# INTRINSIC NATURALISM:
# A TYPE-F MONIST ACCOUNT OF PHENOMENAL CONSCIOUSNESS

By

LUKE ALEXANDER GORDON PALMER

A thesis submitted to
The University of Birmingham
for the degree of
Master of Philosophy

Department of Philosophy
College of Arts and Law
The University of Birmingham
September 2010

**ABSTRACT**

The aim of this thesis is to provide a theory of phenomenal consciousness which accords with both the science-friendly spirit of physicalism and the acknowledgement of panpsychism that phenomenal properties may be inextricably linked to entities, but with none of the problems associated with either type of model. Initially, physicalism and panpsychism are evaluated by the lights of their most serious problems, and solutions are offered to these problems from the point of view of a third kind of model: intrinsic naturalism, presented in the final chapter. This model holds consciousness to be among the battery of a functional system's intrinsic (i.e. non-dispositional) properties. A definition is given, and defence made for the existence of these properties, and their compatibility with an otherwise physicalist ontology.

## ACKNOWLEDGEMENTS

Especial thanks to my supervisor, Dr Yujin Nagasawa, for his many patient re-readings of this work, his advice and his guidance.

Further thanks to Amy King, Graeme Brodie, and Izzy Sanders for their moral support during the writing of this thesis.

My deepest gratitude to Dr Augustus Newland, whose unique insight into methods of enquiry brought fresh perspective to this dissertation, and without whose support and friendship the words would have stopped flowing.

Finally, my respect and thanks to Prof David Chalmers, for taking the problem seriously.

## TABLE OF CONTENTS



## LIST OF FIGURES

# INTRODUCTION

In this thesis, I shall attempt to develop a solution to the problem of consciousness: why is it that we have conscious experiences, and how do they fit into the physical world? Dualistic or panpsychist theories seem insufficient, as do traditional physicalist ones. I believe that a type-F monist theory of some sort will provide the solution. My focus will not be on the mind-body problem itself, but specifically on what sort of type-F monist solution to it would be most fitting.

Since we inhabit a world where all phenomena thus far observed are reducible to purely physical terms, it is desirable to have a theory of consciousness which is compatible with this, if not itself physicalist. To that end, it is my intention to develop a panprotopsychist model of consciousness comprising aspects of a number of other models; in particular neutral monism. I shall show that panprotopsychism is less vulnerable to certain flaws than physicalist and panpsychist models of consciousness.

The structure of this dissertation will be as follows. In the first chapter, I shall introduce the problem more thoroughly, and argue that physicalism seems initially unsatisfying. I shall describe the dominant positions. I shall show that while a number of phenomena previously believed to be non-physical have since been shown to be reducible to physical principles, consciousness remains the exception.

In the second chapter, I shall examine panpsychist theories, and the problems entailed by them. I shall discuss the problems shared between panpsychism and panprotopsychism, and argue that most of these problems can be solved far

more easily from a panprotopsychist point of view than a panpsychist one. I shall argue in particular that consciousness can only reasonably be considered a property of functional systems.

In the third chapter, I shall discuss the problems with traditional models of physicalism, and show how a type-F monist position can avoid these problems. I shall discuss the difference between physical and what I call 'intrinsic' properties, and argue that a theory which appeals to intrinsic properties in order to explain consciousness is not a physicalist one in the standard sense, though it may be considered to demonstrate a physicalistic attitude.

In the final chapter, I shall examine a theory of my own, called 'intrinsic naturalism', which I believe both offers useful ways of thinking about the problem, and remains compatible with an otherwise physicalist ontology. Here, I shall specify the kind of functional system which I believe can possess consciousness as a property (and argue for the qualifications which make something such a system), and finally I shall attempt to anticipate potential worries which the theory may engender.

# CHAPTER ONE

# CONSCIOUSNESS IN A PHYSICAL WORLD

## The problem of consciousness

The term 'consciousness' has many definitions, including wakefulness, awareness, access to one's own cognitive states, and the possession of cognitive and emotional faculties. Throughout this dissertation, however, I shall use the term to refer to phenomenal consciousness: the intrinsic quality of experience. Experience consists of qualia; these are elements of experience such as the sensation of seeing red, the taste of salt, the smell of roses, the feeling of embarrassment and so forth. The term can even encapsulate broader concepts such as 'what it is like to see a friend'; there are potentially infinite kinds of qualia. In fact, because it is impossible to communicate the nature of qualia to others, classifying or quantifying them is a fruitless activity.

