the case that the zombie, a creature that has no consciousness and therefore no feelings, could take the exact same decision in whatever circumstances obtain as would a person for whom such a decision depends on a complex array of feelings. We put that thought aside for the purposes of this section.

One approach to the conceivability argument takes the view that it is only the second proposition, the proposition that if something is conceivable then it is possible, that is problematic (see, for example, Crane, 2001, p 100). This proposition is viewed as being problematic because there *can* be situations which are in a sense conceivable while not actually being possible. An example in which this is the case arises in an argument that has been developed in the literature relating to the explanatory gap (see Levine, 1986, and Section 3.3.4) and derives from the work of Kripke.[48]

Saul Kripke is an authority on necessity statements, an issue he addressed in his book called *Naming and Necessity* (1972). Here is the three-step argument put forward by Kripke to explain why, since there is *not* an analogy between the hypothecated identity between mental states and brain states and the identity between water and H_2O (*ibid.*, pp 148-151), the second proposition above can be expected to hold in the pain/stimulation of the 'pain neurons' case:

[47] The conceivability argument is akin to Descartes' argument in support of substance dualism, see Section 3.2.2.
[48] It might seem to be conceivable that water is not H_2O if someone experienced the watery qualia in the presence of something other than water, but it is not possible for water not to be H_2O.

- First, Kripke says: "the usual view holds that the identification of heat with molecular motion and of pain with the stimulation of the C-fibers are both contingent" (*ibid.*, p 148).[49]

- Second, he says: "When someone says, inaccurately, that heat might have turned out not to be molecular motion, what is true in what he says is that someone could have sensed a phenomenon in the same way we sense heat, that is, feels it by production of the sensation we call 'the sensation of heat', even though that phenomenon was not molecular motion", (*ibid.*, p 150). Note that here the terms 'heat' and 'molecular motion' continue to pick out the same thing; so this identity continues to be true. The only change is that someone happens to be wired up in a different manner from the way in which we are wired up so the *qualia* that are experienced by these creatures in the presence of heat are different from those we experience. Different laws of biology operate regarding this person.

- Third, Kripke then asks whether there is something that can be said that is analogous to this story about heat and molecular motion that applies in the case of the hypothecated identity between pain and stimulation of the 'pain neurons'. He says: "Now can something be said analogously to explain away the feeling that the identity of pain and the stimulation of C-fibers, if it is a scientific discovery, could have turned out otherwise?" (*ibid.*, p 151). The question he asks is: "is it analogously possible that stimulation of the C-fibers should have existed without being felt as pain?" (*ibid.*, p 151). But he finds: "Such a situation would be in flat out contradiction with the supposed necessary identity of pain, and the corresponding physical state, …" (*ibid.*, p 151).

So whereas the second proposition, above, may not hold in the case of the identity between heat and molecular motion, or water and H_2O, Kripke finds that it does hold in the case of the proposed identity between pain and stimulation of the 'pain neurons'.

Levine has pointed out (see Levine, 1983, and Section 3.3.4) that there is a difference in our intuitive response to the statement "Heat is the motion of molecules",

[49] Reference to 'C-fibers' in this context is of course figurative. In neurology, 'C-fibers' is the name given to a class of nerve fiber which carries signals from the periphery to the central nervous system; these nerve fibers do not occur in the brain and they carry other messages as well as those relating to pain. The reference here is not to those C-fibers. In this branch of philosophy of mind the term 'C-fibers' is used figuratively and refers to the fact that damage, perhaps to the toe, is spoken of by philosophers as being experienced (as pain in the toe) as the result of stimulation of a fictional part of the brain called "C-fibers" (*nb.*, it is, of course, very well-known that damage to the neurons does not cause pain), see Puccetti, 1977. Here, I replace this term 'C-fibers' with the phrase 'pain neurons' wherever reasonably possible.

which people generally feel provides a satisfactory and complete explanation of the phenomenon of 'heat', and our intuitive response to the statement "Pain is the firing of the 'pain neurons'" which people generally feel fails to give a totally satisfactory explanation of the phenomenon of pain because it doesn't explain what it actually feels like to experience pain. This difference is relevant to the conceivability argument because as Crane remarks, in the case of water is H_2O or heat is the motion of molecules: "it might be that one could conceive of a situation where water is not H_2O, but what one is conceiving is impossible because water is necessarily H_2O" (2001, p 100). This means, of course, that there *can* be situations which, in Crane's words: "are in some sense conceivable but not truly possible", in contradiction of the second proposition of the conceivability argument. However, as Crane then points out: "One explains away what one is conceiving when one conceives that water is not H_2O by saying that one is conceiving of something which only *seems* like water but isn't. But when one imagines the zombie, one cannot be imagining something for which it only seems like it lacks the feeling of pain (say), but really is in pain: for anything that really is in pain can never be said to lack the feeling of pain!" (Crane, 2001, pp 100-101). The conclusion from this consideration is, therefore, that in the case of mental states that second premise does hold good, so the conceivability argument looks to be sound. Crane remarks, in conclusion: "There are many ways to respond, but it seems to me that none of the responses are adequate. Physicalism fails as a theory of consciousness" (p 101).

However, there is also another point that can be made about the conceivability argument, in the context of the approach spelt out in Section 1.2.3 of this thesis. We can accept that the first proposition of the conceivability argument holds, but only so long as the use of the word 'conceivable' is taken to imply nothing more than that a statement is *not incoherent*, that it is not obviously ruled out by *logic* (Chalmers says in his presentation of the conceivability argument: "There is a *logically* possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold", 1996, p 123, my italics). But suppose there were a law of nature that ensured that consciousness is a property of matter (perhaps ontologically reducible to the properties of matter studied in physics, perhaps not). If that were the case it might seem to us that zombies were possible until experimentation revealed to us that there was a law of nature that carried the implication that they are not possible; but once we did know that, the existence of zombies would no longer seem conceivable. So there is a big proviso here; the consequence

argument only demonstrates that the proposition "physicalism is true" cannot be established *as a matter of logic*.

The conclusion drawn from this argument, which is that if zombies are conceivable then physicalism is false, is therefore only of interest to us if we had required that physicalism hold *as a matter of logical necessity*, so the truth of the proposition that physicalism is true could be known without experimentation.

2.6.3 The hard problem

The hard problem of consciousness, labelled such by Chalmers (1995), is the problem of explaining how/why consciousness arises in the context of matter. This is often referred to as an 'explanatory gap' (see Levine, 1983). In Chalmers' words: "The explanatory gap comes from considering the question, Why, given that P is the case, is Q the case? …" where P is: "the complete microphysical truth about the universe" and Q is: "an arbitrary truth about phenomenal consciousness" (Chalmers 2007, pp 168 and 169).

The proposal explored in this thesis is that there is a fundamental fact, akin to the fundamental facts regarding the existence of the properties of matter that are studied in physics such as electric charge (see Section 1.2.3), that says that if a certain type of material state exists a certain type of phenomenal consciousness will be instantiated as well as the types of physical (non-mental) properties that are instantiated (although the precise details relating to the particular material state, much sought by some scientists, remain to be discovered).[50] If that is the case, the question raised by the "hard problem" would be analogous to the question: Why is there such a thing as electric charge? The obvious answer is simply: it is one of the fundamental facts relating to the universe that this is so. If the fact that mind exists is a fundamental fact of this sort, and if the endpoint of our search for explanations is the set of fundamental facts that make the world operate in the way it does, the answers to the 'why' questions, then it seems likely that we have reached the final explanation. For if we think that mind is one of the fundamental properties of matter, just as is reckoned to be the case with electric charge, we would not, as scientists, ask how it can possibly be that mind is instantiated in this way, which is what philosophers do when contemplating the so-called hard problem, we would simply accept that as being

[50] If you prefer the bundle theory approach to matter (see Section 1.2.4) rather than the substratum view, you will postulate that a fundamental fact ensures that consciousness, or something that causes consciousness, can be associated with the properties in the "bundle", as is electric charge.

the case. Levine in fact provides that answer to his own question (2007, p 147): "… what is there to say in answer to why water is H_2O but that that's just the way it is? A similar answer, it is alleged, is appropriate in the psychophysical case".

However, such an explanation of the existence of mind is unlikely to satisfy everyone. Christof Koch, in 2012, wrote: "The endpoint of my quest must be a theory that explains how and why the physical world is capable of generating phenomenal experience" (Koch, 2012, p 114-115). Perhaps philosophers tend to think of "matter" as being *defined* as being "lifeless", and therefore as being "consciousness-less" and therefore see the hard problem as a problem simply because of their implicit definition of the word "matter" (see Section 1.2.4).

2.7 Conclusion

In many ways the assumption that mental states are instantiations of a property of matter that can be causally effective and are not reducible to the properties of matter studied in the science of physics is an appealing hypothesis. It constitutes a coherent description of the ontology of mind, and it explains why mind can be thought of as existing in the context of the material in the brain, suggesting that the perceived existence of an explanatory gap is unnecessary. It also deals with both the knowledge argument (because it implies that phenomenal facts do exist) and the conceivability argument (because the conceivability argument only implies that physicalism cannot be true as a matter of logic, it says nothing about the possibility that physicalism could be true because mental states are instantiations of a property of matter different from those suitable for study in physics). The hypothesis seems to tick all the boxes except one. If mental causation is effective *per se*, then one particular type of causal process, that in which a creature decides to perform an action, has a mental cause (rather than a physical, in the sense non-mental, cause), so physics is not causally complete. An incomplete physics in this sense has long been considered by many philosophers to be an unacceptable notion.

But the approach is consistent with the proposition that science is complete, in the sense that all the causes and all the effects of all the laws of nature could be included within it. The implication is simply that there may be another property of matter, different from those that are studied in physics, and there may be other causal processes besides those of which we are currently aware; so science may be wider in its scope than the physics that some had thought might explain everything. What we are doing is postulating the existence

of an additional property of matter, not presently considered by physicists, that can interact causally with the properties of matter already incorporated into physics, in order to explain a phenomenon that we know exists and that is not explained by physics. Perhaps we should think again about the causal completeness of physics.

Chapter 3 Approaches to the nature of mind

3.1 Introduction

As we saw in Chapter 1, there are three main possibilities concerning the relationship of mind and matter, and they can broadly be described as: substance dualism, reductive physicalism and non-reductive physicalism. In this chapter I summarise the main *pros* and *cons* in the philosophical literature regarding each of these approaches to the ontology of mind.

First, in Section 3.2, I consider the case of substance dualism, in which it is assumed that there is some "essence of mind", perhaps "consciousness" itself, which is not instantiated by the substance we call 'matter', so consciousness would be outside science; that means that consciousness would be unrelated to the laws of nature that would constitute the body of knowledge of a finished "science" (see Section 1.2.2). Thus, there could be another substance besides the one we call 'matter'. If this were the case, there is a question as to whether an event relating to one substance can cause an event to take place in another. It seems to be impossible to rule out the dual substance approach as a matter of logic, or for other reasons, at least given the current state of knowledge. However, in this thesis it is not discussed outside Section 3.2 apart from a brief mention in the context of free will (see Section 5.2.3).

If it were to turn out that there is only one substance, matter, with matter often referred to as being "physical", then this thing we call consciousness must be in some way related to this so-called "physical" substance ("one 'substance' and we call it 'matter'" combined with determinism or universal probabilistic causation implying, of course, that the occurrence of every event can, in principle, be explained by the laws of nature, see Section 1.2.4). A question then arises as to whether mental states can be wholly accounted for in terms of the properties of matter we study in the physical sciences, the second of the three main approaches to the relationship of mind and matter (discussed in Section 3.3), or whether there could be another property of matter, not suitable for study in the physical sciences, associated with the existence of minds. This is the third main approach, and the one in which our causally effective property dualism belongs (discussed in Section 3.4).

In reductive physicalism (Section 3.3) it is assumed that mental states derive from the properties of matter that are studied in physics (clearly this position requires clarity

regarding what is suitable for inclusion in the physical sciences and what is not, see Section 1.2.1). Other possibilities that may require nothing over and above physics include "eliminativism", according to which it is assumed that some types of mental state (or, indeed, all of them) simply don't exist, and "behaviourism" according to which the only real thing about mental states is the behaviour with which they are associated (see Section 3.3.6).

The relation between mind and matter for which I am putting the case in this thesis is a form of non-reductive physicalism, the third of the three possibilities mentioned above, and the varieties of non-reductive physicalism are discussed in Section 3.4. In this case, again, there is just one substance, matter; however, as well as the properties of matter studied in physics this approach assumes that matter instantiates another type of property, mind, which is not ontologically reducible to those that are studied in physics. There is then a question as to how this other property arises. It may be 'emergent', arising as basic living creatures evolve into more complex ones (see Section 3.4.10), or panpsychism may be the answer, the proposition that consciousness has pervaded the universe since the beginning of time, or perhaps it is micro-psychism, the proposition that some fundamental entities are intrinsically experiential, in which case consciousness would be one of the fundamental properties of matter, as are mass, electric charge, and so on.

3.2 Substance dualism and Vedānta

A natural starting point for thinking about the relation between mind and matter seems to be to consider the possibility that the two are separate in some way. Thus to the extent that ancient philosophers identified the existence of consciousness, or of mental states, as something that required an explanation they tended to postulate that substance dualism must be the case, that is, they thought that there was probably some special substance, something different from "matter" (which is extensive), and which, in some sense, *is* consciousness or is something that one might call "spirit".

The ancient Indian philosophers, for example, following Samkhya, the oldest school of Indian philosophy (probably around the 6th Century BC), assumed there was a form of spiritual energy, or a 'being', which they called Purusha; they thought this must have pervaded the entire universe for all time (see Bernard, 1947). Thus the name Purusha was originally a mythical being, but it came to mean "person," "self," "spirit," or "consciousness", or "the eternal, authentic being". Purusha could be set opposite Prakriti

in Indian philosophy, where Prakriti is the basic matter constituting the universe; an early version of two substances approach. There is still a school of thought, known as Vedānta that signs up to the proposition that consciousness is the 'substance' from which everything derives (although different schools within Vedānta hold different views regarding the question of monism/dualism *etc.*).

Modern Vedāntic philosophers maintain that, whereas in the far-distant past philosophers had no problem in accepting the existence of consciousness in the same way as they accepted the existence of the physical world, since the advent of Newtonian mechanics, Western scientists and philosophers have accepted the materialistic conception of reality (excepting, of course, Berkeley, 1710, who developed a theory named "immaterialism" which denies the existence of matter and maintains that objects like trees and stones only exist in the minds of those who perceive them). The Vedāntic view has taken the opposite stance assuming that the origin of everything, either material or immaterial, is "sentient". It sees sentient life as primitive and "reproductive of itself"; life comes from life (see Shanta, 2015). Consciousness, which it sees as being fundamental, then must "manifest itself" in everything that has either a sentient or an insentient nature. So, in contrast to the idea of the evolution of bodies, as proposed by Darwin and his followers, Vedānta signs up to the proposition that the evolution of consciousness is the principle underlying the development of the world.

Interestingly, starting from the materialist point of view, Western philosophers have postulated the bundle theory approach to the nature of a "substance" (see Section 1.2.4), assuming that there is just a bundle of properties with no substratum. If consciousness is taken to be one of the properties that may be in the "bundle", as mass or electric charge might be, this could transform materialism into an approach somewhat akin to this ancient Indian view, effectively a form of pan- or micro-psychism, in which consciousness would be a property that has existed since the beginning of time, just as is postulated in Western philosophy for the properties studied in the physical sciences.

Spinoza, a Dutch philosopher, proposed an approach broadly similar to this Indian concept in 1677. His proposal takes matter and mind to be two of a possibly infinite set of attributes of something all-pervasive that one might call God or Nature.

The closest the ancient Greeks came to a concept of consciousness seems to have been encapsulated in the word 'nous', a key term in the philosophies of Plato and Aristotle (see, for example, Perelmuter 2010). In Plato's later dialogues 'nous' refers to the highest activity of the human soul. A particular value is attributed to things relating to nous as

opposed to things that are perceptible or corporeal. Like the ancient Indian philosophers, the Greeks also saw nous as pervading all things and believed that it had existed from the beginning of time.

The existence of two substances ("the ghost in the machine", see Ryle, 1949) is probably still where most people start when pondering the nature of consciousness.

3.2.1 The existence of mind

Descartes, right at the start of the 'Age of Enlightenment', began to consider the nature of mind, because he felt it was relevant to his study of science. Over the years he had become aware that there were a number of things that he was taking for granted without giving their validity any serious consideration, so he decided to sit down and give the whole matter of the fundamentals of 'science' and 'being' some serious thought. In the first of his *Meditations* (1641) he describes how he began by asking himself whether there was anything that he knew for certain. For example, could he be absolutely certain that he was sitting by the fire, and that the things he could perceive with his senses as he sat there were really there? Was it possible that all the things he could perceive were actually a figment of his imagination, or a dream? He decided there was no way in which he could be absolutely certain they were there. He thought, as surely many others must have done before and since, that it could be the case that a demon was deceiving him and that everything, including his perceptions relating to his body and the messages he was receiving from his senses, might be nothing more than illusions. He could see no way of ascertaining whether this was or was not the case.

But then, if there is nothing that he could be certain exists, surely it ought to be the case that he, himself, also might not exist. Could he be certain that he existed? But the answer to that question was indubitably yes, that he could – and that was because it was he that had realised that all these things his senses received messages about might not exist; it was he that had the idea that he might have been deceived by some demon, or some strange situation. Clearly, he could think, and that meant that it had to be the case that he existed. This, he concluded, is true beyond all doubt: whenever a person is aware of thinking, or feeling, it must be the case that he or she exists.

A variant on this thought of Descartes' constitutes our first fundamental proposition (see Section 1.4.1): the one thing that we know to be true beyond all doubt is the fact that we think, that we experience something that we call 'mental states', and that we therefore know that mental states exist.

3.2.2 An argument for substance dualism

Descartes continued his project and reached another important conclusion, one of lasting significance although it is now thought that his reasoning was at fault. He decided that although he was inclined to believe that there was a space-occupying, extended body to which he was clearly very closely conjoined, he himself, by which he meant his mind because that is what he viewed 'himself' as being, was a thinking and non-extended thing, distinct from his body. It must therefore be the case, he concluded, that his mind could exist without his body. Mind cannot therefore exist only in the context of matter. There must be another 'substance'; in Descartes' view this must be a substance which can, as a matter of logic, exist on its own.

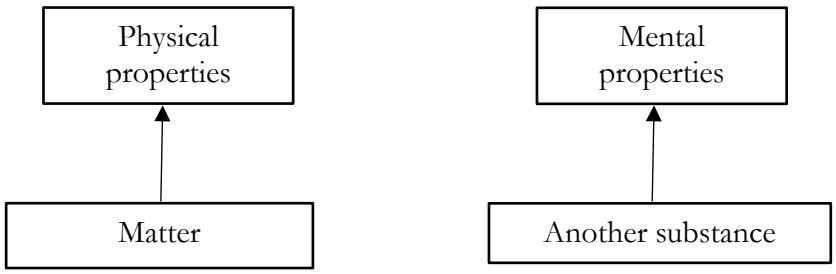


Figure 3 Substance dualism

In substance dualism mental properties exist in the context of a different substance from that of 'matter', which instantiates the properties studied in physics. This option is viewed as leading to problems relating to mental causation – how can a property of one substance have effects on a property of another substance?

Descartes' justification for his view that mind relates to a substance different from that studied in physics is the following (see his *Meditations VI*, paragraph 9):

> "And, firstly, because I know that all which I clearly and distinctly conceive can be produced by God exactly as I conceive it, it is sufficient that I am able clearly and distinctly to conceive one thing apart from another, in order to be certain that the one is different from the other, seeing they may at least be made to exist separately, by the omnipotence of God; and it matters not by what power this separation is made, in order to be compelled to judge them different; and, therefore, merely because I know with certitude that I exist, and because, in the meantime, I do not observe that aught necessarily belongs to my nature or essence beyond my being a thinking thing, I rightly conclude that my essence

> consists only in my being a thinking thing [or a substance whose whole essence or nature is merely thinking]. And although I may, or rather, as I will shortly say, although I certainly do possess a body with which I am very closely conjoined; nevertheless, because, on the one hand, I have a clear and distinct idea of myself, in as far as I am only a thinking and unextended thing, and as, on the other hand, I possess a distinct idea of body, in as far as it is only an extended and unthinking thing, it is certain that I, [that is, my mind, by which I am what I am], is entirely and truly distinct from my body, and may exist without it."

This can be summarised in a manner broadly similar to the summary of the conceivability argument (in Section 2.6.2), thus:

i. I know that all that which I clearly and distinctly conceive can be produced by God exactly as I conceive it *(if something is conceivable it is possible)*

ii. I know with certitude that I exist and that my essence consists *only* in my being a thinking and unextended thing; while my body, with which I am closely conjoined, is *only* an extended and unthinking thing *(zombies are conceivable)*

iii. Therefore, I am distinct from my body and can exist without it *(physicalism is false)*

The problem with this, of course, is that although the proposition that Descartes' 'thinking thing' cannot exist without its body does not run counter to logic, so can be conceived in the sense that it is not internally inconsistent, there could be a law of nature which has the implication that the situation he is considering is impossible. If that were accepted as being the case, the proposal that he could exist without his body would be *nomologically* inconceivable and the second point above would then not follow from the first point, it would *not* be possible that his mind could exist without his body.

3.2.3 Mental causation with substance dualism

A problem that arises with substance dualism concerns the possibility of events relating to one substance influencing events relating to the other. If mind has no physical features such as spatial location, mass, volume, electric charge and so on, as is taken to be the case in substance dualism, then how can it interact with matter? This issue did not escape thinkers in Descartes' time. The Princess Elizabeth of Bohemia famously exchanged a

correspondence with Descartes about his writings. In her first letter, dated 6 May 1643 (see Bennett, 2009), she set down her question as follows:

> "Given that the soul of a human being is only a *thinking* substance, how can it affect the bodily spirits, in order to bring about voluntary actions?"

and continued:

> "The question arises because it seems that how a thing moves depends solely on (i) how much it is pushed, (ii) the manner in which it is pushed, or (iii) the surface-texture and shape of the thing that pushes it. … The first two of those require *contact* between the two things, and the third requires that the causally active thing be extended. Your notion of the soul entirely excludes extension, and it appears to me that an immaterial thing can't possibly *touch* anything else"

Similar arguments are still being made. Kim (2005) argues that: "… our idea of causation requires that the causally connected items be situated in a space-like framework" (p 84) and uses that presumed fact in making the case that the proposition that events in one substance can influence events in another substance must be incoherent

Indeed, two of the problems most commonly cited that arise regarding mental causation in the particular case of dual substances are: the problem of spatial contiguity and the problem of energy conservation. But neither of these seems fatal. Although spatial contiguity is generally a feature of causation in physics one could argue that there is no reason to assume that mental causation, a notion external to the physical sciences, will follow the conventions of physics (indeed, "quantum entanglement", where the quantum state of one particle cannot be described independently of that of another, even when they are separated by a large distance, appears to be an exception to this proposition). As for energy conservation, there seems to be no reason why energy conservation should not hold for all processes not involving mental causation while cases of mental causation could result in increases in the total energy in the universe, unless offset by sinks elsewhere (see Section 2.3.3). The proposition that mental causation operates along with dual substances is known in the literature as 'interactionism', but it is not widely advocated.

Interestingly, Leibniz addressed this issue in his 1695 essay *A New System of the Nature and Communication of Substances*. He accepted that events relating to mind and body appear to develop in total harmony with each other, as if there were some causal process

between the two, and he took the view that this situation arose because of pre-programming by God:

> "God ... created each soul ... in such a way that everything in it arises from its own depths ... with no causal input from anything else ... and yet with a perfect conformity to things outside it. ...
>
> "The organised mass in which ... the soul lies ... is ready, just when the soul desires it, to act of itself according to the laws of the bodily mechanism ... the motions that correspond to the passions and perceptions of the soul.
>
> "All this happens without either body or soul disturbing the laws of the other; it is a mutual relationship, arranged in advance in each substance in the universe" (para. 14, translation by Bennett, 2004).

The idea that God has set the system up so that mental states always seem appropriate to the material state that is instantiated seems singularly unappealing as an explanation of the seeming efficacy of mental causation nowadays.

3.2.4 Conclusion on substance dualism

It seems that many Western philosophers now reject the substance dualism view. This may be largely because they think that everything related to the material world can be 'explained' by things to do with physics, that is, they believe in the completeness of physics (see Sections 2.3 and 2.4), and that leaves no causal role for a second substance. But substance dualism is not obviously false. It cannot be rejected as the result of any of the fundamental principles set out in Chapter 1, although it certainly seems an unlikely explanation of the existence of consciousness in this materialistic age given our increasing understanding of the manner in which mental roles are associated with different structures in the brain. Whereas the problems with mental causation may make substance dualism seem implausible, there are problems relating to mental causation associated with all the other potential explanations for the existence of consciousness. I don't discuss the substance dualism approach in any further depth in this thesis, but we will encounter it again in Chapter 5, in the context of the questions relating to free will.

3.3 Reductive Physicalism

3.3.1 Introduction

An approach to the nature of mind that has been popular with philosophers, neuroscientists and physicists in recent years, and gets around the problem of mental causation, is reductive physicalism, the proposition that mental states can be ontologically reduced to the properties of matter that are studied in physics.

In this section we first examine the identity thesis. Then we turn to the reductive physicalism proposed by Place (1956) and Smart (1959): the proposition that the states and processes of the (so-called) mind can be "ontologically reduced to" the properties of matter that are studied in the physical sciences. A problem that arises then is the possibility of an explanatory gap (Levine, 1983): the idea that although mental states are ontologically reducible to the properties of matter studied in physics, we are unable to explain perfectly all of their features in terms of those properties, so there is an explanatory gap. This suggests that that the approach known as 'concept dualism' should be considered, and the section ends with a description of that approach.

As we have seen (Section 1.2.5) it is usually the case nowadays that when philosophers use the term 'physicalism' they mean reductive physicalism. Figure 4 shows a comparison of reductive physicalism, in which there is no such thing as a mental property that is not ontologically reducible to physical properties, and supervenience physicalism, in which, although mental properties are related to the physical properties of matter, they are ontologically different. Using the words of C. D. Broad (1947) when describing "emergentism" (see Section 3.4.9 for the complete quote) this can be described as the case in which: "… the characteristics of the [mental state] cannot, even in theory, be deduced from the most complete knowledge of the [physical properties upon which it supervenes]" (my insertions).

Place (1956) and Smart (1959) were among the first to take seriously the proposition that mental states can be completely understood in terms of facts relating to the properties of matter that are studied in physics; indeed, they are often cited as being the first proponents in modern times of what is often called "the identity thesis".

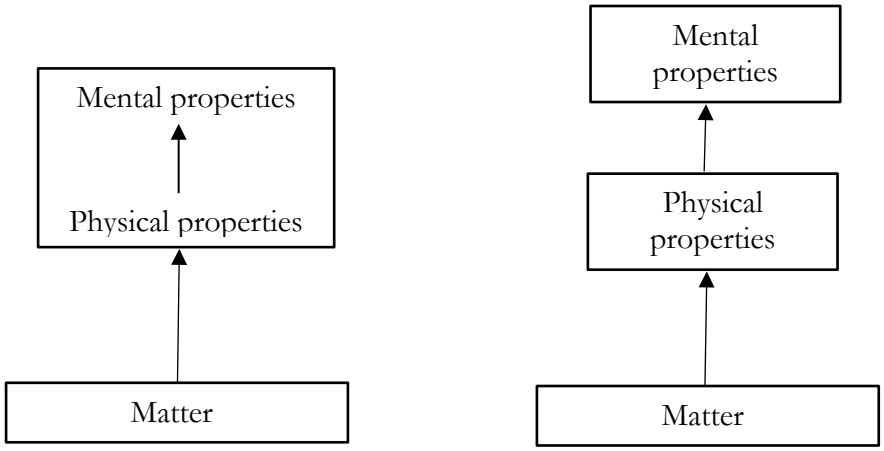


Figure 4: Reductive physicalism and supervenience physicalism

In the case of reductive physicalism, the mental properties are simply a combination of physical (non-mental) properties by another name. In supervenience physicalism the mental properties are related to the physical (non-mental) properties but cannot be ontologically reduced to them.

3.3.2 The identity thesis

Can mental states be identical to brain states?

Many philosophers nowadays apparently take the view that it could be the case that mental states are *identical* to brain states, in the sense of these being two names for the same thing. It seems unlikely that it would ever occur to most ordinary people pondering about these things as they go about their business that this could be the case. Mental states, states of which a creature could be conscious, consist only of thoughts and feelings. Thoughts and feelings are private and internal; they can be directly observed only by the person who experiences them. The brain, however, is a lump of matter; it occupies space and has mass and looks a certain way when extracted from the skull and set down upon a table; it has a number of properties at any point in time, such as electric currents and chemical reactions as well as weight and texture, that are public in the sense that, in principle, they could be examined and noted by any number of people. So a brain state is a particular situation

regarding that lump of matter and a description of a brain state might include any or all of the properties that obtain regarding that lump of matter at some point in time.

If mental states were identical to brain states, the attributes of each would have to be exactly the same; as Quine (1987, p 89) says: "A thing is identical with itself and with nothing else …". And whereas the proposition that thoughts and feelings are an aspect of a brain state is probably acceptable as a possibility to most of us, the proposition that chemical reactions and electric currents are an aspect of what we call a "mental state" is unacceptable, because what we refer to as a "mental state" consists *only* of thoughts and feelings and these are private and can *only* be accessed directly by the person experiencing the mental state.