Because phenomenal consciousness is not described in terms of its causal role (there are very few coherent views about what this role might be, although many people have an intuition that it must have one), it seems to invite further explanation on top of the explanations given for various cognitive functions. If someone were to say, "You have explained how neurons arrange themselves to adapt to new information, but you have not explained learning", the response would be a simple semantic one, that learning is a concept encapsulated entirely by the explanation of that functional process. However, even when such functional processes have been described, the question of what it is like subjectively to experience such a process remains pertinent. This implies that consciousness may not come about by the same physical and functional processes as cognition, emotion and so forth, and that its nature may be

separate from the otherwise physicalist ontology which modern science has largely adopted. In simpler terms, the mystery at hand is how the obviously phenomenal can be emergent from what we can reasonably presume to be non-phenomenal matter[1].

It does seem as though phenomenal facts are supervenient upon physical facts (which is to say that two people could not be identical physically without sharing the same experience), however the precise nature of this relationship must be qualified. Chalmers (1996) distinguishes between global and local supervenience. There does not seem to be any reason to suggest that two physically identical individuals within this world would be phenomenally distinct; that would entail a counterintuitive lack of correlation between neural states and their accompanying qualia. It therefore seems the case that phenomenal facts supervene upon physical facts locally (if at all), rather than globally. That is to say that as long as the two physically identical individuals are within the same world, they should be phenomenally identical. Chalmers also distinguishes between logical and natural supervenience, and it is the application of this distinction to the problem of consciousness which generates the greatest controversy between physicalists and non-physicalists. Physicalists, of course, argue that phenomenal facts are logically (that is, necessarily in all worlds) supervenient upon physical facts; one cannot have two worlds identical with regards to their physical facts but that differ in regard to their phenomenal (or indeed any) facts. Chalmers, and many non-physicalists, argue that in fact this supervenience is only natural (in other words it is contingent, and only happens to be the case in some worlds such as our own). It happens to be that in this world, phenomenal facts are (likely to be) fixed by physical facts, due to whatever relationship exists between them. It is,

[1] This is not always presumed to be the case. See the section on Panpsychism later in this chapter.

however, conceivable that there is a world physically identical to ours which lacks phenomenal properties. It is, naturally, arguable that such a conception may be caused by a misunderstanding of what physical identity entails, however this seems unlikely. We have a good understanding of what it means to say that 'water is physically identical to $H_2O$' and as such the notion of having two worlds which are $H_2O$-identical but not water-identical is not conceivable (as long as one knows everything physical about water).

The problem of consciousness is a problem precisely because qualia do not have characteristics similar to any other physical phenomenon (most conspicuously, they are only available first-person; an external observer simply does not have access to the phenomenal properties of another individual), nor do they seem to fit easily into a physical world (being difficult to fit into a causal picture of the world). I shall argue that although phenomenal facts cannot be reduced to physical facts, as can the cognitive processes which they accompany, they are not completely independent from the physical world.

I am not going to argue against physicalism *per se,* but I believe that it has problems at its core which can be solved through non-physicalist models. First, it will be instructive to examine some historical examples of phenomena which were thought to be non- or extra-physical. Later, I shall compare the non-physical consciousness arguments with these historical positions.

**Historical challenges to the supremacy of the physical**

A number of phenomena have been suspected, as consciousness is now, to be the result of non-physical processes or properties. Indeed, several physicalists often point to these now debunked theories, meaning to infer that non-

physicalism about consciousness will soon be such a theory. I believe, however, that there is a very good reason that consciousness has remained the last natural phenomenon to have its physicality called into serious question. Before I address my reasons for believing this, however, it will be instructive to explain why some of these other theories have since been made redundant by physicalism.