If what the identity theorists assert is that brain states and mental states are identical, in the sense of the referents these two terms being the same thing, then given the way the two terms are generally used they are plainly in error. The fact that people *do* assert that they are identical may be the result of a paper by Lewis, published in 1966, entitled: 'An Argument for the Identity Thesis', see below. But before proceeding we should note that the identity thesis is a *logical* impossibility; the usual definitions of the terms 'mental state' and 'brain state' are such that these terms *cannot* have the same reference, just as a bachelor cannot be a married man.[51]

Lewis on the identity thesis

Lewis, in his 1966 paper in support of the identity theory appears to disagree with the argument presented above, saying: "… it is pointless to exhibit various discrepancies between what is true of experiences as such and what is true of neural states as such" (p 19). The following is Lewis' argument in his own words:[52]

> "The identity theory says that experience-ascriptions have the same reference as certain neural-state-ascriptions: both alike refer to the neural states which are experiences. It does not say that these ascriptions have the same sense. They do not; experience-ascriptions refer to a state by specifying the causal role that belongs to it accidentally, in virtue of causal laws, whereas neural-state-

[51] In this context it is interesting to note that it seems that this was the view taken by many when Place, wrote his seminal 1956 paper 'Is Consciousness a Brain Process?' in support of the identity thesis. We know this because of the comment he made regarding: "The all but universally accepted view that an assertion of identity between consciousness and brain processes can be ruled out on logical grounds alone …" (*ibid.*, p 45, and see Section 3.3.3).
[52] Smart made a similar point, see Section 3.3.3.

> ascriptions refer to a state by describing it in detail. Therefore the identity theory does not imply that whatever is true of experiences as such is likewise true of neural states as such, nor conversely. For a truth about things of any kind *as such* is about things of that kind not by themselves, but together with the sense of expressions by which they are referred to as things of that kind. So it is pointless to exhibit various discrepancies between what is true of experiences as such and what is true of neural states as such. We can explain those discrepancies without denying psychophysical identity and without admitting that it is somehow identity of a defective sort" (*ibid.*, p 19, see later for Lewis' footnote).[53]

But if the identity theory holds, so the terms 'experiences' and 'neural states' both have the same referent, there would, of course, be no discrepancies between what is true of that referent when one term is used and what is true when the other is used. In the case of a numerical identity, the situation in which two terms pick out one thing, *all* the properties of the thing referred to must be *exactly* the same, by definition, regardless of which term is used to describe it, because the thing is one thing referred to by two words.[54] So in spite of what Lewis says here, it *has* to be the case that if one can exhibit discrepancies between what is true of experiences and what is true of neural states that is enough to establish that the identity theory does not hold. So what is wrong with the argument put by Lewis, above?

In the footnote relating to the above paragraph Lewis refers to Frege's paper: 'On Sense and Reference', saying "Here I have of course merely applied to states Frege's doctrine of sense and reference (see Frege, 1892, translated by Geach and Black, 1960)". So before examining Lewis' argument in greater depth we should check what Frege says on this point. Unsurprisingly, Frege does not present a "doctrine of sense and reference" that enables the 'reference', the thing that is picked out by two different names (or 'signs') that we humans have allocated to it, to differ in its properties depending on the 'sense' of the sign used to refer to it.

---

[53] Kripke appears to support this point of Lewis', if somewhat half-heartedly, in his 1972 book *Naming and Necessity*, when he writes (pp 40-41): "… whether an object has the same property in all possible worlds depends not just on the object itself, but on how it is described. So it's argued."

[54] The word 'identity' has two meanings; it can mean that two words have the same referent (numerical identity), or it can mean two things have the same intrinsic qualities (qualitative identity). In philosophy, if the term 'identity' is not qualified, it takes the first meaning. However, the grammar of saying that two things are 'identical' when the situation is that there are two words for one thing can cause confusion.

Frege starts his paper saying (1892, p 36):

> "Equality gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects? In my … I assumed the latter. The reasons which seem to favour this are the following: a = a and a = b are obviously statements of differing cognitive value; a = a holds *a priori* … while statements of the form a = b often contain very valuable extensions of our knowledge and cannot always be established *a priori*. … What is intended to be said by a = b seems to be that the signs or names 'a' and 'b' designate the same thing, so that those signs themselves would be under discussion; a relation between them would be asserted. But this relation would hold between the names or signs only in so far as they named or designated something. It would be mediated by the connexion of each of the two signs with the same designated something."

He then sets up an example to illustrate his point (*ibid*., p 37). He says: "Let a, b, c be the lines connecting the vertices of a triangle with the midpoints of the opposite sides. The point of intersection of a and b is then the same as the point of intersection of b and c." At this point Frege introduces the concept he calls the 'sense' of the sign, which was picked up by Lewis, above. Frege says:

> "It is natural, now, to think of there being connected with a sign …, besides that to which the sign refers, which may be called the reference of the sign, also what I should like to call the *sense* of the sign, wherein the mode of presentation is contained. In our example, accordingly, the reference of the expressions 'the point of intersection of a and b' and 'the point of intersection of b and c' would be the same, but not their senses. The reference of 'evening star' would be the same as that of 'morning star', but not the sense" (*ibid*., p 37).

Thus Frege makes it quite clear that when one object is picked out by two or more different signs, the 'sense' of the signs will differ, but the properties of the one object to which both of the signs refer, the 'referent', will remain unchanged.

The caveat concerning Lewis' argument, quoted above, relates to his use of the term 'as such'. Clearly, whatever is true of experiences *as such* would not be necessarily true of neural states *as such* whether or not the identity theory holds; each picks out just one aspect of or one situation relating to what one might call a brain state. If we consider Hesperus *as such* we pick out the object in a certain position in the sky at a certain time, and that is certainly not the same thing as Phosphorus *as such*. But if we want to consider whether the object that is picked out by the name "Hesperus" is the exact same object as

that picked out by the name "Phosphorus" we will need to seek discrepancies between the object(s) the two names pick out. If the two names pick out the same object there will be no discrepancies in the properties of the object picked out by the two, different names.[55]

The identity theory says that what is picked out by the term 'mental state' (or 'experiences') is the exact same thing as that which is picked out by the term 'brain state' (or 'neural state'), and addressing that question involves investigating the existence of discrepancies between the *referents* of the two terms; what is true of mental states and what is true of brain states. In the case of mental states and brain states, as usually defined, there are indeed discrepancies; for example, brain states have mass and electric charges, while mental states have neither. There can therefore be no identity between mental states and brain states on the usual definitions.

The proposition that an object out there can change its properties as a result of us changing the name we use to describe it (because the *sense*, the situation, is different) is clearly absurd.[56]

Could a mental state be a property of a brain state?

Since it cannot be that the term "mental state" means the same thing as the term "brain state", it may be that the identity theorists have something else in mind. For example, it clearly could be the case that, just as a cylinder can look like a circular disc from one angle and like a rectangle from another, so a brain state could be viewed as a mental state from one point of view, in that it could instantiate thoughts and feelings, and yet be viewed as being a mush of pinkish-white material with electric currents and chemical reactions from another.

It is obviously impossible for a mental state *per se* (which by definition consists only of thoughts and feelings) to be perceived as having *all* the properties of a brain state, since the brain is an object with weight, chemical reactions and electric currents and occupies space, but if a mental state is related to a brain state, and we are assuming in this thesis that that is the case, it could be that the brain instantiates mental states; it could be that instantiating mental states is one of the many attributes of the matter incorporated into a brain. So a mental state could perfectly well be an instantiation of some property (or

55 The term "lump of clay" could refer to the same object as the term "statue of David" depending on the circumstances of the reference, but "lump of clay as such" would not.
56 Could support for the identity thesis in this form be a real-life example of "the emperor has no clothes"?

properties) of a brain; consciousness could indeed be a brain process, as suggested by Place in 1956 (see Section 3.3.2). A mental state could then be identical to, in the sense of being exactly the same thing as, *one aspect* of a brain state; indeed, in this thesis I put the case that that is exactly what it is. That could be what the identity theorists have in mind. Some particular brain state can then be picked out either by referring to some aspect of the mental state associated with it or by referring to any of its non-mental properties; but the brain state as a whole is no more *identical* to its mental properties than it is identical to any one of its other, non-mental, properties, such as the electric currents that pertain, or its weight.

Indeed, if this thought about the nature of mental states is correct, the mental state could be said to belong in a different category from that of the brain state: whereas a brain state would be the situation regarding *all* the properties of the particular lump of matter we call a brain, the non-mental properties and the mental properties, the mental state would be the situation regarding just one type of property.

Could mental states be explicable in terms of physics?

Finally, what philosophers often seem to mean when they say that mental states are identical to brain states is that mental states are an aspect of a brain state and that this has the implication that mental states are totally explicable in terms of the properties of matter studied in the sciences we call 'the physical sciences'.

But is it actually the case that if a mental state is a property of a brain state all the facts relating to mental states, facts relating to feelings and facts relating to thoughts, must be suitable for study in the physical sciences? Not necessarily, is the answer. There is an implicit assumption in this version of the proposition that mental states are identical to brain states, and it is that everything about what we call brain states is totally explicable in terms of the physical sciences. Clearly, it could also be the case, as is suggested in this thesis, that there is another property of matter, and that the facts relating to this other property of matter are different in some crucial way from those that are studied in the physical sciences, in which case the proposal that philosophers call 'non-reductive physicalism' (Section 3.4) would hold.

If this were the case there could be an aspect to any state of a brain, namely the mental state that is associated with it, which is not suitable for study in the physical sciences, a view that is at variance with the meaning usually attributed to the assertion: "consciousness is a brain process". The identity thesis, then, would not hold.

Conclusion on the identity thesis

It is clear that although mental states cannot be identical with brain states *per se*, they can be instantiations of a property of a brain state; it is also clear that it does not necessarily follow from the statement that a mental state is a property of a brain state that mental states can be ontologically reduced to physics.

3.3.3 Is consciousness a brain process?

Could the mind-brain relation be a case of "the 'is' of composition"?

What U T Place (1956) was suggesting in his famous paper 'Is Consciousness a Brain Process?' was that it is not necessarily the case that: "sensations and mental images form a separate category of processes over and above the physical and physiological processes with which they are known to be correlated" (*ibid.*, p 44). He says that there was, at that time, an: "all but universally accepted view that an assertion of identity between consciousness and brain processes can be ruled out on logical grounds alone" (*ibid.*, p 46, and see Section 3.3.2). Place began to doubt this when he discovered that to a physiologist (Sir Charles Sherrington) the question was not seen to be a matter of logic. Instead Place found that what worried the physiologist was, in Place's words: "the apparent impossibility of accounting for the reports given by the subject of his conscious processes in terms of the known properties of the central nervous system" (*ibid.*, p 48). Place took this as indicating that, to the physiologist at least, such an account of conscious properties could, after all, be a practical possibility, and clearly, if that were so it couldn't be the case that it is ruled out by logic ("if something is conceivable it is logically possible", see Chalmers, 1996, and also Descartes, 1641).

Place starts to spell out his proposal by describing a distinction that can be made between two different types of statement involving the word 'is'. He called these "the 'is' of definition" and "the 'is' of composition", citing "a square is an equilateral rectangle" as an example of the first, and "a cloud is a mass of water droplets or other particles in suspension" as an example of the second; and he noted that in both cases it makes sense to add the phrase "and nothing else". Both of these are distinct from the 'is' of predication, he noted (giving "her hat is red" as an example), which does not concern identities and to which it does not make sense to add the phrase "and nothing else". These two types of statement are strikingly different, he says, for the 'is' of definition is necessarily true, while

the 'is' of composition has to be verified by observation (see Section 1.2.3).[57] Two familiar numerical identities, those concerning water and H_2O, and light and a stream of photons, are examples of the use of the 'is' of composition.

Place then suggests that "consciousness is a brain process" might be an example of the 'is' of composition, in which case it would be analogous to the cases of water and H_2O or light and a stream of photons (in this thesis I call identities of this type, for convenience, "composition identities"; this is because the reason for the existence of such an identity is the discovery, by scientists, of the composition of something that is colloquially described by one of the two terms to the identity). If this were the case it would be, as Place says, a reasonable scientific hypothesis, not necessarily true and not necessarily false; a hypothesis that we should be able to test empirically.

J. J. C. Smart (1959), who participated in the discussions surrounding Place's ideas, in Australia and elsewhere, took a slightly different approach from that of Place but it led him to the same proposition. He reckoned that sensations, or states of consciousness, were the only things that were being left out as developments in science increasingly enabled creatures to be understood in terms of complex physiochemical mechanisms saying: "Such sensations would be 'nomological danglers' …", and remarking that the laws by which they would dangle would be "odd". He said in this context (see also Section 2.2.4): "I cannot believe that ultimate laws of nature could relate simple constituents to configurations consisting of perhaps billions of neurons (and goodness knows how many billion billions of ultimate particles) all put together for all the world as though their main purpose in life was to be a feedback mechanism of a complicated sort" (*ibid*., p 143). So he took the line that the issue would have to be decided on the basis of parsimony and simplicity, or, put another way, using "Occam's razor"; and in his view this overwhelmingly favoured the identity thesis.[58]

Smart therefore signed up to the proposal that the relation between sensations and brain processes could be an identity in which one of the two names picks out as referent a familiar object "out there", that we identify with our senses, while the other picks out what the object out there is composed of, as is the case with a 'flash of lightning' and an 'electrical

[57] The word 'necessary' can have significantly different types of implications, see Section 1.2.2.
[58] We might note, for later use, that the same type of argument can be used in support of mental causation as an explanation for action, as in "Johnny went to get a drink because he was thirsty", where the alternative explanation of Johnny's actions relates to a multitude of chemical imbalances and electric currents involving neurotransmitters and muscle contractions, and requiring goodness knows how many billion neuron interactions.

discharge'. Smart said: "... there are not two things: a flash of lightning and an electrical discharge. There is one thing ..." (*ibid*., p 146) and: "When I say that a sensation is a brain process or that lightening is an electric discharge, I am using 'is' in the sense of strict identity" (*ibid*., p 145). But he also reckoned that: "there is no conceivable experiment which could decide between materialism and epiphenomenalism" (*ibid*., p 155).

Why this is not an identity that relates an object's name to a description of its composition

It is perfectly easy to see that the proposed analogy between the mind state/brain state relation and the identity statements that relate the name of an object to a description of its composition simply does not work.[59]

In the case of the identities that relate an object's name to a description of its composition there are in total *three* descriptions identifying things in the analysis: first, the bundle of sense impressions received by a person in a certain set of circumstances; second, the thing "out there" that is generally thought by us humans to cause those sense impressions; and third, a description of a thing "out there" which, although seemingly it might not have been the case, has actually been found by scientists to be the composition of the thing that causes the particular bundle of sense impressions.[60] We humans identify the thing out there that causes a person to experience the "waterish *qualia*" (what it looks like, feels like, tastes like *etc*.) as what we call 'water', but scientists have explained to us that what it actually is, is a collection of H_2O molecules in their liquid phase. Hence water is the exact same thing as H_2O molecules. So the identity under consideration does not involve our *experience* of water, it concerns two things we have identified as being out there, which it was originally thought might be different, but which have now been found by scientists to be the same thing. Each of the two things that are identical (that is, that are actually the same thing but have been given different names) in cases relating to the identities that arise in this way are things out there; neither of them is a sensation experienced by a person. It is, very obviously, not the case that our experience of watery properties is the same as – identical to – the thing that is called, alternatively, 'water' or 'H_2O molecules'. The same is true of lightning and electric discharges, and of all the other identities that arise in this way and occur in the literature, as well as any new ones that might be dreamt up.

[59] See also Kripke (1972).

[60] By a thing "out there" I mean a thing that we believe exists in the world independently of a person being around to note its existence.

In the case of the brain state/mental state relation there are only two descriptions identifying things. One is a collection of sensations (thoughts and feelings), the other is a thing out there (a collection of neurons and other material). How could that possibly turn out to be analogous to the water/H_2O identity? If the two are identical (and, *ab initio*, it seems as if they could be) it will simply because the same object has been given two names (as is the case for "Phosphorus = Hesperus"). It could not be an identity in which one of the terms describes the composition of the thing picked out by the other term so that we can say: "the thing which we perceive through sensations S and have named A, is identical to/is another name for, the thing scientists/others have named B".

It is clear that if mind is a property of matter, and if the matter that instantiates this property is in the brain, consciousness is, indeed, a process in the brain, as Place proposed. However, consciousness could be said to be a brain process if *any* form of physicalism were to hold (so long as mind is a property of the matter in the brain rather than in some other bodily part), and there is no need for reductive physicalism to be true for Place's point concerning consciousness being a brain process to be correct. So there remains a question as to whether, given that the proposed relation between mental states and brain states is not a case of identifying the name of an object with a description of its composition, it is nevertheless the case that mental states are ontologically reducible to the properties of matter studied in physics.

An issue relating to communication

Both Place and Smart identify the communication of facts relating to mental states as being an important issue in the consideration of the ontological relation between mental states and brain states. Place writes:

> "I am not trying to argue that when we describe our dreams, fantasies and sensations we are talking about processes in our brains. That is, I am not claiming that statements about sensations and mental images are reducible to or analysable into statements about brain processes … To say that statements about consciousness are statements about brain processes is manifestly false. This is shown (a) by the fact that you can describe your sensations and mental imagery without knowing anything about your brain processes or even that such things exist, (b) by the fact that statements about one's consciousness and statements about one's brain processes are verified in entirely different ways and (c) by the fact that there is nothing self-contradictory about the statement "X has a pain but there is nothing going on in his brain" (*ibid.*, p 44-45).

So Place is suggesting that whereas he finds that *statements* about consciousness are patently not reducible to *statements* couched in the terms of the physical sciences, consciousness is not *explainable in terms of* the physical sciences, it could perhaps, nevertheless, be the case that the *facts* about consciousness are determined by the *facts* relating to the physical sciences, consciousness is ontologically reducible to the physical sciences. What he is describing here is known nowadays as 'concept dualism', see Section 3.3.5.

Although Smart argues that sensations *are* brain processes (for reasons of parsimony) he also explicitly makes clear that in his view: "It is not the thesis that, for example, 'after-image' or 'ache' means the same as 'brain process of sort X' (where 'X' is replaced by a description of a certain sort of brain process)" (*ibid.*, p 144). Whereas Smart takes the view that the sensation associated with the after-image *is* a certain sort of brain process, he also sees the use of the term 'after-image' as picking out a different aspect of the sensation/brain process to that which is picked out by the term 'brain process of type X'. Like Place, he is suggesting that there is a distinction between the way in which *facts* about mental states and brain states relate to one another, and the way in which *statements about those facts* relate to each other.

3.3.4 The explanatory gap

In 1986, Levine pointed out that people who believe in materialism, in the sense that they believe that mental states are nothing more than physical (pertaining to physics) states of the brain, have a problem with what he called "the explanatory gap" (of course the thought that motivates this problem was not new, see, for example, Nagel, 1974 and Kirk, 1974).

Levine started by re-iterating Kripke's argument against physicalism (see Kripke, 1972), saying that a physicalist is committed to identity statements, statements of the "two words for one thing" variety that take the form in his words: "pain is the firing of C-fibers", and that the statement: "heat is the motion of molecules" is a similar type of identity statement (Levine, 1986, p 354).[61] He then notes that there is what he calls "a felt contingency" (*ibid.*, p 355) about both of these statements. And, indeed, neither can be known to be true without some experimentation; so, neither can be said to be true *a priori*,

[61] See footnote in Section 2.6.2 on the meaning of 'C-fibers', a term used figuratively throughout this particular area of philosophy of mind to represent the part of the brain that enables us to experience pain, and which I refer to here as 'pain neurons'

both can only be discovered to be true, if they are true, following experimentation. However, Levine takes the view that there is an important difference between them that relates to this "felt contingency". When one finds it conceivable that heat is not the motion of molecules, in the sense that one can imagine heat existing in the absence of the motion of molecules, he says that what one is really imagining is that there is: "some phenomenon that affects our senses in the way heat in fact does, but is not the motion of molecules" (*ibid.*, p 355). But, he says, this sort of explanation will not work in the case of pain and the firing of 'pain neurons': "for the simple reason that the experience of pain, the sensation of pain, counts as pain itself. We cannot make the distinction here, as we can with heat, between the way it appears to us and the phenomenon itself. Thus, we have no good account of our intuition that [the pain/'pain neurons' firing identity] is contingent unless we give up the truth of [this identity] altogether", (*ibid.*, p 355, my insertions).

Levine refers to this argument, set down by Kripke (1972), as providing an "intuitive resistance to materialism" (Levine, 1986, p 356), and takes the view that it: "should not be shrugged off as *merely* a matter of epistemology" (his italics). Indeed, he suggests that the underlying difference between the two statements, which he thinks explains the difference identified by Kripke, is that in the case of heat and the motion of molecules the statement: "expresses an identity that is *fully explanatory*, with nothing crucial left out" (*ibid.*, p 357), while in the case of pain and the so-called 'C-fiber' (pain neuron) firing he says that the statement does: "seem to leave something crucial unexplained, there is a 'gap' in the explanatory import of [this statement]" (my insertion). More definitively he says: "… there is nothing we can determine about C-fiber firing that explains why having one's C-fibers fire has the qualitative character that it does …", and: "We don't have the corresponding intuition in the case of heat and the motion of molecules … because whatever there is to explain about heat is explained by its being the motion of molecules" (1986, p 359). Levine calls this phenomenon, that relates to explanations of mental states, "the explanatory gap".

Clearly, both the conceivability argument ("that there is no contradiction or incoherence in the extreme zombie hypothesis, the idea of a microphysical duplicate of one of us but with no consciousness", see Block and Stalnaker, 1999, p 29) and the knowledge argument ("that Jackson's Mary cannot deduce the facts about what it is like to see red from the microphysical and functional facts", *ibid.*, p 29) depend on the existence of an explanatory gap. But both of these are taken to carry the implication that "physicalism is false", that consciousness is not ontologically reducible to the properties

of matter studied in physics; so the obvious question that arises from this proposal made by Levine is: if there is an explanatory gap of this nature, does this mean that there is also an ontological gap? Are Chalmers and Jackson right to assume that the conceivability argument and the knowledge argument carry the implication that there is no ontological gap – that mind, and therefore mental states, in fact cannot be explained in terms of physics? Or could it be that the problem is merely that we humans are unable, for some reason (see later), to explain mental states in terms of physics in a manner that satisfies us?[62]

Block and Stalnaker (1999) address the question of whether the proposition that the existence of an *explanatory* gap carries with it the implication that there must be an *ontological* gap, (note that Levine, also, later, suggested that: "the explanatory gap is primarily an epistemological problem, not necessarily a metaphysical one", *Purple Haze*, 2001). They start by splitting the assertion that water is H_2O into two steps. First, they note that experts have given the name 'water' to the stuff out there that can be described, in Jackson's words (1993, p 39), as that which: "falls from the sky, fills the oceans, is odourless and colourless, is essential for life". Block and Stalnaker summarise Jackson's description as "the waterish stuff". So, this definition can be stated as:

Water = the waterish stuff.

Second, they take it that the fact that H_2O has the properties that affect our senses in a certain way (it is odourless, colourless, *etc.*) can be deduced from its atomic structure using the laws of nature that make up physics (in their words: "is … derivable from microphysics", see p 12).[63] It follows that scientists should be able to tell us that the stuff they call H_2O must have the properties that will affect our senses in the manner Block and Stalnaker have designated for the stuff out there which they call "the waterish stuff"; therefore:

H_2O = the waterish stuff.

It follows immediately from these two statements, one a definition, the other the consequence of the laws of physics, that:

[62] See McLaughlin (2010) for a description of how type physicalism can be justified given the existence of a conceptual gap using inference to the best explanation.
[63] These laws of nature cannot be known *a priori*, of course, they do not follow from the definitions of words, and they cannot be known from introspection; they are *a posteriori*, see Section 1.2.3.

Water = H_2O.

Block and Stalnaker (*ibid.*, p 12) then raise the question of whether the same analysis applies for consciousness; whether there is any reason why the facts relating to consciousness cannot be derived from the laws of nature that make up physics in spite of the fact that, according to them: "the concept of consciousness cannot be given an analytic definition in functional or physical terms" (*ibid.*, p 13); put another way, the question they are exploring is whether reductive physicalism, ontological reduction of consciousness to the properties of matter studied in physics, can hold in spite of the fact that there is an explanatory gap.

Accepting for present purposes what Levine says about the perceived existence of an explanatory gap, if we wish to maintain the possibility that reductive physicalism also holds there are various possibilities. First, it could be the case that whereas at this point in the development of science we are unable to derive what it is like to experience something from the laws of physics, as science develops further we may become able to do this; the explanatory gap we perceive today may become closed in the future. Second, it could be that the explanatory gap Levine has identified results from an inherent inability on the part of us human beings to fully comprehend consciousness (see van Gulick, 1985, McGinn, 1989 and others). Third, it is worth considering the possibility that in spite of Levine's case for the perceived existence of an explanatory gap, we are, in fact, mistaken in thinking that there is one (see Papineau, 2011).

McGinn (1989) makes an excellent case for the possibility that: "we are cut off by our very cognitive constitution from achieving a conception of that natural property of the brain (or of consciousness) that accounts for the psychophysical link" (p 350). He introduces the concept of 'cognitive closure', saying: "A type of mind M is cognitively closed with respect to a property P (or theory T) if and only if the concept-forming procedures at M's disposal cannot extend to a grasp of P (or an understanding of T)". He then gives examples of what one might call perceptual closure (noting that the invisible parts of the electromagnetic spectrum are just as real as the visible parts although we are unable to perceive them) and of cognitive closure from the animal kingdom, citing monkey minds and the property of being an electron. "Nothing …", he says, "in the concept of reality shows that everything real is open to the human concept-forming faculty …" (*ibid.*, p 351).

Indeed, he says he finds it: "… deplorably anthropocentric to insist that reality be constrained by what the human mind can conceive" (*ibid.*, p 366) and adds that: "… in the case of the mind-body problem, the bit of reality that systematically eludes our cognitive grasp is an aspect of our own nature" (*ibid.*, p 366).

Papineau (2011) puts the case against the existence of an explanatory gap. He says: "The feeling of a 'explanatory gap' arises only because we cannot stop ourselves thinking about the mind-brain relationship in a dualist way" (2011, p 18). He lists a number of reasons which might explain why this is the case, and the more convincing of these are somewhat similar to McGinn's thesis.

First, he says that it could simply be that our culture implicitly assumes that dualism is the case so we are all brought up to take it for granted. But Papineau is not convinced by this; he feels that there is, actually: "some feature of our cognitive architecture that forces the intuition of dualism on us" (p 15).

The next reason he gives for why we might think there is an explanatory gap when, actually, there isn't one, he calls 'the antipathetic fallacy'. The idea is that people might implicitly assume that introspection ought to reveal the true nature of consciousness. But when we think about a phenomenal experience, such as the experience of seeing something red, we typically recreate that experience in our imaginations. So introspection does not reveal mental states to be physical. So people conclude that mental states could not be physical. Therefore, Papineau says, people think that the statement: "The experience of seeing something as red = such-and-such neural activity in V4", cannot be true (*ibid.*, 2011, p 16). But Papineau finds this unconvincing because people who have not thought much about the mind-body issue would then be less likely to take a dualistic view, and this appears not to be the case.

Then, there is the possibility that we have two distinct cognitive systems, one for thinking about mental processes and the other for thinking about physical processes, as described by Paul Bloom is his book *Descartes' Baby* (2004). Paul Bloom, a psychologist at Yale University, has been exploring the development of intuitions relating to moral responsibility by looking at morality in babies. He says:

> "We can explain much of what makes us human by recognising that we are natural Cartesians – dualistic thinking comes naturally to us. We have two distinct ways of seeing the world: as containing bodies and as containing souls. These two ways of seeing the world interact in surprising ways in the course of development of each child, and in the context of a community of humans they

> give rise to several uniquely human traits, such as morality and religion." (Bloom, 2004, p xii).

Here we are getting close to the possibility that there *is* an explanatory gap and it results from our inability to switch information between our two cognitive systems.

Papineau's final suggestion for why we think there is an explanatory gap when there isn't follows from a thought of Melnyck (2003). Melnyck, points out that, in Papineau's words: "… what happens when we accept any identity claim of the form a = b is that we 'merge the files"' (*ibid*., p 18). So what if for some reason we aren't able to merge the files of mind-brain identities? If phenomenal concepts and physical concepts are realised in different parts of the brain it may be that we cannot merge the files. Perhaps this could explain why we think there is an explanatory gap.

The conclusion seems to be that, although there appears to be an explanatory gap, this does not necessarily mean that there must be an ontological gap; it still could be the case that mental states are identical to brain states. And if that were the case, concept dualism would be a natural approach to take to the problem of the nature of mind.[64]

3.3.5 Concept dualism

What Place and Smart were suggesting (see Section 3.3.3) was in fact an early version of the approach to reductive physicalism that nowadays takes the name 'concept dualism' and was later enunciated by McLaughlin (2010), Loar (1990/97), Papineau (2002) and others. As Loar puts it as he sets out the rationale for concept dualism at the beginning of the 1997 version of his paper: "… the present objective is to engage anti-physicalist arguments and entrenched intuitions to the effect that conscious mental qualities cannot be identical with ordinary physical properties, or at least that it is problematic to suppose that they are so".

The idea is that it may be the case that whereas conscious mental properties, thoughts and feelings, can be identical with ordinary physical properties (that is, 'ontological reduction' can hold, see Crane, 2001, p 54), there is no good reason why there should not, at the same time, be an 'explanatory gap': it could be that mental events, thoughts and feelings, cannot be *explained* in terms of the theory relating to the science of physical properties (that is, it may be that 'explanatory reduction', which requires that

[64] So long as the existence of an explanatory gap without an ontological gap is consistent with the reproducibility of experiments relating to the properties of matter studied in physics.

physical theory can explain propositions relating to mental states, does not hold, see Crane, *ibid*.). In Chalmers words: "consciousness cannot be reductively explained, but is physical nonetheless" (Chalmers, 1996, p 166).