From ancient times, the philosophy of vitalism was a popular way of thinking about biological life. The two tenets of vitalism are: that the functions of an organism cannot be reduced to biology alone, and must evoke some sort of inherently 'living' property; and that the processes within all living things cannot be explained in terms of physics and chemistry alone, because life is a fundamental property of the organism. Of course, this view is no longer held because the mechanisms of life have been elucidated increasingly since the invention of the microscope. It is now a matter of scientific orthodoxy that the cellular metabolic mechanisms that constitute life are entirely explicable through chemical, and ultimately physical, principles. In short, there is no need to resort to the invocation of a fundamental non-physical property of 'life', the so-called *élan vital*.

A similar theory emerged from the era when alchemy was evolving into chemistry, and is known as 'phlogiston theory'. Before scientists had learned how to detect the presence of gases, it was not known how processes such as rusting and combustion occurred. A massless, colourless, and otherwise undetectable substance called phlogiston was claimed to exist in the late seventeenth century, and was said to be the substance contained within all materials, released upon combustion. Now, chemists understand the chemical processes involved in

oxidation, and the energy transfer between molecules. A combination of these factors entirely accounts for combustion without the need for a pseudo-physical factor the only 'evidence' for whose existence was combustion itself.

More recently, scientists are coming to show that the very origin of the Universe and the life therein can be explained purely in terms of physical, self-organising matter without invoking an intelligent creator being. Human intelligence, personality and emotion are all increasingly explicable through biological and ultimately physical terms without recourse to concepts such as a soul. Of course, Descartes' famous substance dualism becomes progressively less relevant as our understanding of neuropsychology increases, and the need for a separate non-physical source of thought and emotion decreases. These are but a few examples of how science has produced a physicalist worldview. I shall argue that consciousness is, perhaps uniquely, an exception to the rule that all observable phenomena can be explained with physical principles. Furthermore, unlike the above-mentioned phenomena, consciousness may indeed require some sort of non-physical property to be added to the inventory of the Universe's properties.

**Why consciousness is almost certainly non-physical**

One of the reasons that vitalism is no longer a position held by scientists is that it no longer serves to explain anything. In other words, it was only ever subscribed to because the 'vital spirit' was thought to be the only phenomenon which could explain the functions of life. Now that scientists have near-complete models of metabolism and reproduction, nothing remains to be explained. Occam's razor stipulates that one should avoid the introduction of entities into an explanation unless they are strictly necessary, and the *élan vital* is not necessary to explain the functions of life once one understands cellular mechanisms.

Phenomenal consciousness serves no function. The only causal role it might have is in giving us the knowledge that we possess it, and this is potentially problematic. There are ways to show that we can know we are conscious without admitting consciousness to be causal, however. Chalmers (2004) and Brogaard (2010) both argue that phenomenal experience has intentional content, such that in having the experiences we thereby come to know we are having them. Chalmers argues, further, that neither phenomenal nor intentional properties can be asserted to be more fundamental than the other. If phenomenal experiences inherently contain the concepts by which I, as the experiencer, come to recognise them (as I do when I say that I know I am seeing red), they do not need to be causal in such a way as to threaten a position which requires consciousness not to be causal.

Proponents of the conceivability argument would suggest that even my zombie twin (that is to say, someone who is physically identical to me but who lacks phenomenal consciousness) would consider himself conscious and as such would debate the matter with other zombies. Other than that tenuous claim to causal function, there is no other strongly coherent sense in which qualia have a causal effect on us. Even pain, which seems a particularly vivid example of an evolutionary *purpose* for phenomenal consciousness, has a neuropsychological aspect to it which *entirely explains* our behaviour when in pain, and has presents no need to invoke qualia. I shall return to this example in my third chapter about physicalism. For now, let me simply state that one can build a complete causal model of a human being's entire repertoire of behaviour without needing consciousness. As Chalmers (1996, 5) puts it, "…consciousness is *surprising."*

It is difficult to prove the point, but when I respond in some way to a pain sensation, there is no obvious reason why the entire sequence of events from receipt of stimulus to production of response is not entirely within the realm of the biomechanics of my nervous system. There does not seem to be any need for pain to be phenomenally painful; indeed, as Chalmers (1996) argues, my zombie twin would still think himself to be feeling pain, and if we were to extract from a living organism enough of a nervous system to detect pain, we would not suppose that this array of nervous tissue had any qualia. Let us also recall the evidence I discussed in the panpsychism chapter, concerning my brain's control over my body; these outputs are not accompanied by phenomenal feels, yet there is a distinctly obvious causal mechanism in place.

I shall employ a parallel to Chalmers' vanishing qualia argument to make my point more plainly, and perhaps in a way which will be more persuasive to physicalists. Let us imagine a simple system wherein a man, S, lifts a weight by means of a pulley. Once S has pulled on the cord, the explanation for the elevation of the weight is purely mechanical; unless we adopt panpsychism, we have no motivation to add consciousness to this picture (note my emphasis on the moment *after* S has pulled the cord; S himself may have consciousness, but it is not relevant here). Suppose we increase the system's complexity so that there are a few more weighs and pulleys. Again, once S has applied force to the cord, there does not seem to be any necessity to add consciousness to the mechanism. Now we shall increase the complexity yet more; S now stands on a platform which is on wheels and attached to the mechanism. When he pulls the cord, the system of ropes and weights is arranged such that S is withdrawn sharply from his initial position, and now cannot reach the activation cord. Here we have a rather simple model for a reflexive action; S, playing the part of a

painful stimulus, is activating this primitive block-and-tackle nervous system, and being repulsed by the system. Does the system need to be conscious? Let us finally increase the complexity of the system so that it is a full functional analogue of a basic reflex arc. At what point has the system, or any part of it, required consciousness in order to withdraw the stimulus from its initial position? I do not think there is any need for such a property to be added to the system, because mechanical principles are just as adequate to explain the chain of events now as they were when we had nothing more than a single weight and pulley. I am not contradicting my previous assertion that such a functional system could be conscious; indeed it still could. Yet that consciousness would have no *explanatory power*. We could even imagine the construction of a reflex arc from the actual biological components, from its most crude instantiation to the swift and well-evolved system most animals have. At no point would anything other than biochemical processes (themselves a kind of mechanical process) be required to explain reflexes. Why, then, should we think qualia to be causally efficacious? There seems to be no strong support for such a notion, and as I argued in the first chapter, we would not think it necessary to consider consciousness at all were it not for each of us having first-person experiences of it.

If one accepts the above argument, one must either deny consciousness outright (with all of the problems of *a priori* physicalism) or accept that consciousness itself is non-causal. If the physical is causally closed, and a phenomenon exists which is causally inefficacious, then it seems that there are properties of physical systems which are themselves non-physical. My final theory will acknowledge that consciousness is not causal, and in fact *requires* that this be so.

If phenomenal consciousness is non-causal, why should we not discount any notion of there being phenomenal consciousness, just as we have the *élan vital*? It is, after all, not a phenomenon which is observable to science, nor one which is required in order to explain anything, and seems to be precisely the kind of superfluous entity which Occam's razor would bid us to exclude. However, there is one piece of evidence which causes us to keep our belief in consciousness: *our own consciousness.* Each one of us has (certainly I have, and I presume at least all humans have) first-hand and hardly disputable evidence of the existence of qualia in our own case, which it is impossible for us to show to anyone else. I could be a zombie, writing about properties which I do not possess. I know I am not, but there is no way for me to convince someone of this; we take it as a matter of intuition that since others have a psychology similar to ours, it is reasonable to presume that they have phenomenology similar to ours.

A number of questions present themselves at this point: questions about the existence, origin and extent of consciousness. What concerns us here is the origin of consciousness; it seems reasonable to suggest that our phenomenology is tied to our physical properties, but the particular relationship seems to have conceptual problems. For instance, it seems very strange that something physical should be unobservable in principle from the outside. There is nothing subjective in standard physics; everything has an externally observable causal impact on everything else. Also, human brains are obviously the sort of physical system which can produce consciousness, but since we have no way to know for sure that even other brains are conscious, how can we begin to find out if other systems or objects are conscious? I believe that there are a few educated guesses which can be made in this area, but guesses may be all that we can make.