Papineau, spelling out the idea carefully in his book (2002), proposes that there are two different types of 'concept' that can be used to refer to conscious properties: 'phenomenal concepts' and 'material concepts'.[65] In Papineau's words, "The general idea is that when we use phenomenal concepts, we think of mental properties, not as items in the material world, but in terms of *what they are like*" (his italics, *ibid*., p 48). He suggests we think of what it would feel like if the dentist's drill slips and hits the nerve in your tooth. 'Material properties', on the other hand, are the familiar physical (non-mental) properties that are associated with objects in the external world, such as dogs, or houses. Examples of material properties associated with a dog might be "hairy", or "four-legged", but they could also be something physical related to the condition of the dog's neurons at some point in time. The material concept associated with a conscious property is then a description of what happens physically (non-mentally) when the creature experiences the particular conscious property at issue, that is, what happens physically, inside the creature's body, when it sees the colour red or feels thirsty. Papineau emphasises that: "Material concepts are those which pick out conscious properties *as* items in the third-personal, causal world" (his italics, p 48).

However, Papineau also signs up to the proposition that everything relates to the properties of matter studied in physics, saying: "I believe that in the end the materialist argument wins", and by 'materialist' Papineau means physicalist in the narrow sense (*ibid*., p 14). The reason is that he accepts the requirement that physics is complete in the sense that all physical effects are fully caused by purely physical prior histories. But he also

[65] The term 'concept' has been much discussed in philosophy (see Margolis and Laurence, 2014, for a discussion of the meaning of 'concept'). In this thesis, the term 'concept' picks out those bundle of entities (or a single entity) which get stored in the long-term memory as a bundle and turn up in the working memory on a fairly regular basis. This arises when the bundle is relevant to some material or abstract, real or imaginary, object. In such a case it is convenient, for communication purposes, to attach a verbal label to the bundle (or single entity); we call such bundles 'concepts'. The words 'red', 'dog', 'tree', 'stone', 'happiness', 'unicorn' can be viewed as picking out concepts. The inclusion in the bundle of some element relating to what it is like to experience something is necessary if a concept is to be classed as a phenomenal concept ('red' and 'happiness' are phenomenal concepts; 'stone' is not; for 'dog', 'tree' and 'unicorn' the question of whether or not they pick out a phenomenal concept could depend on the context).

assumes that conscious mental states can have physical effects and that there is no "over-determination". So he claims that mental states must be physical:

> "Many effects that we attribute to conscious causes have full physical causes. But it would be absurd to suppose that these effects are caused twice over. So the conscious causes must be identical to some part of those physical causes" (*ibid.*, p 17, but see Section 4.3).

Concept dualism therefore rests on two assumptions:[66]

1. Conscious mental experiences are physical events – as Papineau says of Mary when she emerges from her black and white room: "There are no new experiential properties in the offing" (*ibid.*, p 51); in Crane's words: "all mental events are physical events" (Crane, 2001, p 55).

2. Conscious mental experiences, however, can be *referred to* in two quite different ways, one using phenomenal concepts (that pick out what it is like to experience things, so phenomenal facts must exist) and the other using material concepts (that pick out the physics that takes place when we experience those things), and: "there is no explanatory link between the mental theory and the physical theory" (Crane, *ibid.*, p 55).

In concept dualism, therefore, we have two, quite different, ways of thinking about mental properties – in the one case we can view them "from the inside" (using phenomenal concepts), and in the other we can view them "from the outside" (using material concepts); and along with there being a difference between these two types of concept, it is assumed that reductive physicalism holds.

3.3.6 Eliminative materialism and behaviourism

Eliminativism and behaviourism are two related approaches to the nature of mind that developed in the middle of the twentieth century (although 'psychological' behaviourism

[66] In Loar's words (1997): "It is my view that we can have it both ways. We may take the phenomenological intuition at face value, accepting introspective concepts and their conceptual irreducibility, and at the same time take phenomenal qualities to be identical with physical-functional properties of the sort envisaged by contemporary brain science." In Papineau's words (2002, p 47): "Conscious properties are identical to material properties …", and "I think that we have two quite different ways of thinking about conscious properties … I shall call these two kinds of concepts 'phenomenal' concepts and 'material' concepts."

began much earlier). Both minimise or eliminate the significance of mental states in the scheme of things.

The aim of any approach called 'eliminativism' is to make the case that something that we think might exist, does not exist. So, the aim of eliminative materialism is to argue that something relating to consciousness does not exist. The most radical suggestion, that of eliminating consciousness itself, was touched upon briefly by C D Broad (1925, p 5) when he considered what he termed: "Behaviourism, taken quite strictly", and judged it "silly" in the sense that "only an inmate of a lunatic asylum would think of carrying it into daily life" (*ibid.*, pp 5-6). Nevertheless, Georges Rey reiterated the possibility of radical eliminativism more recently, saying: "I certainly cannot rule out the possibility that I am not conscious right here, now …" (1983, p 22, seemingly a rather strange statement; what could he have meant by 'consciousness'?). Further, both Dennett (1988, Chapter 12) and Lewis (1972) have suggested that '*qualia*' is a word without a meaning, a view apparently also held by a number of other philosophers. But as Descartes pointed out (although he didn't put it this way), it is just plain obvious, to everyone, that conscious mental states do exist.

Behaviourism comes in various flavours. 'Psychological' behaviourism is a theory that assumes that human behaviour is learned through positive and negative feedbacks. It is viewed as beginning with the work of Pavlov (see, for example, 'The Experimental Psychology and Psychopathology of Animals', for which he was nominated for a Nobel Prize in 1903). A classic paper in psychological behaviourism is that of Watson (1913), while arguably the most influential contribution to the approach is that of Skinner (see, for example, Skinner 1974).

However, according to Graham (2016) and others (*e.g.*, Smith, 1986, Putnam, 1963, p 326, or Mace, 1948), 'logical' behaviourism, also known as 'analytical', or 'philosophical' behaviourism, "traces its historical roots to" the logical positivism developed in Vienna in the early-to-mid twentieth century. Logical behaviourism is a theory about the meaning of mental terms. The philosophers in Vienna at that time were developing the philosophy of science; given the fact that science deals only with statements that can be verified experimentally they proposed that mental concepts should be viewed as picking out behavioural tendencies and that mental terms, such as 'believe', should therefore be replaced by the behaviour with which they are associated. Science depends on experiments and on the observations generated by the experiments, which are public and available to all. If mental phenomena were to be subject to a science analogous to physics, they needed

to depend on observations. Therefore, they proposed, behaviour must be what matters for a science of the mental. As Hempel said (1945): "all psychological statements which are meaningful, that is to say, which are in principle verifiable, are translatable … into statements which do not involve psychological concepts but only the concepts of physics", p 18, while Carnap (1932, p 166) maintained that: "Every psychological term is translatable into a statement about the physical state of the body of the organism" (see also William O'Donohue and Richard Kitchener, 1998). Underlying this was the idea that it is fundamental to science that people can report their experiments and the results of their experiments to others, in language, for reproducibility.

Thus, behaviourism as a theory of mind arose following the search for a scientific approach to mental states. It had the great advantage that it allowed philosophers to abstract from the difficult question of mental causation; physical inputs of various sorts led to physical outputs – behaviour.

However, whereas behaviourism is fine when it is dealing with psychology, with scientists examining the behaviour of other people, as a theory of mind it ignores the fact that we also have direct access to our own minds, and the fact that we know that what goes on in our minds is relevant to what we do, to our behaviour. Analytical behaviourism takes it that assertions concerning mental states only have meaning if they are translated into statements about behaviour or "dispositions to behave in some manner". The ontological question at issue in this thesis (whether mental states are properties of a substance different from matter, reducible to physics, or non-reducible, *etc.*) simply doesn't arise for a behaviourist. But this is not because there is no such thing as a mental state (behaviourists refer to "mental states" when they maintain that assertions relating to them can be translated into statements about dispositions to behave), it is because for analytical behaviourism, questions relating to the ontology of mental states are irrelevant.

Gilbert Ryle was an influential proponent of behaviourism. His book, *The Concept of Mind* (1949), published when interest in behaviourism was at its height, held that mental states should be viewed as being nothing over and above the person's actions, the behaviour which is initiated. Behaviourism was also taken up with enthusiasm by the American philosopher Carl Hempel (1949) and was influenced by the thinking of Wittgenstein (1922). A difficulty often cited in connection with behaviourism comes from Putnam (1963): it is the proposition that any particular mental state can be associated with different types of behaviour depending on the characteristics and the particular situation

of the individual experiencing it; it may be that a person who is experiencing pain will strive to behave as though he or she is not experiencing pain.

However, we reject behaviourism here for a different reason: it runs counter to our first fundamental proposition, Descartes' first fundamental finding in his Meditations, the proposition that the one thing that we know for sure is that consciousness, that is, awareness of feelings and thoughts, exists. If this is the case, there is more to mental states than the behaviour with which they are associated. Each of us can make direct observations regarding our own mental states, and behaviourism ignores that fact.

Other forms of eliminative materialism relating to consciousness concern language. Sellars (1956) suggests that a person's understanding of his or her own mind does not actually come from direct experience, instead it is picked up from the culture of the community he or she lives in. It certainly seems likely that some of the terms we use to describe mental states have no clear, precise referent; we have already noted that the words 'consciousness' and 'mind', although useful for communication purposes, have no precise referent. Also, different languages pick out slightly different mental concepts. Elimination of some of the terms we use to describe the so-called mental states to which we refer, is clearly a possibility.

More radically, Feyerabend (1963) suggests that the type of discourse that one might call mentalistic language could be eliminated in favour of physical language, saying: "There is … not a single reason why the attempt to give a purely physiological account of human beings should be abandoned" (1963, p 65); thus a person looking at a patch of red paint on the wall might describe the situation in terms of what he or she thought was happening in the parts of the brain related to seeing. Quine wrote: "the bodily states exist anyway; why add the others?" (1960, p 264). Rorty sees this as a possible consequence of the establishment of 'materialism', writing: "… the sort of empirical results that would show brain processes and sensations to be identical would also bring about changes in our way of speaking" (1965, pp 24-25). Paul and Patricia Churchland take this proposition further, suggesting that: "our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience" (see Paul Churchland, 1981, p 67, for example). But in order to be helpful as a means of communication, replacing mentalistic language with physical language would require that we had all already experienced every possible feeling, so we knew what it felt like when some obscure structure in the brain was

stimulated, and we would have to know exactly what happened in the brain when we had had the experience. Although logically possible, the proposition seems totally unrealistic and anyway does not get around the fact that if someone has not had some experience it is impossible to communicate to them in words exactly what it is like.

3.4 Non-reductive physicalism

3.4.1 Introduction

In 1974, in his paper 'What is it Like to be a Bat?', Thomas Nagel spelt out the reason why he sees the mind-body problem, the problem of the relationship between mind and body, as being different from problems such as that relating to water and H_2O, rejecting what he dubs "the recent wave of reductionist euphoria" (*ibid.*, p 435). He says:

> "If physicalism is to be defended, the phenomenal features must themselves be given a physical account. But when we examine their subjective character it seems that such an account is impossible. The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view" (*ibid.*, p 437, Nagel uses the word 'physicalism' in its narrow sense).

The final sentence of this quote brings to mind the "epistemic gap", and the existence of this is the nub of the case for non-reductive physicalism for those who are reluctant to accept the combination of ontological reduction and an epistemic gap.

Non-reductive physicalism generally takes it that although mental states do not reduce ontologically to the properties of matter studied in physics, they nevertheless 'supervene' on those physical properties (see Figure 1 in Section 1.1); this is known as 'supervenience physicalism' (see Section 3.4.3). The view in which mental states are instantiations of a property of matter that does not relate to physics but has a similar ontological status to the properties studied in physics, called 'property dualism' in this thesis (see Section 3.4.7), is rarely considered. In any case, non-reductive physicalism

accepts that mental states are outside physics, and therefore that the physical sciences are, in that sense, incomplete.[67]

A reason that is sometimes given for the rejection of the idea that mental states, although arising in the context of brains, could have a different nature to the things that are studied in physics is the proposition that if a mental state arises in the context of a brain state, it must surely be the case that anything the mental state can do could just as well be done by the 'physical' aspects of the brain state in question. Quine (1985, p 5) argues that: "… a dualism of mind and body is an idle redundancy." He points out that corresponding to some particular mental state will be a bodily state which might be specified as "the state of accompanying a mind that is in that mental state". Turning that around, so the mental state is the state accompanying the body that is in that bodily state then suggests that, in Quine's words: "The mind goes by the board, and will not be missed". Chalmers (1996) takes a similar view, asserting: "The physical world is more or less causally closed" (p 150), and then saying: "The very fact that experience can be coherently subtracted from any causal account implies that experience is superfluous in the *explanation* of behaviour, whether or not it has some subtle causal relevance" (1996, p 158-159, his italics). I suggest in this thesis that the ability to experience feelings enables the use of an effective mechanism for the taking of certain types of decision, and that this type of decision could *not* be taken in the same way in the absence of a gadget such as 'feelings' (see Sections 1.4.3 and 2.5.2 and Chapter 4). If the mind were to "go by the board" the ability to take decisions in this way would also "go by the board", but it would be sorely missed (see also Section 1.5, which points out the impossibility of mind "going by the board" in the manner suggested if mental states are instantiations of a property of matter).

3.4.2 Lewis' case against the existence of phenomenal facts

In his interesting 1988 paper, 'What Experience Teaches', David Lewis sets out a number of arguments designed to suggest that we should reject the idea that there is such a thing as phenomenal information, information that is not reducible to the type of information studied in the physical sciences (the distinction between the two being, of course, that the

[67] As we have seen (Section 2.3), the 'completeness of physics' is generally defined in philosophy in a rather limited way as: "all physical effects are fully caused by purely *physical* prior histories" (where 'physical' means "pertaining to the physical sciences" on both occasions); this definition of 'completeness' does not prevent a physical situation causing the existence of something, *eg.*, a mental state, that is outside physics.

type of information studied in the physical sciences can be communicated perfectly between people linguistically while in order to know a phenomenal fact, what it is like to experience something, it is necessary actually to have the experience). In this section I examine his arguments to see if any of them can make a convincing case against our thesis that mind could be a non-reducible property of matter, which of course would carry with it the implication that phenomenal facts do exist.

Is experience the best teacher?

Lewis starts by trying, ever so gently, to shake one's faith in the idea that phenomenal information can only be learned through experience. First, he suggests that maybe either neurosurgery or magic might enable the learning of what an experience is like (with neurosurgery viewed as a possibility, but raising a question about the interpretation of what exactly is learned, while the proposition that magic could enable this learning to take place simply ignores the existence of the laws of nature). Or, he says, possibly some new type of science lessons or science learning? After which Lewis concludes (*ibid*., p 79): "It's not an absolutely necessary truth that experience is the best teacher about what a new experience is like. It's a contingent truth … we just might be in for surprises".

Here we take the view that Lewis is wrong on this: that we know *from introspection* that phenomenal information can only be learned from experience (see Section 2.4.2 for the case in support of the existence of phenomenal information).

Lewis' hypothesis of phenomenal information

Lewis then spells out what he calls "the hypothesis of phenomenal information" (*ibid*., p 84) as the hypothesis that: "… besides physical information there is an irreducibly different kind of information to be had: *phenomenal information*." He says:

> "The two are independent. Two possible cases might be exactly alike physically [materially], yet differ phenomenally. When we get physical information we narrow down the physical possibilities, and perhaps we narrow them down all the way to one, but we leave open a range of phenomenal possibilities. When we have an experience, on the other hand, we acquire phenomenal information; possibilities previously open are eliminated; and that is what it is to learn what the experience is like" (my insertion).

But in the course of spelling out this hypothesis Lewis has restricted the meaning of the term phenomenal information to cover *only* such information as is independent of all so-called "physical information". For information to rank as phenomenal information in Lewis' sense it has to be independent of physical information: it has to be possible for two "cases" to be exactly alike physically while differing phenomenally. But if mind is a property of matter, if two cases are exactly alike regarding the *matter* involved, and therefore regarding their physical (non-mental) properties, they will also be exactly alike in terms of their phenomenal properties. When we narrow down the physical possibilities, the possibilities regarding the physical properties, to one we also narrow down the phenomenal possibilities to one (except in the exceptional case in which a change in the matter produces a change in the mental state but no change in the physical properties). It is only in the case of *ontological dualism* (see Section 3.2) that phenomenal information and physical (non-mental) information would be independent in the manner suggested here by Lewis.

By defining phenomenal information in the way he does, Lewis rules out the possibility that that there might be phenomenal information associated with a property of matter not suitable for study in physics; and he gives no justification for doing this.

Lewis then demonstrates that if phenomenal information of the type implicit in what he calls the hypothesis of phenomenal information exists in the context of Jackson's knowledge argument, materialism cannot hold. If 'materialism' holds, he says, any two possibilities that are exactly alike physically (alike in terms of the matter involved) will be alike in every way; and this is true. So, he says, if there were another kind of information, different from the type of information that arises in physics, which picked out one of the two possibilities but not the other, 'materialism' (as he defines it) would not hold; and this is also true. But this does not rule out the possibility that there is another type of information, different from the type of information that arises in physics because it is what we know as 'phenomenal' information, and it arises in *each* of his two possibilities because there is another property of matter that is not studied in the science of physics and this type of information relates to this other property.

Throughout this paper Lewis never explicitly considers the possibility that phenomenal information could be information about an instantiation of a property of matter that is different from those that are studied in physics apart, perhaps, from when, in the middle of his presentation of this particular argument, he refers to phenomenal information as "… something … which is alien to this world – an alien kind of thing, or maybe an alien fundamental property of nonalien things" (*ibid.*, p 89), and the words "an

alien fundamental property of nonalien things" could clearly be used to describe mind being a property of matter.

Betting against the truth of physics

Next, Lewis suggests that phenomenal information is a "funny sort of information". With physical (non-mental) information, when you don't know some piece of information you typically have some idea of what the relevant alternative possibilities are. With phenomenal information, if you don't know what it's like to have an experience, you also have no idea of the alternative possibilities (*ibid.*, p 94). Which is true, up to a point (think magneto-reception), but is not relevant to the possibility of the existence of such information.

Then, more seriously, he makes his case against the possibility that physics as we know it could be at fault, saying (*ibid.*, p 95):

> "if something non-physical sometimes makes a difference to the motions of physical particles, then physics as we know it is wrong. Not just silent, not just incomplete – wrong. Either the particles are caused to change their motion without benefit of any force, or else there is some extra force that works very differently from the usual four. To believe in the phenomenal aspect of the world, but deny that it is epiphenomenal is to bet against the truth of physics. Given the success of physics hitherto, and even with due allowance for the foundational ailments of quantum mechanics, such betting is rash!"

Lewis is quite correct: this would be betting against the truth of physics as we know it today. But I think he exaggerates the odds against such a bet turning out to be correct. In this thesis I suggest that there could be another force relevant to science, as well as the 'usual four', see Section 2.5, but if there were, there is no reason why it should have to work "very differently from the usual four". Such a force would operate only on a very small scale and only inside the brains of certain higher order creatures and it seems highly unlikely that scientists would have tripped over it at this point in the development of our understanding if they were not actually looking for it. It is unlikely that it would have much effect on any of the macro-truths that are currently believed as a result of physics as we know it today (although it would have had a profound effect on the events in the world). Perhaps more likely, Lewis' "something non-physical" could have its effect on the motions of physical particles through one of the already known forces.

The idea that a mental state could affect the motion of a particle is potentially an exciting development for science as well as being a bet against the truth of physics as it is

understood today; both physics and neuroscience would be fundamentally changed. But, after all, science is developing all the time.

The ability hypothesis

David Lewis then suggests that the question regarding the existence of phenomenal information, illustrated so vividly by the Knowledge Argument (see Section 2.6.1), can be explained by what he calls the "ability hypothesis". In a nutshell, the ability hypothesis says that, in spite of the fact that Mary knows everything there is to know about seeing red from studying the relevant physics and neuroscience *etc.* in her black and white room, it is only when she actually sees the colour red that she *acquires the skill* of being able to visualise red (in Lewis' words, the skill of being able to: "remember, imagine and recognise" red). This hypothesis was originally proposed by Nemirow (1990) who said: "some modes of understanding consist, not in the grasping of facts, but in the acquisition of abilities…. I understand the experience of seeing red if I can at will visualize red" (p 475).

The proposal that Mary learns a new ability concerns the way in which a person can "know how" to do something. We actually know quite a bit about "knowing how" (see Section 1.3.3, where I also make the point that different types of knowing can be expressed in many different ways). We know that learning how to perform a skill, such as riding a bicycle or playing a piano sonata involves neuronal changes that take place in the cerebellum and we know that it involves encouraging the growth of new neuronal connections through repeating actions; it involves practice (see, for example, Kandel, 2006, p 202). Lewis was aware of this, so he said: "Information very often contributes to know-how, but often it doesn't contribute enough. That's why music students have to practice" (*ibid.*, p 100). But learning what it is like to see red is not like that. Once a person has seen red, they know what it is like; no practice is needed.

Finally, the way Lewis describes his ability hypothesis it reads very much as if it is a description of "what it feels like" under another name, a different terminology. He explains that: "… knowing what an experience is like just *is* the possession of these abilities to remember, imagine and recognise" (*ibid.*, p 100). But surely, knowing what an experience is like just *is* knowing a phenomenal fact; and it is only once the phenomenal fact is known by the person, that the person can remember, imagine and recognise it. Lewis says: "These abilities to remember and imagine and recognise are abilities you cannot gain (except by super-surgery, or by magic) except by tasting Vegemite and learning what

it's like" (*ibid.*, p 99). Quite so. If there actually *is* such a thing as phenomenal information; we don't need Lewis' ability hypothesis to explain Jackson's knowledge argument.

Conclusion on Lewis' argument against phenomenal information

Lewis says that "for materialists it is essential to reject it [his hypothesis of phenomenal information]" (*ibid.*, p 97, my insertion), and that is correct. For if his hypothesis of phenomenal information were true of the world, given the way in which it is worded, materialism (in the sense that everything pertains to matter) would not hold. But this does not mean that we should reject the proposition that phenomenal facts exist (using the term 'phenomenal' as it is usually understood, as opposed to Lewis' definition of it), or that we should reject the proposition that these phenomenal facts either supervene on physical facts or relate to instantiations of a property of matter different from those studied in physics. After all, we know that mental states exist (if we know anything we know that, see our first fundamental proposition), and we know that what our mental states are like to experience is impossible to communicate perfectly to another person in language (see our second fundamental proposition); and if that is the case we know that phenomenal information, what some experience feels like, exists and is different from public information.

Of the four points Lewis produces that support the proposition that there is no such thing as phenomenal information:

- the idea that some new type of "science lessons" could enable the accurate communication of phenomenal information runs counter to our introspective knowledge that in order to know what it feels like to experience something we need to experience it;

- the 'phenomenal information' of Lewis' "hypothesis" is not the type of phenomenal information we are concerned with in this thesis; he stipulates with his meaning of the term that it only applies to the situation in which two cases can be exactly alike physically (materially) and yet differ phenomenally, and that restriction does not apply to the type of phenomenal information we are concerned with here;

- the proposition that admitting the existence of phenomenal information raises issues relating to physics as we currently view it is absolutely correct, but contrary to the view of Lewis, the problems that might arise seem unlikely to cause any great upset to physics as it is currently understood; and,

- Lewis' "ability hypothesis" provides a possible, but implausible, explanation of how the situation described in the knowledge argument might arise *if there were no such thing as phenomenal information*. It is implausible because it does not comply with what is known about the process of acquiring skills (the necessity for repetition, so new neurons develop). So it seems more likely that if there were no such thing as phenomenal information the 'Mary' situation would simply not arise.

3.4.3 Supervenience physicalism

'Supervenience physicalism', originally proposed by Donald Davidson in his 1970 paper), is the type of non-reductive physicalism most often cited in the recent philosophy literature. The formal definition of supervenience has it that a set of properties 'A' is said to 'supervene' upon another set of properties 'B' if and only if no two things can differ with respect to their A-properties without also differing with respect to their B-properties. In David Lewis' words, "We have supervenience when there could be no difference of one sort without differences of another sort" (1986, p 14). In the case of supervenience physicalism, an A-property would be some type of mental state (an instantiation of the property of matter we have called 'mind') and the B-properties would be the physical (non-mental) properties associated with the collection of neural states that arise in the context of that mental state; supervenience physicalism says that there cannot be a difference in the mental state without there also being a difference in the associated physical (neural) properties of the brain.

When philosophers identify some particular situation as involving the supervenience relation they rarely seek to identify the reason why the supervenience relation should hold in those particular circumstances, as one would imagine a scientist might. So there is a question as to why a supervenience relation between mental properties and physical properties might arise. If the supervenience relation is deemed 'necessary', as is generally the case, it could hold as the result of logic (as in "economics supervenes on the behaviour of people regarding the production, distribution and consumption of goods and services", which is true because of what we mean by the term 'economics'). If it is not necessary as the result of logic, it could be necessary because of the existence of a law of nature which ensures that some mental property is instantiated whenever a particular configuration of the properties suitable for study in physics arises (see Section 1.2.3): that is, there could be a law of nature that says that the B-properties instantiate the A-properties,

as in Figure 1a (in Section 1.1, see the arrows); then the mental properties would be instantiated *because* the physical properties are instantiated, and obviously, every time the particular configuration of physical properties occurs the mental property will too. If that were, indeed, the case, assuming that supervenience physicalism holds would simply involve assuming that mental states are properties of the physical properties of matter (rather than being properties of matter in their own right, see Figure 1b in Section 1.1) because a law of nature requires that that is so.

But supervenience as defined above can also arise because of features that just happen to be the case regarding matter, instead of as the result of a necessary relation between mental properties and physical properties. This could happen in the case of property dualism, see Figure 1b, where mental properties are a property of matter and have a similar ontological status to the properties studied in physics. In that case, a particular mental characteristic is instantiated every time there is a particular configuration of *matter* (the type of matter and/or its arrangement). But if it so happened that matter was such that it was not possible to change its configuration in such a way as to change the mental characteristics without *also changing the physical properties*, then it would be the case that there could be no change in the mental characteristics without a change in the physical properties. But in this case there would be no law of nature *directly* connecting the mental characteristics with the physical properties of matter; the supervenience would appear to have an element of happenstance.

I take the term 'supervenience physicalism' to refer to the thesis that mental properties are determined by the physical properties of matter as the result of a law of nature, as in Figure 1a; one might call this 'nomological supervenience'. This seems to be the case in many philosophers' specifications of supervenience physicalism, see Kim (2005, p 33), Papineau (who finds: "the notion of supervenience more trouble than it's worth", 2002, p 37), Wilson (2005, p 433), Gardner (2005, p 191), and many more.[68]

If materialism is true, so there is only one substance and it is 'matter', and if mental states are instantiations of something that derives from matter in some way (as do physical, non-mental states), then whether consciousness is reducible or non-reducible, emergent,

[68] If it was possible to change the constitution of matter in such a way as to change the mental characteristics but not the physical (non-mental) properties then, so long as we interpret the word 'physicalism' in a broad sense, the proposition that mental states are instantiations of a property of matter different from those studied in physics but of a similar ontological status could be considered to constitute a counter-example to Kim's important statement that: "mind-body supervenience can usefully be thought of as defining *minimal physicalism* …" (2005, p 13).

pan-psychic, micro-psychic or identical to non-mental properties, it would certainly be the case that there cannot be a change in the mental state without there also being a change regarding the matter itself; although, of course, there might be no change to the physical properties.

Before leaving the subject of supervenience we should note that the term 'supervenience' has another meaning in philosophy besides the value-neutral one described above (see Horgan, 1993). The notion of supervenience nowadays often carries with it the idea that the thing that does the supervening is in some sense a 'higher level phenomenon' than the thing on which it supervenes. And, further, the supervenience of mental states on neural states is often taken to imply that consciousness 'derives from' or 'is grounded by' the non-mental properties of neural material in some special, mysterious way rather than that it is simply another property of matter, as some non-mental property of matter such as electric charge might be viewed as being a property of matter. I can see no reason why there should be value judgments relating to 'higher' and 'lower', or to 'first level' or 'second level' associated with the supervenience relation. The first meaning given here of the term supervenience ("no difference in A without a difference in B") clearly does not necessarily imply the other – the "hierarchical values", or "mysterious grounding" relation – although there are many situations in which both might hold.

3.4.4 Physicalism, or something near enough

Jaegwon Kim finds, in his book *Physicalism, or Something Near Enough* (2005, p 1), that: "… although we cannot have physicalism *tout court*, we can something nearly as good".

Kim sets down what he identifies as the two big problems of theory of mind: the first being: "how mentality can have a causal role in a world that is fundamentally physical", and the second being: "can we give a reductive physicalist account of consciousness?" (*ibid.*, p 1). He sees the first as requiring that causative mental states are reducible to physics because, like so many others, he signs up to the philosophers' specification of the causal completeness of physics. However, he views some types of mental state as being "functionalizable" (*ibid.*, p 165), in the sense that they can be: "defined or characterized in terms of their causal work". So he takes these types of mental state to be reducible. In his view there are insurmountable problems with causation operating from one substance to another (see Section 3.2.3), so he rules out the possibility of causally effective mental states that are instantiations of some property of another substance ('immaterial minds'). On the grounds that he has accepted both the causal completeness of physics and the exclusion

argument, the possibility of there being another property of matter, one that is not suitable for study in physics but that *can* interact causally with the properties of matter studied in physics never arises.[69]

On the second big problem Kim finds that there are reasons why we should reject the proposition that all mental states can be wholly accounted for in terms of physics. He says: "As far as we now know, the only way to create a system with conscious experience is to duplicate an appropriate animal or human brain" (*ibid.*, p 169). Where a mental state is not functional, as with sensory states (or *qualia*), Kim therefore accepts that these could well be ontologically irreducible.

His conclusion, after a great deal of careful analysis, is that: "intentional/cognitive properties are reducible, but qualitative properties of consciousness, or '*qualia*', are not" (*ibid.*, p 174). But if this is the case '*qualia*', which would be instantiations of something outside physics, cannot have any effect on the intentional/cognitive properties because those are reducible to physics, and they would not be reducible to physics if the phenomenal information of the *qualia* influenced them. So *qualia* are epiphenomenal.

Kim's proposal is problematic if we accept that we know, *from introspection*, that it is *qualia* that are the key to the taking of those complex decisions in which the effects of the alternative choices are not directly comparable, see Section 1.4.3. So the idea that the "qualitative properties of consciousness", the *qualia*, are epiphenomenal, with its implication that feelings must therefore have no effect whatsoever on a person's actions, is not a proposition that is acceptable in the light of our third fundamental proposition (it would mean, for example, that it is not because you don't like the taste of spinach that you choose not to eat it; instead, it is the result of something to do with your physical, that is non-mental, attributes – a situation somewhat reminiscent of Leibniz, 1695, see Section 3.2.3).

[69] The exclusion argument says that where a mental property that appears to be causally effective supervenes on some particular set of physical properties, it will be the physical properties that cause the behavioural effects; there will be no causal work left for the mental state to perform.

3.4.5 The rejection of 'over-determination'[70]

The issue

We have seen in Section 3.3 that it is often suggested that the only way to reconcile mental causation with the causal completeness of physics (as defined in 2. below) is to be a 'type identity theorist', that is: "to identify mental properties with physical properties" (Crane, 1995, p 230). Whereas we may come across some situations that appear to involve psycho-physical causal processes (mental states *per se* causing physical events), a type identity theorist would reckon that it must be the case that this is either because the mental properties are identical to physical properties, or it is because the mental state is supervenient on a physical state and it is that physical state that does the causing; as Kim says: "at some point, purely physical causal processes take over" (Kim, 1984, p 264).

Many have spelt out the assumptions underlying the problem of mental causation in broadly the same way, see in particular Crane (1995, p 229), Lewis (1966), Papineau (2002, pp 17-18) and Kim (2005, pp 21-22):

1. There is mental causation

2. All physical effects are fully caused by purely *physical* prior histories

In Lewis' case the first premise is "… the definitive characteristic of any experience as such is its causal role" (*ibid.*, p 19) and the second "all phenomena [are] describable in physical terms" (*ibid.*, p 23). From these he goes straight to the conclusion that: "from the two premises it follows that experiences are some physical phenomena or other" (*ibid.*, p 24). Crane, Papineau and Kim add a third assumption:

3. There is no over-determination

In this section I examine this third assumption. Obviously, its implications depend on what you mean by 'over-determination'. If the meaning of over-determination, on the face of it a mildly pejorative term, relates only to 'coincidence' (as in simultaneous bullet killings), then the proposition that there is no over-determination on a regular basis is clearly acceptable. But if the term over-determination means 'dual causation' we might

[70] Here I ignore the proposition that feelings are relevant to the outcome of certain decisions.

not wish to rule it out without further consideration. In this section I take a look at some of the possible implications of this thought.

Dual causation

The first two assumptions above can, of course, be reconciled without signing up to the identity theory; logic allows another alternative. There could be a strict one-to-one relation of nomological necessity between some of the neuroscientific aspects of a brain state and the mental state which it instantiates, a relation that does not require mental characteristics to be identical with neural characteristics but which does mean that you can't have one of these two things, the mental state and the brain state, without the other also obtaining.

If this were the case, an event in a creature's brain that instantiates both a particular mental occurrence and the instantiation of a particular physical (neurological) state, and which causes a physical event (an action) to take place, may cause it as a result of the conscious mental characteristic, operating through a psycho-physical law as in point number 1 above, or through the laws of the physical sciences by the physical prior history, as in point number 2 above, see Figure 5.

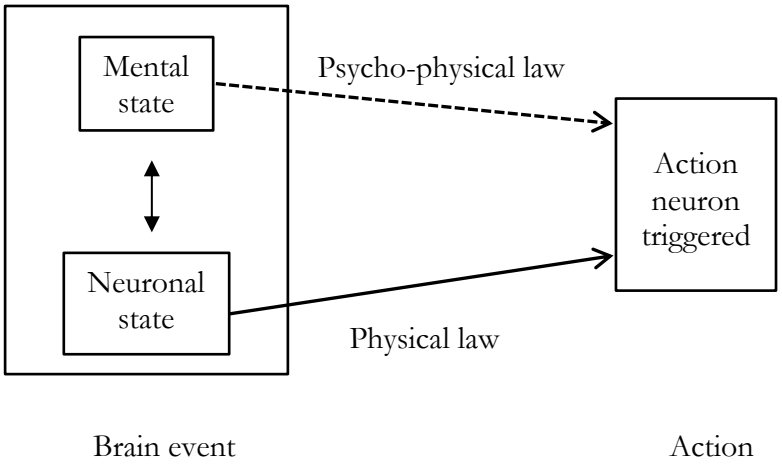


Figure 5: An event with two causes.

In Figure 5 the long arrows represent laws of nature describing a causal process as understood by us humans. The upper, dashed, arrow represents a psycho-physical law. Psycho-physical laws deal in both psychological concepts (such as "feeling hungry") and physical concepts (such as "opening the refrigerator" or the movement of an electron in a

neuron). The lower arrow, undashed, represents a physical law; physical laws deal solely with physical (non-mental) concepts. If the action is caused both by the mental characteristic through a psycho-physical law and by the physical characteristic through a physical law, then the two laws, as well as the two sets of properties of matter, must be inextricably bound together; you can't have one without the other. The action in Figure 5 is caused by the brain event, and the brain event has two different types of property associated with it.

So does the existence of these two 'causes' mean that there is an unacceptable type of over-determination? The answer is: not necessarily. This is because one of these mental/neural characteristics cannot occur without the other also occurring; there would have to be a law of nature that ensures that whenever a mental characteristic acts as a cause of some physical event, operating through a psycho-physical law, a neuronal characteristic operating through the laws of physics also acts as a cause of that event (and *vice versa*). There could be one causal event which has two aspects to it; it is that causal event which we observe resulting in the event that constitutes the effect, the action, but we humans might regard each of these aspects of the causal process as being individual laws of nature (as Davidson says, laws are linguistic).

If this were the case, there would be no reason why we should not use the laws relating to mental events when we identify the thing that causes the action (as in "Joe went to get a glass of water because he was thirsty") rather than the laws relating to physical brain events (involving a multitude of highly complex biological and chemical reactions, electric currents *etc.*). Indeed, to do otherwise would seem to be like trying to "explain" the sequence of pictures on a television screen by specifying the patterns of dots, rather than by citing the story-line depicted (see also Smart, 1959, discussed in Section 3.3.3). In any case, the question of the mechanism through which mental states could have direct physical effects means that the physical has tended to win out in the causality stakes. For example, Honderich, examining Davidson's anomalous monism, (1982, p 64) remarks that it is: "a mental event as physical which causes an action".

However, people have generally tended to view physics as consisting of "strict deterministic laws" (prior to quantum mechanics) while some have doubts as to whether there are any sound psycho-physical laws at all – that is, laws that relate mental states to action (*e.g.*, Davidson, 1970, p 208). If there were a fundamental difference of this sort between the laws of physics and the laws of psycho-physics, we would indeed be able to discover which was the law that caused the action.

Laws of physics and psycho-physics

An obvious problem with the proposition that both the mental and the physical characteristics of some mental event could equally be viewed as being the cause of a subsequent event, is that laws with physical causes are often thought of as being deterministic (100 per cent certain in their outcome) or near-deterministic, while laws with mental causes often appear to be highly stochastic; the predictions associated with them seem to have a large element of uncertainty attached to them (see, for example, Davidson, 1970, who reckons that: "there are no strict deterministic laws on the basis of which mental events can be predicted and explained", p 208)

Where physical laws are concerned, it used to be taken for granted by many that they had no random component attached to them, and that the outcome of a causal process would therefore be determined with 100 per cent certainty. During the last century, however, with the development of quantum mechanics, it seems that there could be a stochastic component in the laws of physics, depending largely on the magnitudes of the objects under consideration. In most cases any stochastic component a process might possess is likely to be an extremely small proportion of the total signal – this is certainly the case for physical macro-events such as those involving billiard balls and the like; but there are also cases for which the random element associated with quantum mechanics could be significant. Whether this might be so for neuronal events depends on the nature and magnitudes of the structures involved.[71] We therefore cannot totally rule out the possibility that quantum uncertainties could inject a degree of randomness into the micro-behaviour of neurons in some circumstances, thereby having a potentially significant effect

[71] The famous two-slit experiment demonstrates that if a particle, such as an electron, is fired towards a screen that has two slits in it, with a detection screen beyond, it is uncertain where the particle will hit the detection screen. The probability of it hitting the detection screen at any particular point does not just depend on which slit it "goes through", as common sense would suggest, it also depends on the relation between the positions of the two slits (Jönsson, 1974). It is as if the one particle somehow goes through both slits. This can be the case with quite sizeable particles. Recently, a group based in Vienna fired a "buckyball" molecule (a molecule with the formula C_{60} which consists of 60 carbon atoms in a cage-like structure), which has a diameter of around 1 nanometer (10^{-9} of a meter), through the double slit apparatus and demonstrated that these particles have no problem in "sailing through both slits simultaneously" (Arndt *et al.*, 1999). The buckyball is similar in diameter to the width of the DNA double helix. A typical axon is around a micrometer (10^{-6} of a meter) in diameter, with some much larger and others much smaller. This suggests that quantum effects in brains is a possibility that cannot yet be ruled out (see also Section 2.5.2).

on the macro-behaviour of the organism (see Hameroff & Penrose, 1996, who propose that this might be the case, and Tegmark *et al.* 2000, who maintain that according to their calculations: "there is nothing fundamentally quantum-mechanical about cognitive processes in the brain", p 12).

Where mental laws are concerned, as with the physical laws underlying a person's behaviour, if one were able to know the full gamut of mental states to which another person is subject, along with all the psychophysical causal processes and the magnitude of any inescapable uncertainty associated with them, we might find that the stochastic element was extremely small, or there might not be any uncertainty at all. In that case one would be able to predict with reasonable or total accuracy the way in which a person would behave in any circumstances. However, quite apart from the current state of psycho-physical science, we can never know exactly what it is like to be another person, so the uncertainty of us humans with regard to psycho-physical causal processes is unlikely ever to be reduced to a level comparable to that of the laws of physics.

It seems that we cannot rule out the possibility of dual causation on the basis that the perceived uncertainly of predictions relating to mental processes is greater than that relating to physical processes.

Conclusion regarding 'over-determination'

On the grounds considered above, there seems to be no reason why we should reject over-determination in the sense of dual causation regarding mental and physical properties. If it is the case that whenever a mental state arises a particular physical state arises also, there is no reason why we humans should not postulate a law of nature corresponding to each of these two aspects of an event. And on the face of it, it could be that we would then be unable to identify either one of these laws of nature as describing the *actual* causal process involved.

However, so long as we sign up to Proposition 3 in Chapter 1, this situation does *not* arise. Given that we know, *from introspection*, that it is the mental state that determines the outcome when a complex decision between alternative actions is being taken; the proposition that the physical aspects of the mental state can "make the judgment" regarding which option will be best for the creature concerned, while not ruled out by logic, just seems too implausibly farfetched on practical grounds.

3.4.6 Naturalistic dualism

As we have seen (Section 2.2.4), an approach that is somewhat similar to the one I support in this thesis is that of naturalistic dualism, put forward by Chalmers and presented in his book "The Conscious Mind" (1996).

Chalmers starts his discussion of naturalistic dualism by pointing out that one cannot, on the basis of logic alone, rule out the possibility of zombies – creatures that are physically identical to conscious creatures in every way but have no consciousness (see Section 2.6.2). This means that, indeed it is simply another way of saying that, consciousness cannot be *logically* supervenient on the physical; for logical supervenience would imply that if two creatures had exactly the same physical properties then, as a matter of logical necessity, they would also have exactly the same properties in regard to consciousness. He then concludes that: "the failure of logical supervenience directly implies that materialism is false: there are features of the world over and above the physical features" (*ibid.*, p 123). He spells this out thus (p 123):

1. In our world there are conscious experiences.

2. There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold.

3. Therefore, facts about consciousness are further facts about our world, over and above the physical facts.

4. So materialism is false.

But as we have seen (Section 2.6.2), if conscious experiences are instantiated by matter, not as the result of logic, but as the result of a law of nature, then materialism in the sense that there is only one substance, matter, would *not* be false. So Chalmers must be using the term 'materialism' in the sense of "everything relates to the physical sciences" that is, mental properties are identical to, or ontologically reducible to, physical (non-mental) properties, when he makes his claim, although he does not make that clear.

In fact, Chalmers subsequently (p 124) endorses the idea that, given the laws of physics and biology that operate in our world, consciousness supervenes on the physical in our world and it is this proposition that he names 'natural supervenience' ('supervenience physicalism' is discussed in more depth in Section 3.4.3). He says, however: "conscious experience involves properties of an individual that are not entailed

by the physical [non-mental] properties of that individual, although they may depend lawfully on those properties" (p 125, my insertion). If mental properties, and therefore consciousness, arise whenever matter is organised in certain ways in the brain then surely they could be viewed as being instantiations of 'just' another, different, property of matter. So his argument only implies that materialism is false if 'materialism' is interpreted as: "everything can be explained by the science of physics"; it does not imply that materialism in the sense: "everything relates to one substance, matter" is false.

Chalmers takes it that, given that we know we *do* have conscious experiences, there must be something other than that which is studied in physics: "The physical facts incompletely constrain the way the world is; the facts about consciousness constrain it further", he says (1996, p 123; but see Lewis' "hypothesis of phenomenal information" and the discussion in Section 3.4.2). He also finds it plausible that if, for example, his own physical [material?] structure were precisely replicated, his conscious experience would be replicated also. And indeed, this would be the case if mental states were instantiations of a property of matter that is different from those studied in physics, as we assume in this thesis. However, Chalmers says, on page 124: "… it remains plausible that consciousness supervenes on the physical. It is this view – natural supervenience without logical supervenience – that I will develop". The difference between 'supervenience physicalism' and the type of 'property dualism' I am exploring in this thesis is illustrated in Figure 1 (see Section 1.1).

Chalmers believes that the approach he calls alternatively 'physicalism' or 'materialism' (for him, both of these mean the same thing, that everything pertains to physics) is false. The reason for this is his belief that: "the existence of further contingent facts over and above the physical facts is a significant enough modification to the received materialist world view to deserve a different label" (*ibid*., p 126). But then, he always uses the term 'physical' to mean "pertaining to the physical sciences". His approach is in fact consistent with something that one *could* call 'physicalism in the broad sense' or 'materialism' because there is nothing in it that requires a second substance; there is nothing that prevents it being the case that everything pertains to matter.

The reason why this difference between Chalmers' naturalistic dualism and the approach I am exploring here matters is because Chalmers' approach, along with his assumption that mind supervenes 'naturally' (not logically) on the physical, lends itself to the assumption that: "the physical domain is causally closed" (1996, see p 161); assuming that mind is a property of matter different from those studied in physics but of similar

ontological status, lends itself naturally to a causal role for mental states (all the other properties of matter are causally effective so why not this one?). Chalmers therefore takes it that epiphenomenalism pertains: that there is no causal role whatsoever for mental states regarding the physical world. He does consider the possibility that experience is: "a fundamental feature of the world, alongside space-time, spin, charge and the like" (p 126), and he says: "Where we have new fundamental properties, we also have new fundamental laws. Here the laws will be psychophysical laws, specifying how phenomenal (or protophenomenal) properties depend on physical properties" (p 127). But then he adds: "These laws will not interfere with physical laws; physical laws already form a closed system. Instead, they will be supervenience laws, telling us how experience arises from physical processes."

3.4.7 Property dualism

As we have seen, many modern philosophers believe that there is only one substance (matter), denying the possibility that there might be immaterial minds that animate material bodies. But while many nowadays identify mental states with physical (that is, 'non-mental') states, many also acknowledge the possibility that there could be two different classes of property, 'physical' properties and 'mental' properties, with mental properties fundamentally different in kind from the 'physical' properties; that is, with the two belonging to different categories. This position is the one I call 'property dualism' in this thesis.[72]

Property dualism thus assumes the existence of a single substance but takes it that this single substance could have two, very different, types of property. The idea is that our bodies have both physical properties such as extension and weight, and mental properties such as thoughts and feelings. It is not just that we might talk of mental states and physical states in different ways; the idea is that there is an ontological difference. And it is not just

[72] McGinn describes an approach somewhat similar to this in his 1982 book (*The Character of Mind*, Chapter 2), but he doesn't address the question of effective mental causation and he later suggests that: "the mind-body problem brings us bang up against the limits of our capacity to understand the world" (1989, p 354). Strawson (eg., 1994, 2008) is perhaps the only other modern philosopher who views "experiential phenomena" as relating to matter in a similar way to that of non-mental phenomena, but he does not explore the possibility of mental causation and its implications. Dubbed "interactionism" (a term which usually describes the situation in which mind is a property of a different substance from matter and causal effects operate between mind and matter) this approach is included in Chalmers' (2002) classification of views on the metaphysics of consciousness under the heading "Type D Dualism".

that mental properties are a kind of strange spin-off from the physical properties, so they make no difference whatsoever to anything physical that happens, now or in the future, because they are epiphenomenal. According to property dualism, mental states cannot be ontologically reduced to the properties of matter studied in the physical sciences, and nor do they arise only as a consequence of the existence of those properties; mental properties have the same ontological status as non-mental properties.

However, although that possibility is allowed by modern philosophy, few modern philosophers actually consider this option sufficiently plausible for it to be worth setting pen to paper and writing an article or a book exploring the implications of those assumptions (see Section 2.2). An exception is Galen Strawson, who has written much on the subject and has been an exponent of this approach for many years. Galen Strawson (2008, p 54) takes the view that what he calls 'real physicalism': "… can have nothing to do with *physicSalism*, the view – the faith – that the nature or essence of all concrete reality can in principle be fully captured in the terms of *physics*." He starts from what he views as being a fundamental fact: the fact that the one thing the existence of which is more certain, more real to us than that of everything else, conscious experience, is 'physical'. The result is that it seems very odd to him that anyone might find it strange that something that is 'physical' is capable of 'experiencing' things. He refers to Eddington who wrote, in his book on the nature of the physical world: "But what knowledge have we of the nature of atoms which renders it at all incongruous that they should constitute a thinking object?" (1928, p 259).

However, although Strawson signs up to the basic ontology of property dualism, to the best of my knowledge he does not address the issue of mental causation in that context.

3.4.8 Functionalism

Functionalism as a theory of consciousness

Functionalism starts from the idea that what matters about the mind is what it makes happen, its function. It is often specified in terms of input-output tables (see Block, 1978). Thus, given the state of some organism (or 'thing') along with an input, functionalism decrees that there will be a specified output and a new state for the organism (or 'thing'). The difference between functionalism and behaviourism is that on the behaviourist view stimulus inputs lead to behavioural outputs whereas, on the functionalist view stimulus *plus*

mental state inputs lead to behavioural *plus* mental state outputs. This enables functionalism to deal with counterexamples to behaviourism in which a person burns his or her finger but does not exhibit classic pain behaviour because of a mental state that suppresses it.

Accepting the proposition that what matters about mind is what it makes happen, the functionalists take the view that it is sensible to *define* 'mental state' in terms of the function that is involved. This means that two mental states are deemed to be the same, in functionalism, if they play the same role (but note that one *experience* of a particular mental state, so defined, may then feel very different from another *experience* of the same so-called mental state). But once that definition has been accepted, the function played by the so-called 'mental state' in making some particular event happen could just as well be played by any gadget other than a mind so long as that gadget was wired up in such a way as to perform the same function; so computers can have these things called 'mental states' as can thermostats, indeed, a particular 'mental state' could be "made of" anything (the population of China, see Block, 1978; beer cans and ping-pong balls, see Chalmers, 2006 p 249), just so long as the "fine-grained functional organisation of the parts" is the same. Then, since we normally think of what we call "mental states" as being a thing that sometimes has consciousness associated with it, a question arises as to whether the so-called gadget, such as the population of China, would instantiate consciousness (see Dennett, 1991, Chapter 14); but of course in functionalism, the definition of 'mental state' is significantly different from the usual one. Clearly, the 'mental states' of functionalism can be multiply realised in terms of the matter involved (see Putnam, 1967) and that is not necessarily the case for mental states on the usual definition of "what it feels like", see Section 4.5.

Functionalism has its so-called 'mental states' dependent on matter and by definition the things it calls 'mental states' can be causally effective. To that extent I can pick no argument with it in the context of the case I am making for the ontology of mind in this thesis. Beyond that, to propose that all that matters about the mind is what it makes happen runs counter to personal experience. Also, and more seriously, to name what might be called a 'functional state' relating to a creature its 'mental state' is seriously at variance with the normal meaning of mental state; and given that people then tend to assume that the normal meaning of the term *also* applies, it results in people proposing fantastical notions, such as the possibility that the population of China, taken as a whole, could experience consciousness!

3.4.9 Multiple realisation

Putnam's influential paper 'The Nature of Mental States' (1967, reprinted in Rosenthal, 1991) is supportive of functionalism in that it supports the view that mental states can be multiply realised. In this paper Putnam addresses the question "Is pain a brain state?" arguing *against* a positive answer in the broad context of this approach known as 'functionalism'. In this thesis I am putting the case for mental states such as pain being properties of the matter in the brain, so in this section I examine Putnam's case against that proposition.

Putnam argues that: "pain is not a brain state, in the sense of a physical-chemical state of the brain (or even the whole nervous system), but another *kind* of state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism" (p 199 as reprinted in Rosenthal, 1991). He then lays down a challenge for a hypothetical "brain-state theorist" who maintains that "pain is a brain state". Putnam says that to "make good his claims" the brain-state theorist must "specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state" (ibid., p 200). Putnam points out that "the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are molluscs, and certainly feel pain), etc." (*ibid.*, pp 200-201).

Before proceeding, we should note that the fact that a wide range of physical states exists to fulfil any particular *function* is a common-place. In biological terms this is something that occurs right across the animal kingdom, and is known as convergent evolution: the situation in which two different structures evolve in genetically unrelated organisms to perform similar functions, or in which organisms that aren't closely related evolve similar traits in order to adapt to similar environments. As Putnam himself points out, there is a limited set of effective solutions to some challenges, and this means that in some cases one solution will develop independently again and again.

There are lots of examples of convergent evolution, here are a few of them.

- *Wings.* Wings have developed in many different forms. It is thought that the last common ancestor of both birds and bats, which have significantly different wing structure, did not have wings. Although there are obvious, significant, differences between the wing structure of birds and bats, they have to be

broadly similar in construction due to the physical constraints to which any type of wing construction is subject.

- *Sight*. Seeing is achieved through a variety of different physical mechanisms. One of the most famous examples of convergent evolution is the camera eye, which arises both in squid and in vertebrates. It is thought that the last common ancestor of squid and vertebrates had a simple photoreceptive spot. However, the eyes of the squid do differ from vertebrates' eyes, in that the squid's eye has the blood and nerve vessels entering from the back of the retina, while in vertebrates these enter from the front. Most insects have a totally different form of eye from that of vertebrates, they have a compound eye which has a much faster 'flicker fusion rate' (the frequency at which a flashing light appears steady); this allows the insect to absorb changes in what it sees far more quickly than humans and making it more difficult for humans to catch them.

- *Anteaters*. The giant armadillo in North America, the giant anteater in South America, the giant pangolin in Africa and the spiny anteater in Oceania all have similar looking heads. This is because they have all evolved to hunt in tall grass and eat ants. But they are not close relatives; their evolutionary origins are all quite different.

- *Antifreeze*. Fish have antifreeze mechanisms both in Antarctica and in the arctic. But the genes that control the mechanisms are quite different. Independent episodes of molecular evolution seem to have taken place in each location, and both have had the same result in terms of function.

It seems likely, therefore, that the reason why there is a range of physical-chemical states of a brain associated with avoidance behaviour (for example), is that if a creature was unable to experience something that caused it to take avoidance action when threatened with damage to its physical state it is very unlikely that the species would survive.

This suggests that we should *expect* evolution to produce a number of *different* mental states (using the term now to pick out feelings and thoughts, its normal meaning) that play the role of causing a creature to take avoidance action, that is, to be in the functional state Putnam associates with the one term 'pain', and there seems to be no reason why these mental states (these feelings) should not be qualitatively quite different, so there would be $pain_1$, $pain_2$, …, $pain_n$, where n is a large number, and each of these mental states could arise in the context of a brain state with its own, individual physical-chemical structure. This would be the case in different types of organism but, indeed, it seems likely that there are also a great number of different types of mental state playing

the role of pain, from time to time, in any one type of creature. The fact that we humans have one generic name for all these states does not mean that every time we use that word we are referring to the same mental state; there are many different kinds of pain. Thus it does seem likely that one so-called functional state could be associated with a great number of substantially different brain states, as suggested by Putnam; but if that is the case, surely it is likely that each of these brain states is associated with a different mental state (defined in the "what it is like" sense rather than in Putnam's function-related sense). There seems to be no good reason for assuming that the mental state, in this sense of the term, associated with each of the different brain states will be the same just because the function is the same.

Now let us look, briefly, at Putnam's proposal. Putnam askes the question: "Is pain a brain state?", arguing that it is not, because "another hypothesis is more plausible" (*ibid*., p 199). But of course it all depends on what you mean by 'pain'. If pain is *defined* to be a 'functional state', as Putnam suggests, then pain is definitely *not* a brain state, as is illustrated by the examples of functional states noted above. But if pain is taken to be a psychological, or mental state of the "what it feels like" variety, then the question is the same ('Is Consciousness a Brain State') as that addressed in some depth by Place (1956) and Smart (1958), see Section 3.3.3.

Had Putnam actually been making the case that there is one, single, unique, *feeling*, named by us 'pain', and asserting that this coexists with a whole range of different *brain* states across creatures, or across species (an idea that has come to be known as 'multiple realisation'), and had he been successful in making that case, then clearly this would have implications for the approach to mind known as the identity thesis (see Section 3.3.2), and possibly for physicalism in general, too. But it seems that that is not what he was suggesting.

In conclusion, multiple realisation of functions seems highly likely to arise as a result of convergent evolution, and a multitude of the different things we call 'mental states' (feelings and thoughts) could perfectly well be involved in the same function, as are a multitude of physical states; and therefore many different versions of the things we usually call 'mental states' could be involved in a single 'functional state' (referred to by Putnam as a 'mental state'). So Putnam's point about multiple realisation, which is true of functional states, is not relevant to the proposition investigated in this thesis, which is about the thoughts and feelings that we call 'mental states'. However, it should perhaps be noted that, although there is not, and never could have been, any means of testing

whether Putnam's proposal of multiple realisation of 'mental states' (as functional states) was in fact also true of 'mental states' (as thoughts and feelings, or psychological states), a number of respected philosophers of mind have signed up to it (see, for example, Yablo 1992, pp 249-50).

3.4.10 Emergentism and pan-psychism

If we are to espouse the proposition that there is another property of matter besides those studied in physics there is a question as to how this property arises: is it instantiated by matter wherever and whenever matter arises, a proposition known as 'panpsychism'; is it instantiated by some types of matter but not by others, as is the case with the physical (non-mental) properties of matter, such as electric charge, or mass, a proposition sometimes called 'micropsychism' (see Strawson, 2008, pp 70-71); or, is this property instantiated only when some particular combination of matter occurs in nature, a proposition known as 'emergentism' (see Broad, 1947)? I briefly consider each of these possibilities in this section.

When considering these possibilities, we should bear in mind that 'matter' is often implicitly defined as being "non-mental" or "non-experiential", which could be viewed as carrying with it the implication that the proposition that matter instantiates consciousness is incoherent. Many philosophers reject emergentism, and maybe that is because given their implicit assumptions about the nature of matter they find it implausible, perhaps even incoherent. But it has to be the case that matter instantiates consciousness if mind is a property of matter, the proposition we are developing in this thesis. So here we have to view matter as being, at the very least, capable of instantiating phenomenal experience.

Emergentism

Emergentism is the idea that new properties of matter might emerge in some way from combinations of matter or properties of matter, and that these new properties of matter might have characteristics that could never, even in principal, be derived from knowledge of the things from which they emerge. C D Broad, in his book *Mind and its Place in Nature* (1925) defined emergentism thus:

> "Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations

> of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not the same as R; and that the characteristics of the whole R(A,B,C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A,B,C)."

He then notes that "The mechanistic theory [the identity theory] rejects the last clause of this assertion." (p 61, my insertion).

As mentioned above, many philosophers reject emergentism, perceiving it to be an impossible situation, perhaps implicitly taking 'matter' to be "non-experiential" by definition. But emergentism is not incoherent, for it is conceivable (it is logically possible): if scientists were to announce, some day in the far distant future, that they had created a conscious creature by combining simple chemicals into some particular structure, we might view it with astonishment, and we might seek alternative explanations, but it would not be immediately apparent that they were lying.

Clearly, if it were found that emergentism held, so mind could be viewed as being a non-reductive property of matter, it would also be the case either that mind 'supervened' on the properties of matter studied in physics, so supervenience physicalism could be said to hold, or that it was a property of matter in its own right (see Figure 1 in Section 1.1). In either case, there would be two classes of property of matter so property dualism would be the case.

Panpsychism

One alternative to emergentism, given property dualism, is the idea that some crucial element of consciousness, or "proto-consciousness", arises everywhere and always (hence the name) alongside the properties of matter studied in physics; this idea is known as panpsychism. If we opt for the bundle theory (see Section 1.2.4) there would be no substratum, only a bundle of properties. In that case some fundamental element relating to consciousness would have to have existed in the universe since time began, as is the case in the set of beliefs known as Vedānta.

But it is difficult to see why we should assume that in some form, consciousness is everywhere, always. The other properties of matter turn up in different combinations; neutrinos have no charge, photons have no mass, and so on, so why should this not be the case for consciousness?