Ultimately, any theory of consciousness will have to make assumptions which cannot in principle be corroborated. If I devise an emergentist theory of consciousness[2], I am assuming that fundamental entities are not conscious; if I argue for a panpsychist approach, I assume that fundamental entities possess a property which I cannot possibly confirm them to have. I believe, however, that one is justified in believing something not to have a certain property when it exhibits none of the detectable properties which are usually correlated with it. In this case, demonstrating cognitive properties is a sign that there may be a phenomenal aspect to an entity. Furthermore, since panpsychist theories add fundamental consciousness to the ontology of the world, it seems that the burden of proof lies with the panpsychist.

Seeking to reconcile the existence of a seemingly non-physical phenomenon with an otherwise physicalist ontology, I shall look for a solution that is physicalist in spirit if not in letter. In short, I believe that it is desirable that any physicalist who believes consciousness to exist (as most do) should have no principled problem with my theory.

Although physicalist and panpsychist theories will be examined in their own respective chapters, it will be informative to give a brief introduction to them now, and how they may be related to a third class of theory, of which mine will be a kind: type-F monism.

[2] For a brief explanation of emergentism, see the section on panpsychism later in this chapter.

**Physicalism**

I shall now outline the dominant positions concerning the problem of consciousness. In their own chapters, I shall examine and evaluate them in detail, showing how the problems associated with them may be applied to other theories of consciousness. Of a great number of positions, the four which concern me are: physicalism; panpsychism; panprotopsychism; and to a lesser extent, dualism. Each of these positions is in fact a category enshrining a number of models, and to an extent there are theories which merge with two or three categories. My ultimate aim is to synthesise such a model; one which is in keeping with our otherwise justified physicalist perspective, but which acknowledges the non-physicality of consciousness for which I shall continue to argue.

Physicalism is predominantly divisible into two strains; *a priori* physicalism and *a posteriori* physicalism (called type A and B materialism respectively by Chalmers (2002)). *A priori* physicalists deny that there is an epistemic gap between physical and phenomenal facts, or believe that if there is a gap then it can be closed easily. In simple terms, the type-A solution to the problem of consciousness is the claim that there is no problem, because there is a self-evident entailment from the physical to the phenomenal. It has several forms, such as eliminativism and behaviourism, which generally either claim that there is no such phenomenon as consciousness, or equate it with behavioural or functional states. The relevant difference between the two physicalist paradigms examined here is that *a priori* physicalists deny even the explanatory gap; once the physical explanations of awareness, access, the ability to report cognitive states and so on are sufficiently detailed, an explanation of what we call consciousness will be present therein.

Immediately, a parallel seems apparent between those who claim consciousness to be something beyond the physical processes of the brain, and a vitalist. Indeed, physicalists often use this analogy. However, as Chalmers (*ibid*.) notes, the analogy does not hold; vitalists claimed that the functions which constitute life would not be explained without recourse to an intrinsically vital property. However, once they had been explained, there was nothing further which required an explanation. For an *a priori* physicalist to argue that the analogy holds, he may have to claim that there is no *explanandum*, which would be extremely counter-intuitive; arguably our consciousness is the only phenomenon of whose existence we can be certain. To do so would be exemplary of an eliminativist position, which is not representative of all type-A theories.

*A posteriori* physicalists admit that there is an epistemic gap between physical and phenomenal facts. However, they then deny that there is an ontological gap. This version of physicalism is (unless stated otherwise) the version which I shall be discussing in future when I use the term 'physicalism', because models within this family are held to be solutions to a hard problem, rather than a denial that the problem exists. A dominant class of theory within this category is the class of 'identity' theories, where consciousness is held to be identical to the functional states of the brain. Importantly, the identity is discovered empirically, in the same way that water was discovered to be $H_2O$ empirically. The two are acknowledged to be conceptually different, but knowing all of the physical facts about water entails knowing its identity with $H_2O$. I shall revisit this argument in my third chapter, with reference to Jackson's knowledge argument.