Micro-psychism

Galen Strawson says that he believes that: "… experiential phenomena cannot be emergent from wholly non-experiential phenomena" (2008, p 70). Like many others, he finds unacceptable the idea, presented succinctly in Broad's definition above, that some particular combination of constituents of matter could possess characteristics totally different from those of the constituents taken on their own or in other combinations. He therefore puts the case for what he calls 'micropsychism', the proposition that at least some 'ultimates' [fundamental physical entities] are intrinsically experiential, the idea that: "… if experience like ours … emerges from something that is not experience like ours …, then that something must already be experiential in some sense or another (p 70).

Strawson then takes the view that if micro-psychism were found to be the case, then in the future: "… the idea that some but not all physical ultimates are experiential might look like the idea that some but not all physical ultimates are spatio-temporal …" (*ibid*., p 71). But one could argue that it would look like the idea that some but not all physical ultimates have charge or mass.

Conclusion

Clearly, one of these three options must be the case if property dualism is to hold. But we have no need in the context of this thesis to take a view as to which of them is the most plausible.

3.5 Conclusion

We have been considering three possible fundamental hypotheses for the ontology of consciousness. First, substance dualism could be the case; consciousness could be a property of a substance different from matter, as proposed by Descartes and still thought to be the case by some. The main problem with this approach, apart from its increasing implausibility (given the development of scientific understanding regarding the relation between the brain and the mind) relates to mental causation: if substance dualism holds, either mental states are epiphenomenal, or some of the causes of the events we observe originate from outside matter. Many find it difficult to conceive how an event in one substance could have an effect on the properties relating to another substance. Originally, this incomprehension was based on the assumption that for one thing to affect another the two had to be close to each other in space; that does not now seem to be enough to

enable us to rule out, once and for all, the 'two substance' approach to the ontology of consciousness. But this possibility is not the subject of this thesis, so I do not consider this option further apart from in Chapter 5, where it arises in the context of free will.

Second, perhaps it could be that reductive physicalism is the correct explanation. In this case, it is taken that everything to do with minds and mental states can be ontologically reduced to the properties of matter that are studied in the physical sciences. Many philosophers subscribe to the proposition that we *know* that we cannot communicate to another person exactly what our feelings are like (see Section 1.4.2), so there must be an explanatory gap, and yet believe that, at the same time, ontological reductionism holds.

The third option is non-reductive physicalism. If this is the situation that pertains, some form of property dualism will be the case regarding the ontology of consciousness, along with either emergentism or pan/micropsychism. The problem here is again that of mental causation. Without effective mental causation, the approach seems counter-intuitive, for it is difficult to see why consciousness would ever have taken the world by storm in the way it has if mental characteristics could not be causally effective (see Chapter 4 for various possible explanations of this). But clearly, if effective mental causation were to hold, physics would not be causally complete in the sense, generally accepted amongst philosophers nowadays, that all physical effects are fully caused by purely *physical* (non-mental) prior histories. And of course, 'over-determination' is not generally considered to be acceptable either (but see Section 3.4.5). But there is no insuperable problem here. With this other property of matter included in the web of laws of nature that constitute 'science' (see Section 1.2.4), science would be complete in the sense that every cause and every effect, mental as well as physical (non-mental), associated with every law of nature would be included.

The most serious problem with this approach is therefore the question regarding the mechanism through which mental states can affect physical (non-mental) properties of matter, addressed in Section 2.5; it remains to be seen whether scientists could, if they addressed the question, make any real progress with identifying a mechanism for effective mental causation.

Chapter 4 The role of consciousness

4.1 Introduction

In this chapter we look in more depth at Proposition 3 in Section 1.4.3, the idea that we all know, *from introspection*, that the ability to experience feelings is *necessary* for practical reasoning: the taking of effective multi-variate decisions (decisions in which the expected outcomes of different choices are not directly comparable because the issues they involve are not commensurable). There are two consequences of that proposition that deserve further examination. The first is that enabling this type of decision to be taken could be the reason why consciousness has taken the world by storm; enabling these decisions could be the *role* of consciousness. The second is that this proposition would have a number of consequences regarding the structure of the brain, and these can be checked against what we know about brains. In this chapter I examine both these issues.

In the next section we examine the possibility that the role of consciousness is to enable the taking of multi-variate decisions. Then in Section 4.3 we consider in more depth whether this proposition is justified given our general understanding of how decisions are taken. Section 4.4 notes that this type of decision-taking can be viewed as being nothing more than a rather sophisticated extension of the biological principle known as 'homeostasis'.

A number of consequences of this hypothesis can then be checked against our current knowledge of neuroscience, see Section 4.5:

- If the hypothesis is true, and if materialism holds (everything pertains to matter), then there must be some part of a person's body, most likely in the brain, that performs this task of action selection. The evidence currently available suggests that this function, which I am suggesting in this thesis involves a mental state of taking a decision triggering a physical action, is performed within the basal ganglia. The basal ganglia have a multitude of connections both to the sensory and to the planning areas of the cerebral cortex, as well as to the motor neurons (see Section 4.5.1).

- If this is where the mental causation actually occurs, damage to the basal ganglia or its connections, perhaps because of illness or accident or perhaps for genetic reasons, would prevent the mental causation taking place; and we would expect that if it was not possible for a person's feelings to influence their actions, the

person would be unable to take sensible and effective multi-variate decisions. It appears that this is so (see, for example, the case of Phineas Gage, described in Section 4.5.2).

- Next, having consciousness, and therefore being able to take effective decisions of this sort, will increase the likelihood that a creature survives and flourishes, and therefore species with this ability would be expected to flourish in evolutionary terms. This means that we would expect that the creatures from which we are descended evolutionarily would also have consciousness of some sort, and that they also, therefore, and their other descendants, would possess a structure in their brain in which the taking of a decision can affect the physics, namely, brain parts that perform the function that basal ganglia seemingly perform for vertebrates. And, indeed, there is evidence suggesting that this is so (see Section 4.5.3).

- Also, we would expect these other creatures to behave in a manner consistent with being able to experience feelings that are similar to ours. Of course, we can never know for certain what it is like to be another creature, but there is evidence that suggests that other creatures might well have feelings much like ours (Section 4.5.3).

Finally, in Section 4.6, I note the implications of this assumption regarding the role of consciousness for the ontology of mind.

4.2 A role for consciousness

One of many important issues relating to consciousness concerns the reason why it has come to be such a striking characteristic of humans, possibly other vertebrates and perhaps other creatures also, given that it remains unclear exactly what role it plays. The most obvious explanation for the huge success of creatures with consciousness would be if, whatever its role turns out to be, that role is of key importance in terms of survival and so has been selected for in the process of evolution. Obviously, if mental states in general, and those that are conscious in particular, play a role that is beneficial to survival, creatures capable of experiencing mental states would have an advantage in the fight for survival over others that were not capable of consciousness; evolution would then favour those capable of consciousness and this would explain why they have flourished.

A commonly held view amongst philosophers (see, for example, Rosenthal, 2008) has it that anything that could be done by a creature as a result of it being capable of consciousness could also be done without involving consciousness. Indeed, Harnad, who

subscribes to the view that consciousness plays no key role in facilitating survival, has set a challenge to those who propose some adaptive advantage for consciousness, requiring that they explain why the same advantage could not be gained in the absence of consciousness "with *exactly the same causal mechanism"* (his italics, 2000, p 3). This proposition has been named 'conscious inessentialism' (see Flanagan & Polger, 1995). If this proposition were correct, consciousness would either be epiphenomenal, with no effect whatsoever on behaviour, or, if it did have some effect on behaviour, it would be possible to acquire the advantage it delivered in other ways, in the absence of consciousness. There would then be no *need* for it to be selected for by an increased propensity to survive. If this were so, consciousness would presumably be a spin-off from some other property, one that is essential to survival; it would be an inescapable by-product of the other property but would have no key functional role of its own. Gould & Lewontin (1979), for example, warn against assuming that something is important for survival in case this might, as they put it, "invert the proper path of analysis", making an analogy with the architectural spandrels of San Marco which are adorned with a design "so elaborate, harmonious and purposeful that we are tempted to view it as the starting point of any analysis" (p 582). And Jackson (1982, p 134) points out that the weight of the polar bear's heavy coat might be viewed as being a hindrance to survival, evidence against the efficacy of evolution, if one were unaware of the enormous importance of having a warm coat to the survival of polar bears.

The idea in our third proposition (Section 1.4.3) is that some of the information needed for practical reasoning is phenomenal information – facts of the type: "it would feel like *that* to experience X". And a creature must be capable of consciousness if it is going to be able to experience feelings and store what those feelings are like in its long-term memory. It is important to note that it is not necessary that these decisions are actually taken consciously in order that consciousness is required for the process to take place; it could be that the process, when not a conscious process, uses material that was conscious at some point in the past and has been programmed for automatic response or is stored in a long-term memory and can be used in unconscious cognitive analysis. Clearly, this proposal meets Harnad's challenge in that, if it were correct, it would be necessary for the taking of this type of decision that the creature *can* experience consciousness; conscious inessentialism would not hold.

Importantly, if it turns out that feelings are necessary for the taking of this type of decision there has to be some way in which the phenomenal information, what the feelings

that are key to this decision-taking process *feel* like, can influence a physical (non-mental) property of matter; the decision is going to cause a physical (non-mental) effect to occur, such as the instigation of the movement of a limb, or speech, so mental causation has to be effective (see Section 2.4). So the proposition that the ability to be conscious of phenomenal information is *necessary* for the taking of this type of decisions – the fact that we know, *from introspection*, that the way we feel about the alternative options affects which action we choose to perform when we take these decisions – has important implications for the ontology of consciousness and for the causal completeness of physics (as we have seen in Section 2.4; although physics would not be causally complete in these circumstances, science could still be).

There have been a number of proposals for the role that consciousness might play that are quite dissimilar to the suggestion in this thesis. For example, Baars (2002) suggests that what is special about consciousness is that it "facilitates widespread access between otherwise independent brain functions" (p 47) and a number of people support the suggestion, notably Morsella (2005). Nichols & Grantham (2000, p 668), perhaps in a similar vein but relating the role to actions, suggest that "phenomenal consciousness serves the function of integrating information in the service of reasoning and action".

Nicholas Humphrey comes up with the interesting suggestion that consciousness gives a creature: "the ability to understand, predict and manipulate the behaviour of other members of his own species …", without which, he says, "a person could hardly survive from day to day" (2008, p 77). This raises the question of whether a species such as ours, along with many others, could get into the situation of mutual dependence which he describes so vividly in the absence of consciousness, and suggests that there may be a further evolutionary role for consciousness over and above that of facilitating multi-variate decisions: that of enabling the benefits of living in herds to be accessed in the interests of survival.

Many people have proposed roles for consciousness that are broadly similar to the suggestion in this thesis, although they take a different view from mine regarding the ontology of consciousness. Peter Carruthers is one of a number of people who think that consciousness plays a role in the enabling of what he calls "increasingly flexible behaviours". He suggests that the "evolutionary gains … come from the increasingly flexible behaviours which are permitted" as a result of "conceptual representations of the environment generated by perception" and "reasoning in the light of those representations" (2000, p 125); but he takes mental states to be "physical states of the

brain" (p xiii). Daniel Dennett says there is a role for consciousness in providing an answer to the "Now what do I do?" question (1991, pp 177, 188), but he envisages that consciousness is a 'functional' phenomenon, with the implication that any physical set-up with a similar structure to the relevant parts of the brain could perform the same role in decision-taking as that of the mind (see Sections 3.4.8). Armstrong (1968), like Carruthers, subscribes to the view that "mental states are nothing but physical states of the brain" (p xi), and he suggests that "in a problem-situation various possible situations may be tried out 'in the imagination' in order to see which response will best fit the agent's purpose" (p 163). Dehaene & Naccache (2001, p 11) suggest that consciousness is required for, among many other things, "the spontaneous generation of intentional behaviour", but also remark that: "Within a materialistic framework, each instance of mental activity is also a physical brain state" (p 3). None of these will pass Harnad's challenge.

The idea here is that consciousness is *necessary* for a broad class of decisions regarding what to do next, rather than just being helpful as in the suggestions noted above. If this is correct, the reason why the ability to experience feelings, and to anticipate experiencing them, has developed through evolutionary adaptation over the aeons would be because it enables creatures to choose the best, or a near-best, action quickly and efficiently given the set of alternative actions available in some situation.

4.3 The role of feelings in practical reasoning

The term used by philosophers to pick out the ability of the mind to select the most desirable action from the set of available options when taking a decision is 'the will'. Having a 'will' enables a person to perform practical reasoning and to implement the action that practical reasoning picks out as having the most desirable consequences, as being the one to choose. Practical reasoning is the process people use when they solve the action selection problem, that is, when they set about deciding what to do next.[73] Where a range of options for action is available, each with a number of possible consequences, practical reasoning involves the consideration of each option in terms of its pros and cons and then the determination of which is most desirable, or the picking of one that seems acceptable

[73] This process is similar to the one economists call 'cost-benefit analysis' and psychologists and neuroscientists call 'action selection'. Descriptions of practical reasoning, broadly similar to the one here, can be found in any text on decision theory. An early explicit use of practical reasoning is Pascal's Wager, described in his *Pensées* (1660, Section 233).

if time is short, through a process of "qualitative evaluation". Qualitative evaluation involves feelings.

4.3.1 The action selection problem

Some decisions are taken by reflex, an automatic response to some stimulation that bypasses the decision-taking parts of the brain; others are programmed into the neural connections in the brain, either before the creature is born or later, in the course of the creature's life. In both of these cases the action initiated by the creature is an automatic consequence of the situation (typically it is of the "if A then B, given situation C" type, although there may be probabilities involved) and it requires no conscious intervention or deliberation. The driving notion in this section is the fact that there are other important types of situation in which a creature is faced with a variety of possible courses of action each with a range of likely outcomes, probably including a large dose of uncertainty given the facts available to the decision-taker, and that in order to choose the most desirable option the creature has to take a view about what the consequences of each option might be like to experience, what they might *feel* like. We call such a decision a 'multi-variate decision' because the decision involves assessing very different types of pleasure or pain (sensory pleasure against risk of pain; benefits to others against future benefits for ourselves) to distinguish it from the other types of decision in which, for example, more of something may be obviously better than less. Choices of this sort vary widely in degree of urgency. The ability of a creature to survive and flourish, and therefore the survival of its species, can depend on it being able to make a quick and efficient decision regarding the available courses of action.

An example of the simplest type of multi-variate decision taken using practical reasoning is that of Peter Rabbit (see Beatrix Potter 1902, *The Tale of Peter Rabbit*). Peter Rabbit is hungry. He has to decide whether to eat the grass on the heath, which doesn't taste that good, or whether instead to go into Mr McGregor's garden to eat the delicious lettuces, but with a risk of being chased, shot, and put in a pot. He has to trade familiar varieties of taste against degrees of risk and fear of pain and death when deciding on his course of action.

4.3.2 What does practical reasoning involve?

The key feature of practical reasoning arises in a situation in which there are various actions that might be chosen and each results in a different bundle of outcomes: if the most desirable, or a reasonably acceptable, outcome is to be identified, then (see Watson, 1975, p 346, and Section 1.4.3) each of these bundles of likely outcomes needs to be 'evaluated' in some way so the one considered most desirable, all things considered, can be chosen and initiated. This evaluation process requires the chooser to imagine what the outcomes might *feel* like, so he or she needs to access memories relating to feelings, and to be able to access memories relating to feelings, to experiencing something, requires that the creature in question has the capacity to feel, that is, it means that the creature must be capable of consciousness.[74]

Decisions of this sort, multi-variate decisions, require the following. First, the creature must be able to identify the alternative courses of action available in some situation and the range of likely outcomes these actions could lead to. Then there needs to be an assessment of the feelings, the satisfactions and pains, which are likely to be associated with each of these outcomes. Many feelings have an aspect to them, known to psychologists as 'motivational affect' (effectively, an incentive effect) according to which a situation is viewed as being desirable (positive) or undesirable (negative) in some degree; 'Motivational affect' is valenced, in the sense that it is experienced in varying degrees, on a continuum. The 'motivational affect' associated with the likely outcomes of a proposed action indicates the desirability, the net intensity of the likely satisfaction or pain expected if that action were initiated, and according to psychologists it is this that determines the outcome of the decision regarding what to do next.[75] In his 1980 paper, 'Feeling and Thinking: Preferences Need No Inferences', Zajonc promotes the role of 'motivational affect' (as opposed to cognitive analysis, but accepting that both are necessary for decision-taking) saying, for example: "People do not get married or divorced, commit murder or suicide, or lay down their lives for freedom upon a detailed cognitive analysis of the pros

[74] Note that if this is correct a philosophical zombie would be unable to take decisions using practical reasoning, because a zombie has no consciousness and so is unable to 'feel' things, and if this is so, and if, as seems likely, decisions taken using silicon do not pick out the same choices as decisions based on feelings, taken by humans, the existence of a philosophical zombie is not conceivable, see Section 2.6.2.

[75] This may be what Crane had in mind when he wrote his book *Elements of Mind* (2001) and spoke of 'intentionality'.

and cons of their actions." If this is correct, it would be the 'motivational affect' associated with the consideration of an option that indicates its desirability and thus it is this that results in a person's actions.

The final step in decision-taking, then, is a comparison of the net intensity of the 'motivational affect' associated with each of the actions available; this will enable the most desirable (or, if time is short, an acceptable) option to be identified; an 'affect comparison' function is required, therefore, in order that a person can perform the necessary task of comparing the 'motivational affect' he or she experiences when contemplating the alternative options available and pick the option perceived to be most desirable. It is difficult to see how physics, which operates in laws of the "If A then B given C" variety, can perform this task. What is needed is a process that can solve the problem:

"choose A_i to maximise F(…), given C"

(see Section 1.4.3). The idea presented in this thesis is that consciousness enables that function to be performed.

If this is the case, one could argue that we should guard against falling into the reverse of the spandrels fallacy: to us, the effect of consciousness is so important that it might lead us to, as Gould and Lewontin (1979) put it, "invert the proper path of analysis". Whereas one might imagine all sorts of weird and wonderful reasons for the existence of something that means as much to us as consciousness does, in fact the reason for its existence could simply be that it enables the taking of the multi-variate decisions that will enhance the likelihood of a creature, and its kind, surviving. The suggestion in this section is that in order for the first, basic, creatures to evolve into more complex and sophisticated creatures it was necessary that this function could be performed, that there was some way in which this role of "choosing" could be fulfilled by the creature. The idea is that physics on its own, dependent of laws of the type "if A then B given C", could not do it. But it turned out that this other property of matter could. So this other property of matter, consciousness, has proliferated in our world. Seen by a Martian, or by some hypothetical sensate being outside our human race, consciousness would then look as if it was merely a gadget that is successful in increasing the likelihood of survival of the creatures that live on this planet.

The theory of multi-variate decision-taking has been developed over a great number of years (see, for example Pascal, 1660; von Neumann & Morgenstern, 1947; Luce & Raiffa, 1957; Markovitz, 1952; Kahneman, 2011). Originally it was thought that the

"expected value" of the outcomes was the key factor that determined the choice, but then it was realised that uncertainty can be highly significant in determining what people choose (both in the form of the variance of the outcomes, see Markowitz, 1952, and in terms of the skew, where there may be a small probability of a very good or a very bad outcome, indicated by the third moment of the probability distribution, see Kahneman & Tversky, 1979). It is likely that there is still much to be understood about risk perception, however. Significant discrepancies arise between risk perception as reported by the general public and predictions of risk based on experts' views of technology and statistical theory when risks of different sorts are compared, as when the risk of death from a nuclear accident is compared with that from a road accident (see Slovik, 1987). Also, a paper by Greene *et al.* (2001) has shown that activity in areas of the brain related to emotion is significantly increased when there is personal involvement in the question at issue compared to when there is not. This may be relevant to the famous trolley dilemma (Foot, 1967), where the decision is whether to allow five people to die or to kill one, other, person by turning a switch, and its partner, the footbridge dilemma (Thomson, 1976), where in order to save the five people it is necessary to push one (other) person to his or her death. It seems likely that this is because the chooser factors in the expectation of being badly afflicted personally after the footbridge dilemma, both by painful memories and by public opprobrium; this would be less of a problem in the case of the trolley dilemma. The emotion associated with this would affect the valuation of the alternatives.

In any case, people's feelings regarding the outcomes of their actions, which derive from their 'values' (the person's dispositions to experience certain types of feelings in certain circumstances), are viewed as being a key factor in the assessment of the alternative possible actions in decisions which depend on the use of practical reasoning.

4.3.3 Multi-variate decision-taking and awareness

A question that now arises concerns the extent to which awareness is involved in the taking of these decisions, and whether awareness is necessary at the time the decision is being taken. The answer seems to be that whereas decisions of this sort can be, and frequently are, taken without awareness of any sort being involved, this can only the case if the results of previous analysis of similar issues relating to the same feelings are already available in the memory.

A number of reasons have been cited in the literature that suggest that consciousness is not necessary for the taking of this sort of decision. First, we now know

that conscious capacity is severely limited in that the working memory, where cognitive processes of this sort are worked out consciously (see Section 1.3.5) can only hold only about seven items, plus or minus two, at a time (Miller, 1956). So we know that much of the processing needed before a multi-lateral decision is taken must be done unconsciously (see also Dijkaterhuis & Nordgren, 2005, on unconscious thought). But since feelings are a necessary part of such decision-taking, and a creature must be capable of consciousness in order to experience feelings, this processing must be informed by conscious analysis that has taken place previously.

Next Rosenthal, who subscribes to the proposition that the taking of decisions of this sort does not necessarily require consciousness, suggests that in order for consciousness to be necessary for practical reasoning it must be the case that intentional action only results from conscious volitions (2008, p 833). But this may not be the case. A creature that has experienced the relevant feelings, and can store information about them in its memory, is able to use that information in unconscious analysis as well as in conscious analysis. It is not necessarily required that such decisions are always taken in the spotlight of consciousness, although given that it is our feelings that determine the outcome of the choice process, the ability to experience consciousness must be a necessary concomitant for these decisions to happen; the memory of previous conscious experiences stored in a creature's long-term memory store can be used to take decisions of this sort without consciousness being needed at the time the decision is taken.

Then there are two experimental results that may be deceptive in suggesting that decisions can be taken without consciousness of feelings ever occurring. First, as we saw earlier (Section 2.3.2), the work of Libet *et al.* (1982) shows that certain rather trivial decisions concerning the timing of actions can be taken before the subject is aware of having taken them, so consciousness can have played no part in the triggering of these decisions. But of course, it is not necessarily the case that, as suggested by Libet *et al.*, *all* our decisions are taken in that manner.

Second, the famous case of the 'hand-shaking' experiment, reported by Édouard Claparède (1911) is often interpreted as indicating that decisions can be taken without consciousness. In this experiment, the subject was unable to remember the situation surrounding the pain she experienced on shaking the psychologist's hand when he had concealed a pin in his hand because she suffered from a form of amnesia. But afterwards she was always reluctant to shake his hand. This has often been interpreted as implying that knowledge that is not available to consciousness can inform decisions. But whereas

it certainly seems to be the case that decisions of this sort can be taken without conscious awareness, this is not necessarily the correct interpretation of Claparède's experiment. The subject could well have associated 'motivational affect' with shaking his hand, and it is awareness of 'motivational affect' that drives conscious decisions, while having no memory at all, conscious or otherwise, of the circumstances that had originally resulted in this recalled 'motivational affect'.

If practical reasoning is actually the way in which this type of decision is taken, feelings can be viewed as being instantiations of the particular mechanism that enables certain types of creature to take multi-variate decisions effectively and thereby enhances the likelihood of their survival.

4.4 Practical reasoning as a form of homeostasis

In order to survive, and therefore reproduce, organisms must maintain a relatively stable situation regarding a number of aspects of their internal environment such as the pH level in their blood, their temperature and their blood pressure. The mechanism by which this is done is the negative feed-back process known as 'homeostasis', named such following the work of Claude Bernard published in 1865. Homeostatic control mechanisms have three components. First, they have a sensing component that monitors and responds to changes in the environment; this is known as the 'receptor'. Second, they have some sort of a 'control centre' that determines whether or not action is required. Third, they have an 'effector'; this is a mechanism which operates to correct any significant deviation from the acceptable range by initiating a change in the behaviour of some muscle or organ or some-such.

However, in the more sophisticated creatures a number of the feedback mechanisms necessary for survival, and for quality of life, are the result of intentional behaviour, that is, they follow from an action taken by the organism seemingly triggered by an awareness of some imbalance (see Panksepp, 2011, who identifies certain 'primal affects' along with their 'comfort zones' and potential negative effects on survival). 'Motivational affect', the component of an emotion/feeling that drives an intentional action, can therefore be viewed as being the trigger for a form of homeostasis (as suggested by Paulus, 2007); cognitive processes are involved in determining the range of possible actions and informing the choice, and motor neurons ensure that the muscles respond to and implement the chosen action. It is fundamental to this process that the emotion

experienced has an incentive effect associated with it, thus causing the person/creature to decide to perform an action. Thus, a person/creature will seek something to eat when it is hungry, something to drink when it is thirsty (the internal feed-back system regarding osmosis and thirst, that depends on the hypothalamus, can only deal with small variations), somewhere to rest when it is tired and somewhere warm when it is cold. For instance, reptiles, which don't have an effective internal temperature adjustment system, often choose to lie on a sun-heated rock in the morning so as to increase their body temperature. A person, and perhaps a creature of another type, may also be concerned to maintain the welfare of loved ones, and to keep their own self-esteem and therefore, perhaps, moral behaviour within acceptable ranges.

Thus consciousness, enabling the experience of a great variety of feelings and thoughts, can be regarded as an essential part of a particular sophisticated biological development of the homeostasis type.

4.5 Evidence for the selection of actions through practical reasoning

If the role of consciousness, and of 'motivational affect' in particular, is to enable the effective taking of multi-variate decisions relating to actions, so this is the reason why consciousness has been so successful in evolutionary terms, there should be some evidence amongst our current understanding of neuroscience that supports this proposition. In this section I report my findings from a search for three types of such evidence: evidence relating to the part of the brain that facilitates this type of decision-taking; evidence relating to what happens when that part of the brain is damaged; and, evidence that suggests that other creatures descended from the same evolutionary ancestors as us have similar brain structures, that these might play a similar role in facilitating decision-taking and that these creatures might therefore experience feelings broadly similar to those which we experience.

4.5.1 The physical correlate of action selection

It is currently thought that the neural correlate of action selection is the basal ganglia. The basal ganglia are part of the limbic system, a collection of structures which includes the hypothalamus, the hippocampus and the amygdala as well as the basal ganglia and is located below the cerebrum in humans. Ever since the publication of Papez' 1937 paper on the

hypothalamus and its connections it has been generally accepted that the limbic system is associated with feelings and emotions.

There is a range of evidence supportive of the view that the basal ganglia are associated with decisions relating to actions. First there was the discovery, some years ago now, that Parkinson's disease involves a deterioration in certain types of neuron in the basal ganglia, and then that in Huntingdon's disease an abnormal form of a protein causes some of the cells in the basal ganglia to malfunction (Marsden, 1981 contains much useful background information on the role of the basal ganglia). Next, the key inputs to the basal ganglia come from the cerebral cortex, which is responsible for processing sensory input, and for reasoning and planning, and the output from the basal ganglia, after first returning to the cerebral cortex, travels *via* the thalamas to the brain stem and then on to the muscles that trigger actions. It has been known for some time that the primary motor cortex has axons which extend all the way to the basal ganglia (see Mink, 1996). All of these findings support the proposition that the basal ganglia are the structures in the brain in which decisions concerning actions are taken. The idea is that whereas the possible actions are identified in the cerebral cortex, it is within the basal ganglia that it is decided which action, out of the set which the cortex has identified as being available, is the one likely to produce the most desirable outcome and so is the one that should be executed.

With this in mind, Redgrave *et al.* (1999) have proposed a model in which systems in the brain that are associated with certain 'affective' emotions, such as those concerned with eating or hunting or sleeping or reproducing, compete for access to the motor resources that are needed for any one of these to achieve their ends. Somehow, the causal mechanism in the basal ganglia (the mechanism that enables a mental state to influence the physics) must enable the selection of the one with the most pressing needs.[76] Redgrave *et al.* identify three methodologies that could solve this action selection problem; one has a pre-ordained priority order across the competing initiating systems, the second postulates inhibitory bilateral connections between each pair of systems and an excitatory link to the shared motor resource, while the third has a central selection mechanism that inhibits access to the motor neurons for the systems that are not preferred and disinhibits access for the one that is favoured. They thought the third one seemed the most likely, partly

[76] A useful definition of the selection problem is the following: "… simply stated, the selection problem is: at any time, which functional system, or sensory event, should be put in a position to direct the shared motor resources – the final common motor path?" (see Redgrave *et al.*, 2011, p 140).

because it is the most efficient (it requires fewer connections for each competing system and for a new entrant), and partly because it is the only system that enables the selection mechanism to evolve without the changes having any effect on other aspects of the control mechanism. Unlike in most other regions of the brain, in the basal ganglia the great majority of neurons use a neurotransmitter that is inhibitory in nature (GABA), so the selection process probably involves ensuring that the inhibition of those channels that are not selected is maintained or increased while the selected channel is disinhibited to enable access to the necessary motor resource. In a later paper, Redgrave *et al.* (2011), it is argued that: "… the re-entrant looped architecture of the basal ganglia represents biological solutions to [the] fundamental behavioural problems of selection and reinforcement" (p 138).

The proposal that it is the basal ganglia that is the brain structure within which this important function takes place gains support from an increasing number of psychology experiments. Structures in the basal ganglia have been found to be active when choices are made, see for example Knutson & Greer (2008), who were seeking the neural correlates of anticipatory affect – the feelings people experience when anticipating outcomes that matter to them. Further, it now seems likely that the bit of the basal ganglia that assesses the intensity of motivation is the caudate nucleus. Delgado *et al.* (2004) conducted an experiment in which the only thing that varied was the intensity of the motivation: they asked their subjects to guess the values of a series of cards with the visual feedback varying between simply recording their success or failure or recording a monetary reward when they succeeded and a monetary charge when they failed. The magnitude of the response in the caudate nucleus varied significantly depending on the type of reward offered.

It seems likely, therefore, that it is within the basal ganglia that the mental influences the physics – that this is where mental causation actually takes place (see Section 4.3), and current knowledge suggests that the caudate nucleus may be the bit of the basal ganglia where the *valence* (the intensity) of the 'motivational affect' (the incentive effect) associated with each of the alternatives is assessed.

4.5.2 What if the ability to experience 'affect' is damaged?

Direct evidence relating to the mechanisms we humans use when we are taking multi-variate decisions is available from the effect of certain brain lesions on people's behaviour; these lesions may arise as the result of accident or illness, or following the medical treatment of brain conditions.

There is one particular brain lesion that seems to leave all the cognitive aspects of a person's abilities intact and working perfectly well, but which means that the person is unable to take multi-variate decisions effectively (see John Darrell Van Horn, 2012). The lesion in question is the one relating to the well-known case of Phineas Gage, who had an iron bar 3 cm thick pass through his head following an accidental explosion when building a railway in the US in 1848. Afterwards he appeared at first to have experienced no damage whatsoever to his mental abilities. But it turned out that he had suffered a puzzling change to his personality that meant that he was no longer able to hold down a job and that his marriage and all his other relationships broke down. The case puzzled the doctors. Many years later, in the early 1970s, Antonio Damasio, a professor of neuroscience in California, was invited to examine a patient, known as 'Elliot', who had had a tumour removed and this had left him with disabilities that had the same effect on his life as Phineas' had had on his. So Damasio was able to perform tests on Elliot to try to discover exactly what it was that was causing the problems (see Damasio, 1994).

Elliot performed perfectly, in the laboratory, on every mental test that had at that time been invented. So the doctors created a new set of tests relating to decision-taking in social situations; these included listing the possible options for action, predicting the consequences of these actions, devising strategies to achieve objectives, and analysing the options in moral dilemmas (for example, should one steal a drug to prevent one's spouse from dying?). Elliot performed well on all these tests, also. However, he is reported as commenting at one point, with a smile, after analysing the possible outcomes in some tricky hypothetical situation in the laboratory, "and after all this I still wouldn't know what to do!" Eventually Elliot himself identified what was wrong with him: he described his total lack of emotional response saying, according to Damasio, that "topics that had once evoked a strong emotion no longer caused any reaction, positive or negative" (Damasio, *ibid*., p 45). Other patients with similar lesions have been found to behave in the same way as Elliot when subjected to these tests. The implication seems to be that effective choice involves an assessment of the feelings attached to the available alternatives, and without being able to experience the emotion, the 'affect' – if one lacks the ability to care – the choice made is arbitrary. Although the person can predict the consequences of the actions they might choose, they seem to be unable to assess how these consequences would *feel*, so they are unable to make a soundly based decision.

The implication of this is important: although cognition plays an important role in the choice of alternative actions and the analysis of the likely outcomes, cognition alone,

it seems, without the emotional responses associated with our 'values', our dispositions to experience certain types of feelings, a soundly based decision, a decision that is consistent with the preservation of the person's long term welfare, cannot be taken. The phenomenal information regarding the feelings experienced is necessary for sound decisions to be taken and there has to be some mechanism through which this informs our actions: mental causation must operate.

As well as supporting the proposition that 'affect' is important for the taking of deliberative decisions, and the idea that there is a structure in the brain which is necessary for such decisions to be taken, this evidence suggests that the experience of 'affect' itself is associated with a particular structure in the brain, thus providing evidence in support of the proposition that consciousness arises in the context of certain particular material structures.

4.5.3 The role of consciousness in evolution

Effective decision-taking is crucial to survival, so if the ability to experience feelings improves decision-taking one would expect that it would be selected for in evolution; one would expect that creatures with the ability to take decisions in this way would be more likely to survive than those without that ability. It could be that this ability to take multi-variate decisions sensibly is unique to humans, or to mammals, or to vertebrates; on the other hand, it also could be that some of the more ancient creatures from whom we are descended evolutionarily also had this ability. Further, if we have inherited this ability from the ancient creatures from whom we are descended evolutionarily it seems likely that some of the other creatures alive today that are descended from the same ancient creatures as us, and so are related to us in evolutionary terms, would also have these brain structures. If so, one might expect that these creatures alive today would then experience feelings broadly similar to some of those we experience, enabling them also to take this type of decision (people are sometimes reluctant to accept that other species of creatures may experience some of the feelings that we experience, but it clearly is a logical possibility). In this section I describe some of the evidence now available that supports the presumption that other creatures alive today, besides us humans, have brain regions *homologous* to our basal ganglia, that is, brain regions that have the same interconnections, use the same neurotransmitters and appear to play the same role in the creatures' survival as they play in ours.

For many years it has been well-established that different parts of the brain perform different functions. It has also been accepted that many creatures have brain regions that are homologous to ours implying that they have shared evolutionary ancestry with us (see Ghyson, 2003, for example). But, until quite recently it was thought that the cerebral cortex, where logical thinking and planning take place in humans, was unique to vertebrates. Also, although we have known for some time that all vertebrates have similar structures in their hindbrain which ensure that autonomic processes (continuous, automatic processes such as breathing, heart-beat, digestion *etc.*) continue to function effectively without the need for conscious control, it was thought that these brain parts, also, might be unique to vertebrates. Now it seems that in fact many other creatures may have brain structures that play the role our cerebral cortex plays. For example, Tomer *et al.* (2010) have found evidence suggesting that brain structures in marine ragworms (annelids), known as 'mushroom bodies', may be homologous to the vertebrates' cerebral cortex, and this is generally interpreted as meaning that there must have been a common ancestor to both vertebrates and annelids that had this feature in its brain structure, with the implication that the origins of the cerebral cortex date back to the Precambrian era (more than 500 million years ago). And it has now been established that the vertebrates' hindbrain structures are very similar, both anatomically and functionally, to a part of the brain in the arthropod (animals with external skeletons, such as insects) which is called the 'sub-oesophageal ganglion' (see Ghyson, *ibid.*, p 559). So it now seems likely that vertebrates, and therefore mammals, have an evolutionary ancestor in common with arthropods. This has been named 'Urbilateria' (see de Robertis & Sasai, 1996), and is thought to have become extant about 600 million years ago, suggesting that some form of consciousness has been around for that length of time.

When it comes to the basal ganglia, it has been accepted for the last couple of decades that structures homologous to our basal ganglia are found in all jawed vertebrates (Medina *et al.*, 1995). But more recently, structures in arthropods, known as the 'central complex', have been found to play the same role as the vertebrate's basal ganglia (Strausfeld and Hirth, 2013). So the presumption now has to be that the basal ganglia, also, derive from the so-called Urbilateria. And if we are right in thinking that the basal ganglia are the brain structures that compare the intensity of 'affect' associated with alternative options when choices are being made, this suggests that some form of feelings, and therefore consciousness, must exist in all the creatures, including insects, crabs and such-like, that are thought to have derived from the Urbilateria through evolution. It seems probable,

therefore, that many of the emotions we experience, such as fear, anger, various 'needs', a desire to approach or seek, and so on, are also experienced in some form or another by the other animals that have these same brain structures. But of course there is no way of knowing what these might actually feel like for the creature concerned.

This is a proposition that has been espoused by Panksepp. He has proposed that our affective feelings derive from areas of the brain that we have inherited through the process of evolution, and that we should therefore assume that other animals that have evolutionary ancestors in common with us would experience feelings broadly similar to the ones we experience; see for example his review article (2011) and his book (2012). So a question arises as to whether there is any evidence that supports the proposition that the 'affective' elements of the feelings we experience as humans have developed in evolutionary terms from those of other animals (Panksepp, 2005).

This question was first put to the test in the middle of the last century. Stimulation of parts of the brain, either with electricity (applied to implanted electrodes) or using drugs, was found to cause animals to behave in a way very similar to the way in which humans behave in the same circumstances, and when humans behave in that way we know, because of what others tell us and also because of what we, ourselves, have experienced, that it is the result of certain pleasurable or painful feelings being induced. Delgado *et al.*, as long ago as 1954, reported that a "fear-like reaction" could be elicited in cats by electrical stimulation of various brain regions; and noted that this reaction "had all the drive properties of a true emotion …" (p 179). Olds & Milner, also in 1954, showed that rats would press a lever that activated electrical stimulation of particular regions in their brains, presumed to give them pleasure, in preference to eating or drinking; some eventually dying of exhaustion. There have been a great many experiments since then that are supportive of the proposition that other animals feel emotions similar to ours when similar brain regions are stimulated. For example, recent experiments suggest that a dog's caudate nucleus is activated when it perceives something pleasurable (Berns, 2015). This certainly suggests that emotions, and 'affect' in particular, play a key role in the selection of action, and therefore in survival, for other creatures besides us.

4.6 Implications for the ontology of consciousness

Finally, if feelings are actually *necessary* for the taking of effective multi-variate decisions, in that such decisions cannot be taken effectively in the absence of feelings (as spelt out above

and in our third fundamental proposition in Chapter 1, and as appears to be the case given the experiences of Phineas Gage and Eliot), mental causation *must* be effective. Further, since feelings are not reducible to the properties of matter studied in physics (physics is concerned only with such things as can be communicated perfectly from one person to another using language, and this is not the case with feelings), there must be something else besides what is studied in physics. So then, if we rule out dual substances, there has to be another property of matter besides those studied in physics and that property of matter has to be causally effective in that it has to be capable of influencing actions, that is, it must be capable of causing changes in the physical properties of matter, those that are studied in physics.

Put another way, if the role of consciousness is to enable decision-taking and we have seen in this chapter that there is considerable evidence in support of that proposition, feelings must be able to influence actions so feelings must be causally effective.

Chapter 5 Mind and the problem of free will

5.1 Introduction

My purpose in this chapter is to examine the question of whether our assumption of causally effective property dualism combined with determinism or 'universal probabilistic causation' (see below) is consistent with some acceptable form of free will: that is, whether compatibilism could hold.

However, the issues involved in the free will problem seem to be an order of magnitude more subtle than is the case for most of the other issues I address in this thesis. I start, therefore, for the avoidance of confusion, by spelling out precisely what I mean by each of the key terms used in the analysis such as 'determinism', 'the will' and 'moral responsibility'. We then turn to van Inwagen's Consequence Argument, which appears to imply that compatibilism, the proposition that both determinism and free will can hold, is incoherent. However finally, after a brief but important digression on the possible meanings associated with the personal pronoun (and the term 'person' and proper names), we establish that compatibilism of a sort is possible.

5.2 The meaning of 'determinism'

5.2.1 Determinism

'Determinism' is the doctrine that every event is the necessary consequence of what has gone before given the causal processes that pertain. According to the *Stanford Encyclopaedia of Philosophy*, the meaning of 'determinism' is given by: "The world is governed by (or is under the sway of) determinism if and only if, given a specified way things are at a time t, the way things go thereafter is fixed as a matter of natural law" (Hoefer, 2016). But if probabilistic causal processes are ruled out of 'determinism' by definition, then given the presumed fact that some such relations *are* probabilistic (*cf.*, the observation of the decay of a radium nucleus along with the presumed validity of quantum mechanics), determinism thus defined is perhaps of little interest. We need to coin a phrase for the situation that allows probabilistic causal processes as well as the 100 per cent certain causal processes of 'determinism'. In this thesis I use the term 'universal probabilistic causation' to include

probabilistic causal processes as well as causal processes that fix the outcome with 100 per cent certainty. Section 1.2.2 addresses more issues relating to causal processes and laws of nature.

If determinism were a true description of the universe, then every event that occurs would be related to some collection of things that could be labelled its (immediate) 'causes', given the situation that existed. Further, if determinism held the set of these causal processes would be 'complete', in the sense that *every* event, mental as well as physical, would have a cause of this kind; science would be complete. In the most obvious, simple, case there would be a law of nature that decreed that if event A were to be instantiated in circumstances C, then event B would follow with 100 per cent certainty.[77] It would then follow that, with perfect knowledge of some situation in the past along with perfect knowledge of all the causal processes every future event could be predicted (and every event that had happened in the past could be known). It could therefore be said that determinism thus defined implies something that one might call 'predetermination': the idea that everything that happens has been 'determined' in the above sense from the beginning of time, since the Big Bang.

There are two alternatives to this theme of strict determinism, and both are generally known to philosophers as 'indeterminism' because in both cases outcomes are unpredictable, or random. But the term 'random' has two distinct meanings: it can be a technical term indicating that something is subject to a well-defined probability distribution; alternatively, it can indicate total unpredictability, that is, the lack of any identifiable cause.[78] This means that there are two forms of 'indeterminism'.[79] The two forms of indeterminism have very different implications for free will so it is important to distinguish them.

[77] Generally, any particular effect has a number of possible immediate causal factors, various subsets of which are necessary for its occurrence, and any particular causal event is likely to have more than one effect. Also, as we shall see, the causes of any event in fact go backwards in time, in a forking chain, to the Big Bang. I state the situation here in the simplest possible terms.

[78] See Eagle, 2016 for a discussion of these terms.

[79] Philosophers often ignore the possibility of this second type of unpredictability and consider only the possibility of an event being determined by a process that incorporates a probability distribution; see for example Robert Kane's 2007 paper; in which a probabilistically determined situation is referred to as being 'undetermined', or see Fischer (2007).

5.2.2 Stochastic causal processes

First consider the type of indeterminism that involves the possibility that whereas there are causal processes associated with every event that occurs, some or all of these are probabilistic in the sense of incorporating a well-defined probability distribution, so that if event A were to occur in circumstances C, a law of nature might decree that event B would follow with some predetermined probability, perhaps with event D being instantiated in the situations in which B is not. Every event would still have a cause, and therefore could be said to be "determined by a causal process", but it would not be the case that, given knowledge of the causal processes and the situation at some point in time, every event could be predicted. Predetermination of the future would not hold. However, events could, in principle, be predicted with an appropriate probability given perfect knowledge of what has gone before along with perfect knowledge of the causal processes.

I call this form of (so-called) indeterminism 'universal probabilistic causation'. The term 'indeterminism' can be reserved for use in the case in which an event has no identifiable cause (spelt out in the next section).

There is nothing incoherent about this more general, probabilistic, interpretation of the causal processes; indeed, I think we can be pretty confident that there can be probability in a causal process: on a macro scale it seems plausible because of the stochastic manner in which the nuclei of radio-active elements, such as radium, decay; on a micro scale, because of the experimental support for the theoretical approach known as quantum mechanics (see Levine, 2001, p 70). However, it should come as no surprise that for many of the most familiar laws of nature, which relate to large-scale physics, such as Boyles law and the laws of motion, the probabilistic element is either extremely small or non-existent; a large stochastic element attaching to such a law would make it very much more difficult for us to discover it. It does not follow, of course, that the causal processes of psychology or economics, if such exist, would have minuscular probabilistic elements associated with them (see Section 3.4.5), and nor do the causal processes that arise in modern medicine and weather forecasting have small probabilistic elements; predicting in those disciplines is more akin to gambling.

This form of indeterminism, which we are calling 'universal probabilistic causation', has the same implications for free will as does determinism: in neither case can the operation of someone's will change the substance of a causal process, including the probability distributions implicit within it. If, because of the causal processes, a person in circumstances C has a probability p of choosing action B and a probability (1 – p) of

choosing action D then their behaviour will be consistent with that fact, and observations would find that to be the case. If universal probabilistic causation holds, the probability that a person will choose one action rather than another given the circumstances and given that person's values (their disposition to experience certain types of feelings in certain circumstances) would be determined by a causal process, but this does not mean that the person's freedom to make choices is fundamentally different in some way from that which would pertain in the absence of the probabilities; a person cannot influence the probabilities implicit in the causal processes.[80]

It has been suggested (see Fischer, 2007, p 44-46) that if it was discovered that the laws of nature had 100 per cent probabilities associated with them rather than 99.9 per cent (say), that might be a finding of key importance to people's views about the existence of freedom and moral responsibility. But the understanding of probabilistic causal processes spelt out in this section suggests that it would make no difference whatsoever.

5.2.3 A lack of causal processes

Now suppose that either determinism or universal probabilistic causation operates for causal processes relating to most of the events that take place in the world, but that it does not operate for certain events involving mental states; an assumption reminiscent of Davidson's anomalous monism.[81] The causal processes would then leave some mental events unexplained: there would not even be a probability distribution underlying the occurrence of these mental states, there would be *no* explanation; science would not be complete (see Section 1.2.4). If this were so, and if it was also the case that mental states can be causally effective, so an agent's mental state involving a decision regarding action (an intentional mental state) could directly cause a physical event (the agent's action) to take place, there would be no identifiable prior cause of the mental state that resulted in the action – the person would be, in the words of Chisholm (1964, para. 11), "a prime mover unmoved" (Pereboom calls this an 'undetermined agent-cause', 2007, p 85). Key

[80] Robert Kane justifies the existence of his self-forming acts (SFAs) on the basis of "tension and uncertainty in our minds" (2007, p 26). He suggests that, when you take a decision in such circumstances, "(t)he indeterministic noise would have been an obstacle that you overcame by your effort" (p 27). On my view, Kane's SFAs could be about second order volition (Frankfurt, 1969), but cannot be about a person's will having an influence over the probability of an outcome in a law of nature because causal processes, *by definition*, are not "up to us" (see van Inwagen, 1983, p 16).

[81] Of course, it could be that determinism also does not always operate for the events that are not related to mental states, but this seems less probable and anyway would add nothing to our story.

aspects of mental states would then be outside science. A circumstance relating to the ontology of mind in which this situation might arise would be if consciousness (mental states) was a property of another substance, different from matter; that is if the 'dual substance' approach to the nature of mind were to pertain (Section 3.2).

Importantly, this case (called 'indeterminism' in this thesis, with some events 'undetermined') can have differing implications for our views regarding free will depending on the point in the causal chain at which the indeterminism arises. The implications for free will depend on whether the undetermined event takes place at the time the decision is being arrived at, simply having an effect on individual actions, or whether it affects the person's values – their dispositions to experience certain types of feelings in certain particular circumstances.

First, suppose that the individual actions initiated by a person are sometimes 'uncaused' (no causal process, not even a probabilistic one). In this case one might interpret such choices as being outside the control of the 'person', and hold that the person who initiated the action could not, therefore, be held to be *morally* responsible for it. This is the 'luck objection' to libertarianism, described in Mele (2006), and Pereboom (2007, p 102), which draws on the writings of Hume (1739/40, Book II, 3.1-3.2). So in this case, indeterminism of this second kind could be interpreted as implying that in these instances a person cannot be viewed as being morally responsible for his or her actions.

The more interesting situation, however, is that in which the person *is* viewed as being morally responsible for his or her actions, because the actions are determined by the person's values and it is the values themselves that have an element that is uncaused. This raises a question as to what it is that determines a person's values. I take it here that people arrive in this world with initial dispositions to experience certain types of feelings in certain sorts of circumstances, that it is these feelings that affect their choices. The values, the disposition to experience feelings, that a person has at any later point in time develop from these initial dispositions given the person's experiences in the course of his or her life. The question that then arises is: "what determines these initial dispositions?" There are two possibilities. A person's initial dispositions could be determined by the causal processes, possibly with some stochastic element, in a manner consistent with the one-substance, deterministic approach to science described in Section 1.2.2. If this is the case, the person's genes are likely to be at least partially responsible for the nature of their initial dispositions, although there could be probabilistic and environmental elements involved also. On the other hand, a person's initial dispositions may not be determined by a causal process, they

may simply pop up, uncaused, in the foetus when the person is *in utero*. An alternative to this would be if the person's values were subject to uncaused influences at some later point in their life. If either of these were so, there would be no cause that we can comprehend that determines the nature of the person's values. In this case the person's values would *not* be determined by some prior happening through a causal process, probabilistic or not, they would arise in a manner beyond our comprehension.

If a person's values were not determined by causal processes (perhaps with a probabilistic element involved) one might then feel able to argue that the person could be called a 'prime mover unmoved' in the sense of Chisholm (1964). In this case neither determinism nor universal probabilistic causation would not hold. This is therefore an example of the case known to philosophers as 'libertarianism' (see also Strawson, 1986, pp 35-36). An uncaused element to the values a person has seems to be the only way in which libertarianism can arise, since as shown above, if we introduce an uncaused element into the manner in which decisions are taken we find that we are not comfortable with asserting that the person is morally responsible for their actions.

5.2.4 Conclusion on determinism

The purpose of this chapter is to discover whether compatibilism is possible, given our causally effective property dualism approach to mind. It is clear from the above that if determinism or universal probabilistic causation operate, this would be consistent with the so-called 'one substance' approach to the ontology of mind, and therefore with property dualism. Libertarianism, with a person's initial values arriving 'uncaused', in a manner beyond our comprehension, carries with it the implication that science is not complete. Such a situation could therefore arise in a dual substance approach to the ontology of mind.

5.3 The meaning of 'free will'

In Section 4.5 we addressed the meaning of the term the 'will', and took a look at practical reasoning and the role of a person's values in determining the choices he or she makes. Here, after summarising the working of practical reasoning and looking in more depth at the role of the person's values, we examine what the adjective 'free' might mean in the context of the will.

5.3.1 The operation of the will

As we have seen in Chapter 4, the term 'the will', in philosophy, picks out a particular aspect of the mind. It describes the aspect of the mind that enables practical reasoning to be used so that a specific type of choice can take place: choice of intended action. This aspect of the mind is one of the fundamental attributes of a person – without a will a person would not be able to operate in the manner we take to be normal for people; he or she could survive only as the result of automatic reflexes pre-programmed before birth and the responses that might become programmed into the system as a result of events in the person's life.[82] Along with making these choices, the person must have the ability to implement the choice he or she has made regarding what to do next, and therefore it is implicit in the proposition that a person has a will, and can perform practical reasoning, that the information relating to the person's feelings, their preferences relating to the options available, can affect their actions, and therefore, that there are mental states that can be causally effective.

In Section 4.3.2, we saw that there are three separate functions associated with practical reasoning. First, the person must collect a whole lot of facts from his or her long-term memory (see Section 1.3.5) and form a view about what the likely outcomes of each possible action would be along with a rough estimate of how probable these outcomes are; these facts and probabilities could be called 'beliefs' and they are based on past experience. Next, the person must identify how he or she feels about these likely outcomes in the light of his or her values including his or her perceptions regarding duties. Finally, the sum total of the person's feelings about the bundle of expected outcomes of the actions, given his or her beliefs, must be compared, one against another, to identify the action deemed likely, on balance, to bring about the best, or at least an acceptable, outcome. The person's conclusion would then be something like, repeating Watson's words: "the thing for me to do in these circumstances, all things considered, is *a*" (1975, p 346). The person will initiate the chosen action at the appropriate time (demonstrating the, implicitly assumed, effectiveness of mental causation).[83]

---

[82] As noted in Section 2.3.2, the fact that Libet *et al.* (1982) have shown that some simple timing decisions can be taken on automatic pilot does not mean that all decisions are taken in this manner.
[83] In his 2007 paper, Fischer considers the case in which reductive physicalism holds and practical reasoning takes place but the action identified as best is, in fact, initiated by factors beyond the

There is then a question as to what aspect is it of a person that makes that person choose one action rather than another; what is it that makes a person's actions characteristic of that person.

5.3.2 The role of a person's values

We are assuming that, along with all the other attributes people have at any point in time, such as their physique, hair colour and eye colour, their musical, sporting and intellectual abilities and skills and so on, a person possesses three attributes, which can be separately identified, that determine their choices. First, we assume that a person has the ability to compare the 'motivational affect' associated with one option, the incentive effect associated with each available option that indicates its desirability, with that of another (see Section 4.3.2). Second, we assume that people have values. Values are the dispositions a person has to experience certain feelings in certain particular circumstances; and according to Watson (1975), as we have seen, they are: "the set of considerations" that cause a person to choose one course of action rather than another. When a person is considering whether one thing is in some way better or worse than another, more or less desirable, the way the person *feels* about those things is determined by the person's 'values'. A person's values therefore incorporate their desires as well as their moral nature. Third, we assume that people acquire a certain knowledge of facts and probabilities regarding the likely consequences of certain types of action during the course of their lives as the result of their experiences and that this knowledge about the working of the world enables them to make some sort of a prediction of the likely outcomes of their actions and to form a judgment about the probability of those outcomes occurring; for the purpose of this thesis I call this knowledge their 'beliefs'.

Clearly, the component of the will that is important in determining the characteristic choices of some person is not their ability to make a comparison of the 'motivational affect' associated with different options. If the ability to make such comparisons were swapped between two people, leaving each person's values and beliefs unchanged so the 'motivational affect' associated with each of the options remained exactly

control of the person concerned rather than by the person. In this case the proposition that the use of the will in performing practical reasoning is effective in determining action would, rather obviously, be nothing more than an innocuous fantasy; mental states would be epiphenomenal.

the same, there would no significant change in the choices the person made. It would be much the same as if one person's heart or kidney were swapped with another's.[84] It is therefore the person's values and their beliefs that are the factors determining the characteristic choices a person makes: what the person *believes* will be the likely consequences of their chosen action the way, and the way the person *feels* about those various possible consequences.

The feelings which we experience as a result of our values, which influence the judgment of a person who is choosing between alternative actions, come in a great range of shapes and sizes. They concern our subjective experiences and involve consciousness of sensory perceptions, feelings, thoughts and emotions. On a trivial level, I may dislike vinegar while you may love it, perhaps because my taste buds or their connections in my brain operate in a different way from yours so the 'what-it's-like' experience for each of us is different, or perhaps because of different experiences we have had in the past. Similarly, some key aspect of the feelings that we experience as the result of our value systems may differ: you may tend to give the effect of your actions on other people a greater weight than I, discounting them on average, relative to the effects on yourself, perhaps by only 10 per cent compared to my 50 per cent, say; or I might tend to leave effects that are likely to occur only in the far distant future out of the reckoning, while perhaps you generally take them into account, applying some appropriate internal discount rate. Our past histories may differ, so I may care deeply about the welfare of person X while this person is of no particular importance to you. Some people base their actions on what they deem to be their duty on appropriate occasions while others, perhaps, do not. Some of these important aspects of people's values may have their foundations in the genes they inherited, or they may have some source beyond our comprehension (see Section 5.2.3); but in both these cases the person's values are likely to have been heavily influenced by the person's experience, by all that has happened to the person during the course of his or her life. Clearly, all these things can be crucial in determining which action is chosen in some situation.

The possession of a will therefore enables a person to make a certain type of choice, the type of choice that requires practical reasoning because the alternative options include a number of different types of outcome (such as sensory pleasure, risk of pain or effects on other people the chooser may care about) which are not directly comparable

[84] See Bernard Williams (1970), where he postulates the switching of two people's brains.

one with another. The feelings associated with the likely outcomes therefore have to be evaluated thoughtfully and deliberated upon carefully if it is going to be possible to make a sound decision regarding which option appears, at the time the decision is taken, and to the person concerned, to be the one likely to lead to the best outcome.[85]

5.3.3 When is the will free?

With this background, we can address the question of what we mean by the term 'free will'. If a person has a 'will' he or she has the ability to perform the various functions we have identified that enable first the choice and then the initiation of a preferred action, and of course the will would not be considered to be free if either of these functions was constrained.[86] So I now identify the circumstances in which we might wish to say that a person was *not* free in his or her choice and/or initiation of an action.

Most obviously, the person does not have free will if he or she is unable, for some external reason, to perform the action chosen as the one to be preferred. A person can be in a situation in which their action selection function is able to operate freely but he or she is physically constrained from acting on a choice, as when a person's finger is forced upon the trigger of a gun to shoot another 'against their will' (as in Chisholm, 1964) or when a person is constrained unable to move when he or she wishes to help someone in distress. Hobbes (1651) saw this as the only constraint on what he called "free will", saying:

> "from the use of the words free will, no liberty can be inferred of the will, desire, or inclination, but the liberty of the man; which consisteth in this, that he finds no stop in doing what he has the will, desire, or inclination to do."

But nowadays we consider that freedom to perform a chosen action is only one aspect of freedom of the will. There are also situations in which one might argue that a person's *choice* (the identification of that which he or she "has the will, desire, or inclination" to do) is not free, perhaps because of interference from other people, perhaps because events have distorted some aspect of the person's choice mechanism. This may be because the person's values have been interfered with in some way, by other people or by events;

[85] It may seem tempting to identify some part of the will with the 'self' or to suggest that it is within the will that something that one might call the 'agent' component of a person resides (see Velleman, 1992, for example), but we have no need to do this in this thesis.

[86] See Borges' short story 'The Garden of Forking Paths' (Borges,1941), much cited in the recent literature on free will, see Kane (2007, p 6) and Fischer (2007, p 46)

the cases referred to as 'Frankfurt cases' (Frankfurt, 1969), and Pereboom's set of four cases involving Professor Plum (Pereboom, 2007, pp 94-97) are cases in point, as are hypnosis, drug-administration and addiction. Or it may be that the person's beliefs concerning facts and probabilities have been tampered with; as when someone is deliberately convinced by another of the validity of false information. Or, again, it may be that the ability to remember relevant facts has gone (as in dementia), so the data relating to the facts and probabilities needed for the weighing of alternative options are simply not available to the chooser or that the ability to experience the feelings that relate to values have been damaged by some illness or accident (as in the cases of Phineas Gage and others, see Damasio, 1994). In all these cases we might argue that a person's will is not 'free' in the sense required.

But here we are concerned with the question of whether it is *ever* possible for a person to have something that could be called 'free will', so we can leave aside those interesting cases in which the will might be said to be constrained or distorted or disabled in some way, and therefore not free, and in the rest of this chapter I take it that in the absence of any of these afflictions a person's will can be treated as being 'free' – in the sense that the person is free to use practical reasoning to choose and initiate an appropriate action from some set of available alternative actions, *given* that their values and their perception of the facts of the case have not been tampered with.

However, there is a question as to whether this is really the 'free will' that has so often been referred to by philosophers, that we associate with moral responsibility and which is sometimes described as being the situation in which: "the ultimate sources of our actions lie in us and not outside us in factors beyond our control" (Kane, 2007, p 5). So before taking a look at the so-called free will problem, we need to consider what it is we mean when we refer to 'moral responsibility'.

5.4 The meaning of 'moral responsibility'

In this section I first address the question of the meaning of 'responsibility' for the occurrence of an event, then I examine the meaning of 'moral responsibility'.

5.4.1 Responsibility

Where we have a causal process we have a cause and an effect. It is often said that the event that is viewed as being the cause was 'responsible' for the event that was the effect. The phrase "was the cause of" appears to be synonymous with "was responsible for" in this context. Thus the church clock was responsible for waking my guest, and the hurricane was responsible for blowing down your house. But causal processes come in forking chains running backwards in time to the Big Bang, so whatever it was that caused some particular event to take place was itself caused by some prior event or set of events. 'Responsibility' can thus be tossed around from the church clock to the vicar to the church committee to the maker of the church clock, and so on, forking into many branches along the way.

Clearly, there is never just one event that is responsible for something happening; there is always a forking chain of causes; if determinism (or universal probabilistic causation) holds the forking chain will go back to the beginning of time, the Big Bang. The meaning of the terms 'responsible for' and 'caused' are deceptive in their implications in this regard.

5.4.2 Moral responsibility

Now consider the question of how a person comes to be 'morally' responsible for something that happens rather than just being responsible for it as, one might say, some link in a set of forking chains of causes is responsible.

Neither the clock nor the hurricane would be charged with moral responsibility, it is going to be a 'person' that plays that role. And, as we have seen above, it is the person's values and beliefs that determine the choice a person makes, so it is the person's values, given their beliefs, that are the aspect of a person that makes him or her *morally* responsible for their actions. And this is so regardless of what has made those values what they are. It is irrelevant whether we think that the person's values are wholly determined by all that has gone before along with the laws of nature, as would be the case if compatibilism held (where the person's initial disposition to experience certain types of feelings is determined by the laws of nature, which may be probabilistic, and what has gone before), or whether we think of their values as being in part 'uncaused' as with libertarianism (where the initial disposition to experience certain types of feelings appeared out of the blue, with a cause beyond our understanding). It also suggests that these judgments regarding the moral

behaviour of others might remain the same even if a person's values are due to deliberate human intervention, as in Pereboom's Professor Plum cases (2007, pp 94-97), where neuroscientists and others create a person with obnoxious values, although here there may be moral opprobrium *also* for those concerned with creating a person with such values (see later, it is not necessarily the case that only one person is involved in initiating an action that results in moral approval or disapproval from others).[87]

With this approach, in effect, when we pass moral judgment on a person's behaviour what we are doing is praising or blaming the person for the perceived quality of the values that underlie their actions; we are saying: "I like your values and so I have much respect for you", or: "I dislike your values and so I have little respect for you". Whether we are justified in punishing a person who performs actions that we deem to be based on values we find obnoxious or rewarding a person of whose actions we approve is another question, and one that is beyond the scope of this thesis; here we go no further than the attempt to understand the nature of moral responsibility.

But if we accept that it is the person's *values* that are the issue here it raises the question of whether, when we praise or blame a person in this way for the actions that he or she has taken, we are praising or blaming the person's values *per se*, or the person, or whether the concept of a person necessarily encompasses their values, so the two come to the same thing. Is it necessarily the case that the person's values are included in the concept of a 'person'? Would it make sense to refer to a person, or use a personal pronoun, in a situation in which the thing referred to could have no values? I address this question, which turns out to be of considerable importance to the clarity of our thinking on these issues, in Section 5.6. First we must take a look at the Consequence Argument.

5.5 The Consequence Argument

Peter van Inwagen (1983) says, summarising succinctly an argument set down at greater length in his 1975 paper: "If determinism is true, then our acts are the consequences of the

[87] Fischer (2007) argues that so long as a person deliberates over some choice using practical reasoning (in his terminology, the person is 'reasons-responsive') we could say that a person has moral responsibility for their actions even though his or her actions are *actually* determined by external factors, because mental states are epiphenomenal. The moral responsibility is allocated because the action performed is the one the person has chosen. This is somewhat akin to the approach of Leibniz, noted in Section 3.2, where God has arranged things so that the mental and the physical act in harmony "without either body or soul disturbing the laws regarding the other".

laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us" (1983, p 16). This thought is known as the Consequence Argument, and had previously been spelled out by Carl Ginet (1966), among others. The implication is that if determinism, or universal probabilistic causation, holds, it follows as a simple matter of logic (so long as the consequences of things that are not up to us cannot be "up to us"), that our own acts are not "up to us"; and from that it follows that the proposition that we have moral responsibility for what we do is ruled out, so compatibilism is impossible. If the Consequence Argument is correct, not only is everything determined by what has gone before along with the laws of nature, but also our practical reasoning and decision-taking are totally irrelevant to those causal processes: future events are, quite simply, not "up to us". Many philosophers have found this argument persuasive (see Fischer, 2007 p 56, Kane, 2007 p 10, Pereboom, 2001, pp 33-37, and many others).

However, now suppose that one of the laws of nature referred to in the first two sentences of van Inwagen's Consequence Argument is about how a person takes a decision, based on the third of our fundamental propositions (see Sections 1.4.3 and 5.3.2). Such a law of nature might say, paraphrasing somewhat: "a person uses practical reasoning in order to decide what to do next and implements whatever action he or she decides should be chosen". If this law pertained, it would be the case that when we take a decision the consequences of that decision *are* "up to us": the consequences would be the direct result of a decision that is taken by us using practical reasoning and is then implemented by us. Given this law, these decisions would be part of the causal processes that determine the effect which the events in the past and present have on what happens in the future. But where a decision has been taken in this manner it is a *person* who has determined what should happen in the future in this regard, a *person* who has affected the future course of the world, using his or her own values and beliefs to make a difference to what happens in the world. There is a sense, therefore, in which these things we call 'persons' do, actually, influence the future course of the world in a manner that depends on their own values. In these circumstances, therefore, we *can* say, *correctly*, that the choice is "up to" the person, and we *can* say that there is a sense in which a person is free to choose; for the outcome of the choice depends on the particular person's values (indeed, a person with different values might well choose a different option). We *can*, therefore, hold a person responsible for the consequences of his or her actions.

There is no reason in logic why such a law should not operate, and so no reason why it should not be one of the laws of nature referred to in van Inwagen's summary statement of the Consequence Argument (above). Therefore, there must be something wrong with van Inwagen's argument.

5.6 The fundamental facts

Before proceeding, it may be helpful to summarise the situation we are now considering. The two fundamental facts that we take as given when addressing this issue raised by the Consequence Argument are the following:

> First, that determinism (or universal probabilistic causation) holds, so every event is determined by causal processes along with events in the remote past.
>
> Second, that a person takes decisions about his or her actions using practical reasoning, and the values and beliefs that determine the outcome of his or her decisions have, of course, been determined by causal processes and events in the remote past.

Our question is then:

> Given these facts, can a person be held to be morally responsible for the consequences of his or her actions (so long as circumstances such as those noted in Section 5.3.3 do not pertain)?

In the next section we see that if a person is to be held to be morally responsible for the consequences of his or her actions, a constraint has to be imposed on the meaning that can be attributed to the word 'person' (and to any personal pronouns or proper names that are used in its place). We also find that, given that this meaning is attributed to the term 'person', there is a question as to whether what is then implied for the meaning of the term 'free will' is acceptable.

5.7 The concept of a 'person'

We start by considering Chisholm's famous example of the man who was forced by another man to shoot someone (Chisholm, 1964, p 4). Chisholm says:

> "… if what we say he did was really something that was brought about by a second man, one who forced his hand upon the trigger, say, or who, by means of hypnosis, compelled him to perform the act, then since the act was caused by the *second* man it was nothing that was within the power of the *first* man to prevent. And precisely the same thing is true, I think, if instead of referring to a second man who compelled the first one, we speak instead of the *desires* and *beliefs* that the first man happens to have had. For if what we say he did was really something that was brought about by his own beliefs and desires, if these beliefs and desires in the particular situation in which he happened to have found himself caused him to do just what it was that we say he did do, then, since *they* caused it, *he* was unable to do anything other than just what it was that he did do. It makes no difference whether the cause of the deed was internal or external; if the cause was some state or event for which the man himself was not responsible, then he was not responsible for what we have been mistakenly been calling his act" (his italics).

In the first case, Chisholm's man was not responsible for shooting the person because he was forced to do it by the second man. In the second case, in just the same way according to Chisholm, the man was not responsible for what he did because he was forced to do it by his own beliefs and desires.

This throws light on the significance of the proviso about the Consequence Argument above: the fact that it is only true so long as the consequences of things that are not up to us are not "up to us". If determinism (or universal probabilistic causation) holds the man's beliefs and desires are not "up to him". And it is the man's beliefs and desires that caused his action. So is the man responsible for his action, or is he not? That depends on whether the word 'man', and the personal pronouns 'his' and 'he' connote the man's beliefs and desires, or whether they do not. If they *do*, the man is responsible for his actions because it is his beliefs and desires, which are part of *him*, that caused the action to occur (the consequences of things that are not up to us *can*, then, be "up to us"). If those pronouns do not connote the man's beliefs and desires, and in Chisholm's case they do not, for Chisholm defines "the man himself" as a thing without beliefs and desires, the man is not responsible for his actions.

Thus the "he" in the final clause of the Chisholm quote, which picks out the first man, has been explicitly stripped of the man's values (of his "beliefs and desires"). Chisholm has explicitly excluded the "they", his "beliefs and desires", from being implied by the "he". So whatever might be associated with the pronoun "he", as used on this occasion, there is certainly nothing there that could possibly be responsible for taking the decision to initiate a particular act. Chisholm's conclusion that something he calls "the

man himself" was not responsible rests on his implicit assumption that a person's values are *not* an integral part of what he means when he uses the words "the man himself".

We can now see that Van Inwagen's Consequence Argument would be valid only if the person's values (which, if determinism or universal probabilistic causation holds, are the result of the laws of nature and what went on before we were born) are *not* connoted by the word 'us' in its final sentence since, given his or her beliefs (determined by previous experience *etc.*), a person's 'values' are the aspect of the person that is responsible for the decision that he or she takes concerning what to do next. The attributes of a person which are connoted by the use of the word 'us' in van Inwagen's Consequence Argument therefore cannot include the person's values, if it is to hold. But if that is the case, the word 'us' in that context picks out a thing that is incapable of taking decisions in any manner other than randomly because it has no values.[88] So the Consequence Argument is then not of any interest.

In casual parlance, even though one may know nothing relating to the person referred to when a name is cited, it is automatically presumed that the person referred to has all the normal parts relating to a person, including the values and the background knowledge that we refer to as 'beliefs' which are necessary for the taking of sensible decisions. But occasionally, when the word person or a pronoun or proper name that refers to a person is used in the context of the free will issue, this is not the case.

The significance of this question about what is or is not associated with the concept of a person is nicely illustrated by Galen Strawson's impossibility theorem (1993, p 5). This states that: "… nothing can be truly morally responsible." Strawson justifies this result by pointing out that "Nothing can be *causa sui* – nothing can be the cause of itself", and then saying: "In order to be truly morally responsible for one's actions one would have to be *causa sui*, at least in certain crucial mental respects." Given this meaning for the phrase "truly morally responsible", Strawson is, of course, absolutely right. But my values are the part of *me* that determines my actions. So if I accept that the values that I happen to hold, whatever the reason for my holding them, are part of what I refer to as *me* (they are connoted by that pronoun) then it follows that *I* am morally responsible for my actions. I do not, then, have to be responsible for *causing* my values to be the way they are in order

[88] There seems to be an implicit assumption when people write about a person, that one could remove all of a person's attributes, which are all determined by causal processes and events in the remote past if determinism (or universal probabilistic causation) holds, and yet something fundamental about the person still remains; perhaps they view a person as having a 'soul'?

to be morally responsible for my actions, as Strawson suggests; all I have to do is accept that my values are an integral part of what I refer to as *me*; so that, although it is my values that determine what I do, it is *me* that is morally responsible for what I do: *I* am morally responsible for my actions so long as the personal pronoun connotes my values.

This question of whether a person's values are or are not viewed as being an integral part of the concept of a person, and therefore of a personal pronoun, is also the key to understanding the significance of van Inwagen's Consequence Argument. As we have seen, he summarises this saying (1983, p. 16):

> "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us".

But of course, if that final "us" connoted the person's values, albeit that they have been determined by what went on before we were born and what the laws of nature are, then so long as the person used practical reasoning to decide what to do next, the conclusion of the Consequence Argument, in the final sentence, would be wrong; the consequences of our own acts could then be said to be "up to" the thing we call 'us'. Given that it is 'our' values that determine our acts, for it not to be the case that 'we' are responsible for them requires that that final 'us' in van Inwagen's argument *excludes* the person's values (which in van Inwagen's case are, indeed, explicitly determined by "the laws of nature and events in the remote past"). So although what he says can be viewed as being true if the word 'us' has no connotations, it would then be of no relevance whatsoever to the free will problem because, as we have seen, if the personal pronoun has no connotations, the thing that is picked out by the personal pronoun is unable to make considered choices and cannot be considered to be responsible for the person's actions.

Another way of making a statement along the lines of that of van Inwagen, in his Consequence Argument, but incorporating the use of a pronoun that connotes the person's beliefs and values, would be to say:

> "If determinism is true, then what went on before we were born along with the laws of nature determine our values. And since our values are the part of us that determines our acts, it can be said that if determinism is true our acts are up to us".

So long as we define a 'person' as something that has a will, and therefore has values, the 'person' will be able to use practical reasoning when taking certain sorts of decision; so the 'person' is able to choose freely and the consequences of the person's acts can then be said to be "up to" the person. [89]

5.8 The free will problem

5.8.1 The apparent paradox at the heart of the free will problem

We have seen that the decisions people make using practical reasoning are part of the unfolding of the predetermined (perhaps probabilistically) path of the universe. But the person doing the choosing, taking the decision, will experience that familiar feeling of having the power, which freedom of the will engenders, to use his or her values to influence the future course of the world. However, if determinism (or universal probabilistic causation) holds, those values are actually the result of the laws of nature combined with things that happened before that person was born

It is this that leads to the seeming paradox at the heart of the free will problem. Given effective mental causation, we are correct in thinking that we can influence the future path of the world with our decisions. But the decisions that we take about what to do next are determined by our values and, if determinism (or universal probabilistic causation) holds, these in turn have been determined by the natural laws and what has gone before, so the future path of the world is, in fact, predetermined (perhaps probabilistically). Therefore, both of the statements:

"the future path of the world is predetermined"

and

"we influence the future path of the world with our decisions"

can be correct at the same time so long as the personal pronouns 'we' and 'our' are understood to connote a person's values and beliefs. As we have seen (Section 5.4.1), there

[89] Arthur Schopenhauer neatly summarised the fundamental aspects of this situation when he wrote: "You can do what you will, but in any given moment of your life you can *will* only one definite thing and absolutely nothing other than that one thing" (1841, see p 24; the precise wording depends on the translation).

is not just *one* cause of an event; responsibility for events comes in chains (with 'moral' responsibility only being involved when the cause of an event is informed by a person's values) and many important developments in the world have been, and are being, caused by people's choices regarding what to do next.

5.8.2 Can the will be said to be free?

We therefore find that, if we sign up to determinism (or universal probabilistic causation), although a person can have this thing we call 'free will' (in the sense of having freedom to choose) it has to be the case that the values that he or she has, that determine the outcome of any choice, have been determined by laws of nature. A question then arises as to whether the word 'free' is appropriate in such circumstances.

If you think of the 'person' as being constrained by values foisted upon him or her by laws of nature (as in Chisholm, 1964), then that 'person' is clearly not free at all. But if you combine the person's values, along with the person's beliefs and their ability to compare the 'motivational affect' associated with the available options, into something called the 'will', and think of that will as being an essential element of the *person*, the thing that is then called 'a person' can operate freely, given whatever constraints operate in the external world, taking the decisions regarding actions that he or she deems appropriate. The set of options is *truly* available and the person is then *truly* free to choose from amongst them, as is apparent from the fact that another person, with different values, might well choose a different option. In this case the 'person', the thing of which the will is an important part, *is* free to choose.

The key issue in deciding whether it can truly be said that a person has free will, therefore, is what exactly you mean by the word 'person', or the proper name or the pronoun that from time to time replaces that word. If the word 'person' connotes the will, and therefore the values, then the thing picked out by that word *can* be said to have 'free will', and *can* be held to be morally responsible for his or her actions. If the word 'person' does not connote the will, then the thing picked out by that word *cannot* be said to have 'free will', and *cannot* be held to be morally responsible for his or her (its?) actions.

5.8.3 Do we have 'ultimate moral responsibility' for our actions?

Responsibility comes in chains, as we have seen. One thing causes another which causes another and perhaps combines with other happenings to cause yet another. The ultimate

responsibility for the hurricane that blew your house down goes back in time to the Big Bang, to the beginning of time.

Moral responsibility however, as defined in Section 5.4.2, rests with persons. This means that one *could* allocate 'ultimate' moral responsibility to the last person whose decision was necessary for the event in question to take place. For example, Vasili Arkhipov, second-in-command of the Soviet Submarine B-59, who refused his Captain's order to launch nuclear torpedoes against US warships at the height of the Cuban Missile Crisis, could be deemed to be 'ultimately' morally responsible for saving the world as we know it. However, Henry VIII, who decreed that Anne Boleyn should be publicly executed in 1536, must surely pick up some moral responsibility for that event (the executioner presumably had committed himself in advance and felt he had no choice) albeit that there is the possibility that his personality changed somewhat after he hit his head in a jousting session, which may modify somewhat our feelings of opprobrium. And where Pereboom's Professor Plum cases (*op. cit.*) are concerned, one could argue that some moral responsibility needs to be transferred, at least partially if not wholly, from Professor Plum to those who created the moral monstrosity that he was designed to be.

We therefore conclude that a person could be deemed to be 'ultimately morally responsible' for some event that he or she has intentionally caused, even though the ultimate *causal* responsibility for the event, the thing that is ultimately responsible for its occurrence, must always be with the Big Bang if determinism (or universal probabilistic causation) holds.

5.8.4 Conclusion on the free will issue

Our conclusion on the free will issue is that, given that determinism (or universal probabilistic causation) operates and so long as the term 'person' carries the implication that the referent has a 'will' (and therefore has values), a person has freedom of choice (unless one of the situations described in Section 5.3.3 operates); and this freedom of choice is what is picked out by the term 'free will' if determinism (or universal probabilistic causation) holds. For some, this meaning for the term 'free will' may not be enough: it may not be acceptable.

However, *de facto*, in ordinary conversation, we hold a person who has freedom of choice to be what we call 'morally responsible', at least partially and often wholly 'morally responsible', for the consequences of the actions that he or she chooses to perform.

5.8 Causally effective property dualism and compatibilism

The purpose of this chapter has been to discover whether, if determinism (or universal probabilistic causation) holds, something that could be called free will is consistent with the causally effective property dualism approach to mind under examination in this thesis.

Suppose that it were to turn out that science is complete, so we are in a situation that can be described as deterministic (or subject to universal probabilistic causation, see Section 5.2.2), that people have feelings so there must another property of matter besides those that are studied in physics, and that mental states can be causally effective. With determinism (or universal probabilistic causation) holding, people's values would be determined by the laws of nature given all that has gone before. There is then a question as to whether anything that could be called free will can operate in such circumstances; is something that we might call 'compatibilism' possible?

Since the key feature that determines the outcome chosen in any choice situation is the person's values, it follows that so long as these are viewed as being an essential part of what we refer to when we use the word 'person', or use a pronoun or a proper name, the person referred to can choose freely within the constraints of the external situation; and therefore there is a sense in which a 'person' so defined *does* have something that one might call 'free will'. Given that practical reasoning is used for taking certain types of decision, it would be the person's values, and therefore the 'person', that takes the decisions which lead to the unfolding of the future course of the world.

Compatibilism (free will combined with determinism) is therefore a viable position in the case of causally effective property dualism. It can rightly be said that 'we' have freedom to choose, and therefore free will, subject only to the requirement that our values are viewed as being part of 'us'. And the fact that, typically, we hold a person morally responsible for the consequences of their actions suggests that, typically, we *do* take the term 'person' and the personal pronouns and proper names to connote the person's values.

Importantly, however, it is also true that if compatibilism holds (and substance dualism does not), causally effective property dualism must be the case. Because it is the aspect of a person that we call the person's 'values' that determines the way the person *feels* about the likely consequences of choosing some option, and feelings are not part of physics (they cannot be communicated perfectly in words, and so are outside our definition of the physical sciences) it follows that if compatibilism holds these mental states, that are associated with the taking of decisions, cannot be the type of thing that is suitable for study

in physics. So compatibilism implies that causally effective property dualism must be the case.

Conclusion

The reasoning that leads to the assumption about the nature of mind investigated in this thesis is quite straightforward.[90]

First, in order to decide whether or not it could be the case that mental states can be totally explained in terms of the properties of matter studied in the physical sciences we need a clear definition of physics. I have defined physics, and the physical sciences, in terms of their most renowned and essential feature: the repeatability of experimental findings. Any fact accepted as true in the physical sciences will be supported by experiments, and it must be possible for those experiments to be repeated by others, with the same results found and recorded. This means that all the facts relating to the properties of matter that are studied in the physical sciences must be public facts, facts capable of being communicated in language, perfectly, by one person to another.

Second, it is not possible to communicate perfectly, in language, to another person what it feels like to experience something that the other person has never experienced. Typically, in philosophy, the truth of this has been presented in terms of Mary (Jackson, 1982), who had never seen the colour red, learning something new when she eventually sees it in spite of the fact that she already knew everything that can be known to the physical sciences relating to the colour red and to seeing. But it would also have to hold for a person who lacked the physical necessities for seeing red: someone who had no eyes could learn from others what it is like to see if what it is like to see could be explained in terms of the vocabulary of physics, and therefore could be communicated *perfectly* in words; we could all learn what it is like to experience magneto-reception, even though we have neither the sense organs nor the brain structures necessary just so long as someone used the correct words in describing it. If it is accepted that communicating these things, perfectly, in words is impossible then, given that mental states are related to matter, rather than being instantiations of a property of another substance, it could be that they are instantiations of another property of matter, one that is not suitable for study in physics. And if that were so, another property of matter would exist besides those studied in physics. However, an alternative hypothesis, supported by many in the academic

[90] See also Section 1.5.

philosophy literature, would be that such an explanatory gap could exist alongside ontological reduction, where mental states are, in fact, nothing more than physical states. The dual property approach is the possibility that is investigated in this thesis.

Third, I suggest that what it feels like to experience something is information that is essential to the taking of decisions regarding one's own welfare and the welfare of others. That this is generally *thought* to be so is apparent from the analysis of choice by economists (cost-benefit analysis), by psychologists (decision theory) and by philosophers (practical reasoning). It is also implicit in evolutionary considerations: actions that are consistent with enhancing the likelihood of survival are typically triggered by feelings of hunger, cold, avoidance of pain, and so on. It is also supported by direct observation of what happens to a person's ability to take sensible decisions when the particular part of the brain which is involved in action selection is taken out by accident, illness or for medical reasons but all the person's cognitive abilities remain, enabling them to identify the available actions, predict the actions' outcomes and identify the effects of those outcomes on the people concerned. But for this information about feelings to influence actions it has to be the case that mental states can be causally effective.

Combining these three propositions – that what it is like to experience something cannot be communicated perfectly to another person in words, that that suggests that there *may* be another property of matter besides those studied in physics, and that if there were, this other property of matter must be capable of causing physical events to take place – we can see that there is a strong case for considering that 'causally effective property dualism' (the proposition that there is another causally effective property of matter besides those studied in physics) may describe the ontology that underlies consciousness.

An important implication of this is that, if it turns out to be correct, there must be some way in which how a person feels about something can influence physics; physics cannot then be causally complete. Maybe one day we will come to understand the manner by which such a process can operate.

References

Armstrong, David (1968), *A Materialist Theory of Mind*, London: Routledge.

Armstrong, David (1989), *Universals: An Opinionated Introduction*, Westview Press.

Arndt, M., O. Nairz, J. Voss-Andreae, C. Keller, G. van der Zouw, and A. Zeilinger (1999), 'Wave-particle duality of C60', *Nature*, **401**, pp 680-682.

Ashcroft, F. (2012), *The Spark of Life*, London: Penguin Books Ltd.

Atkinson, R. and R. Shiffrin (1968), 'Human memory: a proposed system and its control processes', in K. W. Spence and J. T. Spence (eds.), *The Psychology of Learning and Motivation* vol. 2, New York: Academic Press.

Baars, Bernard (1988), *A Cognitive Theory of Consciousness*, Cambridge: Cambridge University Press.

Baars, Bernard (1997), 'In the theatre of consciousness, Global Workspace Theory, a rigorous scientific theory of consciousness', *Journal of Consciousness Studies*, **4** (4), pp 292-309.

Baars, Bernard (2002), 'The conscious access hypothesis: origins and recent evidence', *Trends in Cognitive Sciences*, **6** (1), pp 47-52.

Baddeley, Alan (1996), 'Exploring the central executive', *The Quarterly Journal of Experimental Psychology*, **49A** (1), pp 5-28.

Baddeley, Alan (1997), *Human Memory: Theory and Practice*, Hove and New York: Psychology Press.

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

Baddeley, Alan and Graham Hitch (1974), 'Working memory', in G. A. Bower (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*, 8, New York: Academic Press, pp 47-89.

Baker, Lynne Rudder (1993), 'Metaphysics and mental causation', in *Mental Causation*, John Heil and Alfred Mele (eds.), Oxford: Clarendon Press, pp 75-95.

Ball, Derek (2009), 'There are no phenomenal concepts', *Mind*, **118** (472), pp 935-962.

Beckermann, Ansgar (1992), 'Introduction – reductive and nonreductive physicalism', in Beckermann, A., H. Flohr and J. Kim (eds.), *Emergence or Reduction? - Essays on the Prospects of Nonreductive Physicalism,* Berlin/New York: Walter de Gruyter, pp 1-21.

Bennett, Jonathan (2004), 'A New System of the nature and communication of substances, and also of the union that exists between the soul and the body G. W. Leibniz', *http://www.earlymoderntexts.com/assets/pdfs/leibniz1695c.pdf*.

Bennett, Jonathan (2009), 'Correspondence between Descartes and Princess Elisabeth', *http://www.earlymoderntexts.com/assets/pdfs/descartes1643_1.pdf*.

Berkeley, George (1710), *A Treatise Concerning the Principles of Human Knowledge*, *http://www.earlymoderntexts.com/assets/pdfs/berkeley1710.pdf*.

Bernard, C. (1865), *Lectures on the Phenomena Common to Animals and Plants*, trans. H. E. Hoff, R. Guillemin & L. Guillemin, 1974, Springfield (IL): Charles C Thomas.

Bernard Theos (1947), *The Hindu Philosophy*, New York: The Philosophical Library, pp 69-72.

Berns, Gregory S., Andrew M. Brooks, Mark Spivak (2015), 'Scent of the familiar: An fMRI study of canine brain responses to familiar and unfamiliar human and dog odors', *Behavioural Processes,* **110**, pp 37-46.

Block, Ned (1978), 'Troubles with Functionalism', in W Savage (ed.), *Perception and Cognition: Minnesota Studies in the Philosophy of and Science IX*, University of Minnesota Press.

Block, Ned (1980), *Readings in Philosophy of Psychology*, Cambridge: Harvard University Press.

Block, Ned (1995), 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18** (2), pp 227-287.

Block, Ned (2001), 'Paradox and cross-purposes in recent work on consciousness', *Cognition*, **79**, pp 197-219.

Block, Ned (2006), 'Max Black's objection to mind-body identity', in D. Zimmerman (ed.), *Oxford Studies in Metaphysics, Volume 2*, Oxford: Oxford University Press.

Bloom, Paul (2004), *Descartes' Baby*, New York: Basic Books

Borges, Jorge Luis (1941), *The Garden of Forking Path*s, London: Penguin Books.

Broad, C. D. (1925), *Mind and its Place in Nature*, London: Routledge and Kegan Paul Ltd.

Burge, Tyler (1993), 'Mind-body causation and explanatory practice', in John Heil and Alfred Mele (eds.), *Mental Causation*, Oxford: Clarendon Press, pp 97-120.

Burnham, William H. (1888), 'Memory, historically and experimentally considered', *American Journal of Psychology*, **2**, pp 568-622.

Carruthers, P. (2000), *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge: Cambridge University Press.

Cartwright, Nancy (1983), *How the Laws of Physics Lie*, Oxford University Press.

Chalmers, David J. (1995), 'Facing up to the problem of consciousness', *Journal of Consciousness Studies,* **2**, pp 200-219.

Chalmers, David (1996), *The Conscious Mind*, Oxford: Oxford University Press.

Chalmers, David (2002), 'Consciousness and its place in nature', in Stephen P. Stich & Ted A. Warfield (eds.), *Blackwell Guide to the Philosophy of Mind*, Blackwell, pp 102-142.

Chalmers, David (2007), 'Phenomenal concepts and the explanatory gap', in Torin Alter & Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, New York: Oxford University Press, pp 167-194.

Chisholm, Roderick M. (1964), 'Human Freedom and the Self', The University of Kansas Lindley Lecture, in G. Watson (ed.), *Free Will*, Oxford: Oxford University Press, pp 26-37.

Churchland, Paul (1981), 'Eliminative materialism and propositional attitudes', *The Journal of Philosophy*, **78**, pp 67-90.

Claparède, É. (1911), 'Recognition et Moite', *Archives de Psychologie Geneve*, **11**, p 79.

Cowan, Nelson (1997), *Attention and Memory*, Oxford University Press: Oxford.

Crane, Tim (1995), 'The mental causation debate', *Proceedings of Aristotelian Society*, Supplementary Volume 69, pp 211-236.

Crane, Tim (2001), *Elements of Mind*, Oxford: Oxford University Press.

Crane, Tim and Mellor, D. H. (1990), 'There is no question of physicalism', *Mind*, **99**, pp 185-206.

Crick, Francis (1994), *The Astonishing Hypothesis*, Simon & Schuster Ltd.

Damasio, Antonio R. (1994), *Descartes' Error*, New York: G. P. Putnam's Sons.

Davidson, Donald (1970), 'Mental events', in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory,* pp 207-222, Amherst, Mass., reprinted in Davidson, D., *Essays on Actions and Events*, 1980, Clarendon Press: Oxford.

Davidson, Donald (1993), 'Thinking causes', in John Heil and Alfred Mele (eds.), *Mental Causation*, Oxford University Press, pp 3-17.

Dehaene, Stanislas and Lionel Naccache (2001), 'Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework', *Cognition*, **79**, pp 1-37.

Delgado, J. M. R., W. W. Roberts and N. E. Miller (1954), 'Learning motivated by electrical stimulation of the brain', *American Journal of Physiology*, **179**, pp 587-593.

Delgado, M. R., V. A. Stenger and J. A. Fietz (2004), 'Motivation-dependent responses in the human caudate nucleus', *Cerebral Cortex*, **14**, pp 1022-1030.

Dennett, Daniel (1978), *Brainstorms*, Montgomery, VT: Bradford.

Dennett, Daniel (1988), 'Quining Qualia', in Marcel, A and E. Bisiach (eds.), *Consciousness in Contemporary Science*, pp 42–77, New York: Oxford University Press.

Dennett, Daniel (1991), *Consciousness Explained*, Little, Brown & Company.

Descartes, René (1637), *Discourse on the Method*, trans. Laurence J. Lafleur 1960, New York: The Liberal Arts Press.

Descartes, René (1641), *Meditations on First Philosophy,* Paris: Michael Soly.

Dijkaterhuis, A., and L. F. Nordgren (2005), 'A Theory of Unconscious Thought', *Perspectives on Psychological Science*, **1**, pp 95-109.

Eagle, Antony (Winter 2016 Edition), "Chance versus Randomness", *The Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.), forthcoming
URL = <http://plato.stanford.edu/archives/win2016/entries/chance-randomness/>.

Eddington, A. (1928), *The Nature of the Physical World*, New York: Macmillan.

Efron, Robert (1970), 'The minimum duration of a perception', *Neuropsychologia*, **8**, pp 57-63.

Feigl, H. (1958), 'The "Mental" and the "Physical"', in Herbert Feigl, Michael Scriven, and Grover Maxwell (eds.), *Minnesota Studies in the Philosophy of Science: Concepts, Theories, and the Mind-Body Problem, Volume II*, Minneapolis: University of Minnesota Press.

Feyerabend, P. (1963), 'Materialism and the Mind-Body Problem', *Review of Metaphysics*, **17**, pp 49-66.

Fischer, John Martin (2007), 'Compatibilism', in, John Martin Fischer, Robert Kane, Derk Pereboom and Manuel Vargas, *Four Views on Free Will*, Blackwell Publishing Ltd.

Flanagan, Owen and Thomas Polger (1995), 'Zombies and the function of consciousness', *Journal of Consciousness Studies*, **2**, pp 313–321.

Foot, Philippa (1967), 'The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices', *Oxford Review*, **5**, pp 5-15.

Frankfurt, Harry G. (1969), 'Alternate possibilities and moral responsibility', *The Journal of Philosophy*, pp 829-39, in Watson, G. (ed.), *Free Will*, 2003, Oxford: Oxford University Press pp 167-176.

Frege, Gottlob (1892), 'On sense and reference', *Zeitschrift fur Philosophie und philosophische Kritk*, **100**, pp 25-50, (translated 1960).

Galton, Francis (1883), *Inquiries into Human Faculty and its Development, Part 4*, London: Dent.

Gardner, Thomas (2005), 'Supervenience physicalism: meeting the demands of determination and explanation', *Philosophical Papers*, **34**, pp 189-208.

Geach, Petyer and Max Black (1960), *Translations from the Philosophical Writings of Gottlob Frege*, New York; Oxford.

Graham, G. (2016), 'Behaviorism', *The Stanford Encyclopedia of Philosophy*, E N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2016/entries/behaviorism/, (Fall 2016 Edition).

Ghysen, A. (2003). 'The origin and evolution of the nervous system', *International Journal of Development Biology*, **47**, pp 555-562.

Ginet, Carl (1966), 'Might we have no choice?', in Keith Lehrer (ed.), *Freedom and Determinism*, New York: Random House.

Gould, S. J., & Lewontin R. C. (1979), 'The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme', *Proceedings of the Royal Society B*, **205**, pp 581-598.

Grau, Carles, Romuald Ginhoux, Alejandro Riera, Thanh Lam Nguyen, Hubert Chauvat, Michel Berg, Julià L. Amengual, Alvaro Pascual-Leone, Giulio Ruffini (2014), 'Conscious brain-to-brain communication is humans using non-invasive technologies', *Plos One,* **9**, pp 1-6.

Greene, J. (2001), 'An fMRI investigation of emotional engagement in moral judgment', *Science*, **293**, pp 2105-2108.

Hameroff, Stuart and Roger Penrose (1996), 'Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness', *Mathematics and Computers in Simulation*, **40**, pp 453-480.

Harnad, Stevan (2000), 'Turing indistinguishability and the blind watchmaker', in Fetzer, J. & Mulhauser, G. (eds.), *Evolving Consciousness*, John Benjamins.

Hempel, C. G. (1949), 'The logical analysis of psychology', in H. Feigl and W. Sellars (eds.) *Readings in Philosophical Analysis,* New York: Appleton-Century-Crofts.

Hobbes, Thomas (1651), 'Of the liberty of subjects', Chapter XXI in *Leviathan.*

Hodgson, David (1991), *'The Mind Matters'*, Oxford: Oxford University Press.

Hoefer, Carl (Spring 2016 Edition),"Causal Determinism", *The Stanford Encyclopedia of Philosophy,* Zalta (ed.), *http://plato.stanford.edu/archives/spr2016/entries/determinism-causal.*

Honderich, T (1982), 'The argument for anomalous monism', *Analysis*, **42** (1), pp 59-64.

Horgan, Terence (1984), 'Jackson on physical information and *qualia*', *Philosophical Quarterly*, **34**, pp 147-152.

Horgan, Terence (1993) 'From supervenience to superdupervenience: meeting the demands of a material world', *Mind*, **102**, pp 555-586.

Hume, David (1739/1740), *A Treatise of Human Nature*, ed. Ernest C. Mossner, reprinted Penguin Classics, 1985.

Humphrey, N. (2006), *Seeing Red*, The Belknap Press of Harvard University Press, London.

Humphrey, N. (2008), *The Mind Made Flesh*, Oxford: Oxford University Press.

Jackendoff, Ray (1987), *Consciousness and the Computational Mind*, MA: Bradford Books, MIT Press.

Jackson, Frank (1982), 'Epiphenomenal Qualia', *Philosophical Quarterly*, **32**, (82), pp 127-136. Reprinted in W. G. Lycan and J. J. Prinz, (eds.), *Mind and Cognition: An Anthology*, third edition (2008), pp 657-663.

Jackson, Frank (1986), 'What Mary didn't know', *The Journal of Philosophy*, **83** (5), pp 291-295.

James, William (1890), *Principles of Psychology*, New York: Henry Holt.

Jammer, Max (1974), *The Philosophy of Quantum Mechanics*, New York: Wiley.

Johannson, Lars-Goran (2006) 'Natural necessity', *Uppsala Philosophical Studies*, **53**, 231-246.

Johnston, Mark (1987), 'Human Beings', *The Journal of Philosophy*, **84**, 2, pp 59-83.

Jönsson, C. (1974), 'Electron diffraction at multiple slits', *American Journal of Physics*, **4**, pp 4-11.

Kahneman, D. and A. Tversky (1979), 'Prospect theory: An analysis of decisions under risk', *Econometrica,* **47**, pp 263–291.

Kahneman, D. (2011), *Thinking Fast and Slow*, London, Penguin Group.

Kandel, Eric, R., M. Klein, B. Hochner, M. Shuster, S. Siegelbaum, R. Hawkins, D. Glanzman, V. F. Castellucci, and T. Abrams (1987), 'Synaptic modulation and learning: new insights into synaptic transmission from the study of behaviour', in G. M. Edelman, W. E. Gall and W. Ml Cowan (eds.), *Synaptic Function*, pp 471-518, New York: John Wiley.

Kandel, Eric R. (2001), 'The molecular biology of memory storage: a dialogue between genes and synapses', *Science's Compass*, pp 1030-1038.

Kandel, Eric R. (2006), *In Search of Memory*, W. W. Norton and Company Inc., New York.

Kane, Robert (2007), 'Libertarianism', in John Martin Fischer, Robert Kane, Derk Pereboom and Manuel Vargas, *Four Views on Free Will*, Blackwell Publishing Ltd.

Kim, Jaegwon (1984), 'Epiphenomenal and supervenient causation', *Midwest Studies in Philosophy*, **9**, pp 257-270, reprinted in David M. Rosenthal, *The Nature of Mind*, (1991), OUP: Oxford.

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

Kirk, Robert (1974), 'Zombies v. materialists', *Proceedings of the Aristotelian Society*, Supplementary Volume **48**, pp 135-152.

Knutson, B., and S. M. Greer (2008), 'Anticipatory affect: neural correlates and consequences for choice', *Philosophical Transactions of the Royal Society B*, **363**, pp 3771-3786.

Koch, Christof (2012), *Consciousness: Confessions of a Romantic Reductionist*, Cambridge MA: The MIT Press.

Kornorski, J. (1948), *Conditioned Reflexes and Neuron Organisation*, Cambridge: Cambridge University Press.

Kremers, Dorothee, Juliana López Marulanda, Martine Hausberger, Alban Lemasson, (2014), 'Behavioural evidence of magnetoreception in dolphins: detection of experimental magnetic fields', *Naturwissenschaften*, **101** (11), pp 907-911.

Kripke, Saul (1972), *Naming and Necessity*, Blackwell, 1981 (first published 1972).

Leibniz, Wilhelm Gottfried (1695), *A New System of the Nature and Communication of Substances, and also of the Union that Exists Between the Mind and the Body*, translated by Jonathan Bennett, 2004, www.earlymoderntexts.com.

Levine, Joseph, (1983), 'Conceivability identity and the explanatory gap', *Towards a Science of Consciousness*, **3**, pp 3-13.

Levine, Joseph, (1983), 'Materialism and qualia: the explanatory gap', *Pacific Philosophical Quarterly*, **64**, pp 356-361.

Levine, Joseph (1993), 'On leaving out what it's like', in Martin Davies & Glyn W. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*, Blackwell.

Levine, Joseph (2001), *Purple Haze*, Oxford: Oxford University Press.

Levine, Joseph (2007), 'Phenomenal concepts and the materialist constraint', in Torin Alter & Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge*, pp 145-166, New York: Oxford University Press.

Lewis, David (1966), 'An argument for the identity theory', *Journal of Philosophy*, **63**, pp 17-25.

Lewis, David (1988), 'What experience teaches', *Proceedings of the Russellian Society*, Sydney, Australia: University of Sidney. Reprinted in D. J. Chalmers, 2002, *Philosophy of Mind: Classical and Contemporary Readings,* pp 281-294, Oxford: Oxford University Press.

Libet, B., E. Wright & C. Gleason, (1982), 'Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts', *Electroencephalography and Clinical Neurophysiology*, **54**, pp 322-335.

Libet, B. (1985), 'Unconscious cerebral initiative and the role of conscious will in voluntary action', *The Behavioural and Brain Sciences,* **8**, pp 529-566.

Loar, Brian (1970), 'Phenomenal States', in J. Tomberlin (ed.), *Philosophical Perspectives*, vol. 4; repr. in a revised form in N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness,* Cambridge, Mass.: MIT Press, 1997.

Locke, John (1671), *Essay Concerning Human Understanding*, Draft B, see Peter H Nidditch and G. A. J. Rogers (eds.), *Drafts for the 'Essay Concerning Human Understanding' and Other Philosophical Writings*, 1990.

London, F. and Bauer, E., (1939), 'The theory of observation in quantum mechanics', in J. A. Wheeler and W. H. Zurek, *Quantum Theory and Measurement*, (eds.), 1983, Princeton: Princeton University Press, pp 217-259.

Louie, K. and M. A. Wilson (2001), 'Temporally structured REM sleep replay of awake hippocampal ensemble activity', *Neuron,* **29**, pp 145-156.

Luce, R Duncan and Howard Raiffa (1957), *Games and Decisions*, John Wiley and Sons, Inc.

Mace, C. A., (1948-49), 'Some implications of analytic behaviourism', *Proceedings of the Aristotelian Society*, **49**, pp1-16.

Margolis, Eric and Laurence, Stephen (2014), "Concepts", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.).

Markowitz, H. M. (1952), "Portfolio Selection", *The Journal of Finance,* **7** (1), 77–91.

Marsden, C. D. (1981), 'The mysterious motor function of the basal ganglia: the Robert Wartenberg Lecture', *Neurology*. 1982, **32** (5), pp 514-39.

Marsh, Henry (2014), *Do No Harm*, Weidenfeld and Nicholson, London.

Massaro, Dominic W (1970), 'Perceptual auditory images', *Journal of Experimental Psychology*, **85** (3), pp 411-417.

McGinn, Colin, (1982), *The Character of Mind*, Oxford University Press: Oxford.

McGinn, Colin, (1989), 'Can we solve the mind-body problem?', *Mind*, **98** (391), pp 349-366.

McLaughlin, Brian F (2005), 'A priori versus a posteriori physicalism', *Philosophy-Science-Scientific Philosophy, Main Lectures and Colloquia of GAP,* **5**, pp 267-285.

McLaughlin, Brian F (2010), 'Consciousness, type physicalism, and inference to the best explanation', *Philosophical Issues*, **20**, pp 266-304.

Medina, L. and Reiner, A. (1995), 'Neurotransmitter organization and connectivity of the basal ganglia in vertebrates: implications for evolution of basal ganglia', *Brain Behaviour and Evolution*, **46**, pp 235–258.

Mele, Alfred, R (2006), *Free Will and Luck*, Oxford: Oxford University Press.

Melchert, Norman (1986), 'What's wrong with anomalous monism?', *The Journal of Philosophy*, pp 265-274.

Melnyck, Andrew (2008), 'Can physicalism be non-reductive?', *Philosophy Compass*, **3** (6), pp 1281-1296.

Merickle, Philip M (1980), 'Selection form visual persistence by perceptual groups and category membership', *Journal of Experimental Psychology: General*, **109** (3), pp 279-95.

Miller, George A (1956), 'The magical number seven, plus or minus two: some limits on our capacity for processing information', *Psychological Review*, **63**, pp 81-97.

Milner, Brenda, Larry R Squire, and Eric R Kandel (1998), 'Cognitive neuroscience and the study of memory, review', *Neuron*, **20**, pp 445-468.

Mink, Jonathan W (1996), 'The basal ganglia: focused selection and inhibition of competing motor programs', *Progress in Neurobiology*, **50**, pp 381-425.

Morsella, Ezequial (2005), 'The function of phenomenal states: supramodular interaction theory', *Psychological Review*, **112** (4), pp 1000-1021.

Nagel, E (1979), *The Structure of Science: Problems in the Logic of Scientific Explanation*, Indianapolis: Hackett.

Nagel, Thomas (1974), 'What is it like to be a bat?', *The Philosophical Review*, **88**, pp 435-50.

Neath, Ian and Aimée M. Surprenant (2003), *Human Memory,* Thomson: Wadsworth.

Nemirow, Laurence (1980), 'Mortal Questions by Thomas Nagel', *The Philosophical Review*, **89**, pp 475-477.

Neumann, J. von (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton; German original *Die mathematischen Gundlagen der Quantenmechanik*, Berlin: Springer, 1932.

Nichols, Shaun & Todd Grantham (2000), 'Adaptive Complexity and Phenomenal Consciousness', *Philosophy of Science*, **67,** pp 648-670.

Noë, Alva and Evan Thomson (2004), 'Are there neural correlates of consciousness?', *Journal of Consciousness Studies*, **11**, pp 3-28.

Nordby, Knut (2007), 'What is this thing you call color: can a totally color-blind person know about color?', in Torin Alter & Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge,* pp 77-86, New York: Oxford University Press.

O'Connor, Timothy (1995), 'Agent causation', in *Agents, Causes and Events: Essays on Indeterminism and Free Will*, (ed.) Timothy O'Connor, Oxford: Oxford University Press, pp 173-200.

O'Donohue, William and Richard Kitchener (1998), *Handbook of Behaviourism*, Academic Press.

Olds, J., & Milner, P (1954), 'Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain', *Journal of Comparative and Physiological Psychology*, **47** (6), pp 419-427.

Panksepp, J (2005), 'Affective consciousness: core emotional feelings in animals and humans', *Consciousness and Cognition*, **14**, pp 30-80.

Panksepp, J (2011), 'The basic emotional circuits of mammalian brains: do animals have affective lives?' *Neuroscience and Biobehavioural Reviews*, **35**, pp 1791-1804.

Panksepp, J., & Biven, S. (2012), *The Archaeology of Mind*, New York: W. W. Norton & Company Inc.

Papez, J. W (1937), 'A proposed mechanism of emotion', *Archives of Neurology and Psychiatry*, **38** (4), pp 725-743.

Papineau, David (1990), 'Why supervenience?', *Analysis*, **50**, pp 66-71.

Papineau, David (2002), *Thinking About Consciousness*, Oxford: Clarendon Press.

Pascal, Blaise (1660), *Pensées*, translated by W. F. Trotter, Gutenberg.

Paulus, Martin, P (2007), 'Decision-making dysfunctions in psychiatry – altered homeostatic processing?', *Science*, **318**, pp 602-608.

Peacocke, Christopher (2007), 'Mental action and self-awareness (1)', in Brian P. McCloughlan and Jonathan Cohen, (eds.), *Contemporary Debates in Philosophy of Mind*, Blackwell Publishing, pp. 358-374.

Penrose, Roger (1989), *The Emperor's New Mind*, Oxford: Oxford University Press.

Pereboom, Derk (2001), *Living Without Free Will*, Cambridge: Cambridge University Press.

Pereboom, Derk (2007), 'Hard Incompatibilism', in John Martin Fischer, Robert Kane, Derk Pereboom and Manuel Vargas, *Four Views on Free Will*, Blackwell Publishing Ltd.

Perelmuter, Zeev (2010), 'Nous and Two Kinds of Epistêmê in Aristotle's Posterior Analytics', *Phronesis*, **55** (3), pp 228-254.

Perry, Clint, J., Luigi Baciadonna and Lars Chittka (2016), 'Unexpected rewards induce dopamine-dependent positive emotion-like state changes in bumblebees', *Science*, 353 (6307), pp 1529-1531.

Place, U. T (1956), 'Is consciousness a brain process?', *British Journal of Psychology*, pp 44-50.

Potter, Beatrix (1902), *The Tale of Peter Rabbit,* Frederick Warne & Co.

Prinz, Jesse J (2003), 'A neurofunctional theory of consciousness', in A Brook and K Atkins (eds.), *Philosophy and Neuroscience*, Cambridge: Cambridge University Press.

Prinz, Jesse J (2007), 'All consciousness is perceptual', in Brian P McCloughlan and Jonathan Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, Blackwell Publishing, pp 336-357.

Puccetti, Roland (1977), 'The great C-fiber myth: a critical note', *Philosophy of Science,* **44**, (2), pp 303-305.

Putnam, Hilary (1963), 'Brains and behaviour', in Ronald J. Butler (ed.), *Analytical Philosophy: Second Series*, Blackwell.

Putnam, Hilary (1967), 'Psychological predicates', in Capitan and Merill, Pittsburgh (eds.), *Art, Mind and Religion*, (reprinted with the revised title 'The Nature of Mental States' in *Readings in Philosophy and Psychology*, ed. Block, 1980, Methuen).

Quine, W. V. O (1951), 'Two Dogmas of Empiricism', *The Philosophical Review,* 60, pp 20-43. Reprinted in Quine, W. V. O, *From a Logical Point of View* 1953, Harvard University Press.

Quine, W. V. O (1960), *Word and Object*, Cambridge, Mass.: MIT Press (p 265)

Quine, W. V. O (1985), 'States of Mind', *The Journal of Philosophy*, **87**, pp 5-8.

Quine, W. V. O (1987), *Quiddities*, Cambridge, Massachusetts: The Belknap Press.

Redgrave, P., T. J. Prescott and K. Gurney (1999), 'The basal ganglia: a vertebrate solution to the selection problem?', *Neuroscience*, **89** (4), pp 1009-1023.

Redgrave, P., T. J Prescott, and K. Gurney (2001), 'A computational model of action selection in the basal ganglia', *Biological Cybernetics*, **84**, pp 411-423.

Rees, Geraint, Gabriel Kreiman and Christof Koch (2002), 'Neural correlates of consciousness in humans', *Nature Reviews: Neuroscience*, **3**, pp 261-70.

Revonsuo, A and Newman, J (1999), 'Binding and consciousness', *Consciousness and Cognition,* **8**, pp 123-127.

Rey, G (1983), "A Reason for Doubting the Existence of Consciousness", in R. Davidson, G. Schwartz and D. Shapiro (eds.), *Consciousness and Self-Regulation* Vol 3. New York: Plenum, pp 1-39.

Richet, Charles R (1886), 'Les origines et les modalités de la memoire', *Revue Philosophique*, XXI, p 570.

De Roberts E M and Sasai Y (1996), 'A common plan for dorsoventral patterning in Bilateria', *Nature*, **380**, pp 37-40.

Rorty, Richard (1965), 'Mind-body Identity, Privacy, and Categories, *The Review of Metaphysics*, **19** (1), pp 24-54.

Rosenthal, David M (1986), 'Two concepts of consciousness', *Philosophical Studies*, **49** (3), pp 329-359.

Rosenthal, David M (2008), 'Consciousness and its function', *Neuropsychologia* **46**, pp 829-840.

Rowlatt, Penelope A (2009), 'Consciousness and memory', *Journal of Consciousness Studies*, **16** (5,) pp 68-78.

Ryle, Gilbert (1949), *The Concept of Mind*, London: Hutchinson.

Sakitt, Barbara (1976), 'Iconic memory', *Psychological Review*, **83** (4), pp 257-276.

Schneider, Susan (2013), 'Non-reductive physicalism and the mind problem', *Nous*, **47** (1), pp 135-153.

Schopenhauer, Arthur (1841), *On the Freedom of the Will,* New York: Dover Publications Inc. Originally *Uber die Freiheit des Menschlichen Willens*, 1841, Frankfurt am Main: Joh. Christ. Hermannsche Buckhandlung.

Scoville, William Beecher and Brenda Milner (1957), 'Loss of recent memory after bilateral hippocampal lesions', *Journal of Neuropsychiatry and Clinical Neurosciences*, **20**, pp 11-21.

Searle, John R (1983), 'Why I am not a property dualist', reprinted in *Journal of Consciousness Studies*, 2002, **9** (12), pp 57-64.

Searle, John R.(1991), *The Mystery of Consciousness*, Granta Publications.

Segner, Johann Andreas (1740), *Specimen Logicae Universaliter Demonstratae*, ed. Mirella Capozzi, Bologna: CLUEB, 1990.

Sellars, W (1956), 'Empiricism and the Philosophy of Mind', in Feigl H. and Scriven M. (eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis: Minnesota Studies in the Philosophy of Science*, Vol. 1. Minneapolis: University of Minnesota Press, pp 253–329.

Shanta, Bhakti Niskama (2015), 'Life and consciousness – The Vedāntic view', http://www.tandfonline.com/doi/full/10.1080/19420889.2015.1085138, *Communicative & Integrative Biology*, (retrieved 10/8/2016).

Skinner, B. F (1953), *About Behaviorism*, New York: Macmillan.

Slovic, P (1987), 'Perception of risk', *Science,* **236**, pp 280-285.

Smart, J. J. C (1959), 'Sensations and brain processes', *The Philosophical Review*, **68** (2), pp 141-156.

Smith, L (1986), *Behaviourism and Logical Positivism: a Reassessment of their Alliance*, Stanford CA: Stanford University Press.

Snowdon, Paul (2003), 'Knowing How and Knowing That: A Distinction Reconsidered', *Proceedings of the Aristotelian Society*, pp 1-29.

Soon C. S., Brass M., Heinze H. J., & Haynes J. D (2008), 'Unconscious determinants of free decision in the human brain', *Nature Neuroscience*, **11**, pp 543-545.

Sperling, George (1960), 'The information available in brief visual presentations', *Psychological Monographs*, **74**, pp 1-30.

Spinoza, Benedictus de (1677), *The Ethics*, Penguin Classic.

Spurrett, David & David Papineau (1999), 'A note on the completeness of 'physics'', *Analysis*, **59** (1), pp 25-29.

Stanley, Jason & Timothy Williamson (2001), 'Knowing how', *Journal of Philosophy,* **98** (8), pp 411-444.

Strausfeld, Nicholas J., and Frank Hirth (2013), 'Deep Homology of Arthropod Central Complex and Vertebrate Basal Ganglia', *Science*, **340**, pp 157-161.

Strawson, Galen (1993), 'The impossibility of moral responsibility', *Philosophical Studies*, **75**, pp 5-24.

Strawson, Galen (1994), *Mental Reality*, London: The MIT Press.

Strawson, Galen (2006), 'Realistic monism', *Journal of Consciousness Studies*, **13**, pp 3-31.

Strawson, Galen (2008), *Real Materialism*, Oxford: Oxford University Press.

Suppe, Frederick (2000), 'Definitions', in *A Companion to the Philosophy of Science*, ed. W. H. Newton-Smith, Maldon: Blackwell.

Sutton, John (Summer 2016 edition), 'Memory', *The Stanford Encyclopedia of Philosophy*, URL = http://plato.stanford.edu/archives/sum2016/entries/memory/, ed. Edward N. Zalta.

Tegmark, Max (2000), 'Importance of quantum decoherence in brain processes', *Physical Review E,* **61**, pp 4194-4206.

Tomer, R., Denes, A., Tessmar-Raible, K., & Arendt, D (2010), 'Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium', *Cell*, **142**, pp 800-809.

Tulving, Endel (1972), 'Episodic and semantic memory', in E Tulving and W Donaldson (eds.), *Organisation of Memory,* New York: Academic Press.

Turvey, M. T (1973), 'On peripheral and central processes in vision: inferences from an information-processing analysis of masking with patterned stimuli', *Psychological Review*, **80** (1), pp 1-52.

Thomson, Judith Jarvis (1976), 'Killing, Letting Die, and the Trolley Problem', *The Monist*, **59**, pp 204-217.

Tye, Michael (2009), *Consciousness Revisited: Materialism Without Phenomenal Concepts*, MIT: MIT Press.

Van Gulick, Robert (1985), 'Physicalism and the subjectivity of the mental', *Philosophical Topics*, **XIII**, (3), pp 51-70.

Van Horn, John Darre,l Andrei Irimia, Carinna M. Torgerson, Micah C. Chambers, Ron Kikinis, Arthur W. Toga (2012), 'Mapping Connectivity Damage in the Case of Phineas Gage', *Plos One, http://dx.doi.org/10.1371/journal.pone.0037454*.

Van Inwagen, Peter (1975), 'The incompatibility of free will and determinism', *Philosophical Studies*, **27**, pp 185-199.

Van Inwagen, Peter (1983), *An Essay on Free Will*, Oxford: Oxford University Press.

Velleman, J. David (1992), 'What happens when someone acts?', *Mind*, **101**, pp 461-481.

Von Neumann, J. & Morgenstern, O (1947), *Theory of Games and Economic Behaviour*, 2nd edition, Princeton N. J.: Princeton University Press.

Von Neumann, J (1955), *Mathematical Foundations of Quantum Mechanics*, translated from the German edition by Robert T. Beyer Princeton U.P.; Oxford U.P, 1955.

Watson, Gary (1975), 'Free agency', in *Free Will*, ed. Gary Watson, 2003, Oxford: Oxford University Press pp 337-351.

Watson, Gary (2003), *Free Will*, Oxford: Oxford University Press.

Watson, J (1913), 'Psychology as a behaviorist views it,' *Psychological Review*, **20**, pp 158-177.

Wigner, Eugene, O (1961), 'Remarks on the mind-body question', in I. J. Good, (ed.), *Basic Books*, New York: Heinemann.

Williams, Bernard (1970), 'The Self and the Future', *Philosophical Review*, **79**, pp 161-180.

Wilson, Jessica (2005), 'Supervenience-based formulations of physicalism', *Nous*, **39**, pp 426-459.

Wisdom, John (1952), *Other Minds*, Oxford: Basil Blackwood.

Wittgenstein, Ludwig (1953), *Philosophical Investigations*, trans. G. E. M. Anscombe. New York: Macmillan.

Woollett, Katherine and Eleanor A. Maguire (2011), 'Acquiring "the Knowledge" of London's Layout Drives Structural Brain Changes', *Current Biology*, **21** (24), pp 2109–2114.

Yablo, Stephen (1992), 'Mental causation', *The Philosophical Review*, 101 (2), pp 245-280.

Yoo, Julie (2016), 'Mental causation', *(http://www.iep.utm.edu/mental-c), The Internet Encyclopedia of Philosophy* (retrieved 25/9/2016).

Zajonc, R. B (1980), 'Feeling and thinking: Preferences need no inferences', *American Psychologist*, **35** (2), 151-175.