Another *a posteriori* tactic, rather than making analogies such as the one above, is to assert there to be a brute, necessary link between the physical and the phenomenal. There are various ways to do this, and several arguments fall under the category of the 'phenomenal concept' strategy. What these arguments have in common is that they posit an hypothesis such that we possess certain psychological features, and that these features explain why we perceive an explanatory gap between the physical and the phenomenal. These features are argued to be physical themselves, thus giving us an entirely physical explanation of why there is an explanatory gap. There are several versions of this strategy, and several ways to respond to each. Chalmers (2007) argues that in general, such positions lead to one of two outcomes: either we are compelled to adopt an acquaintance model of how we come to have an explanatory gap (the problems with which I shall examine in my third chapter); or they leave us with insufficiently strong epistemic relations to explain why we perceive an explanatory gap in the first place. I shall examine the epistemic and ontological gaps in my third chapter, wherein I endorse the conceivability-possibility link.

**Dualism**

Although not one of the primary topics of this thesis, I shall spend some time discussing whether my final theory is dualistic. For my purposes, I shall call a model 'dualistic' if it involves a (one- or two-way) causal link between a physical and non-physical domain, where these domains are distinct. I would class sepiphenomenalism as dualistic in this regard, as well as the obvious example of Cartesian substance dualism. Such models are of no use in solving the problem of consciousness because they include an unnecessary causal role for consciousness, and still come no closer to solving the problem of why there should be anything it is like to be in a particular cognitive state.

Even a dualistic theory such as Jackson's epiphenomenalism, which is only dualistic with regards to qualia (as opposed to those which are dualistic with regards to cognitive states), has problems. The causal closure of the physical is a notion that there is no good reason to violate, and epiphenomenalist theories have physical systems imparting causality to a phenomenal realm which cannot in turn cause anything in the physical domain. If one accepts, and again I must make it clear that I do, the debatable position (given in, e.g. Fair, 1979) that causation is identical to the transfer of energy or momentum, the laws of thermodynamics are also under threat from such a theory.

**Panpsychism**

The dichotomy between *a priori* physicalism and complete Cartesian dualism presents only one dimension of the problem of consciousness. Another axis forms its poles at emergentism and panpsychism.

Standard physicalism, epiphenomenalism, panprotopsychism and so forth are all emergentist; the most simple entities in the universe (such as the fundamental particles) are taken to be non-conscious, whereas human brains are taken to be conscious. At some point between these two entities, consciousness emerges, and the puzzle which must be solved is how this emergence occurs. The theories to be discussed shortly claim that even the fundamental entities in the Universe (whatever these may eventually be confirmed to be) possess mental states to some degree. These theories fall within the category of 'panpsychist', however there is a great deal of difference in the definitions of 'mental', 'universe' and 'fundamental' between many of these models.

Broadly speaking, panpsychist theories vary between claiming ubiquitous mentality with only rare consciousness, to the precise reverse. What concerns us here is not the cognitive aspect of the mental, wherein fundamental entities are claimed to posses a mind, or behave in some way which might be called 'rational'. I shall refrain, therefore, from considering panpsychist theories which propose any mental property beyond phenomenal states. Theories which hold phenomenal properties to belong even to non-functional entities are more properly described as 'panexperientialist', however I shall continue to use the term 'panpsychism' as it accurately describes the category in which all such theories are contained.

**Panprotopsychism**

Panpsychism and panprotopsychism are both classified by Chalmers as type-F models of consciousness. This is because the feature which they have in common is the notion that entities may have intrinsic properties which are inaccessible through scientific (i.e. third-person) means. However, the inherent difference between them is that panprotopsychism is an emergentist position, and in some ways is more compatible with physicalism than panpsychism.

Of course there is a trivial, physicalistic sense in which panprotopsychism is true; if one accepts that fundamental entities are non-conscious but that humans are, then one must accept that fundamental entities are proto-conscious in the same way that a person is a proto-nation. In other words, that they may one day be in a position where they form part of a conscious system, but do not necessarily have any properties other than their physical ones which can account for this. Obviously, a physicalist would argue just that.

Chalmers classifies panprotopsychism as a monist position, because his own version of it holds there to be one kind of substance in the world, which is a 'natural' substance consisting of both physical and protophenomenal fundamental properties. Another way in which panprotopsychism may be considered monist is if the physical is all that exists, but that there are (proto)phenomenal aspects to physical entities by way of their possessing an intrinsic character. This is an idea has been explored by Stoljar (2001) and Russell (1927), most notably, and is one to which I shall return when I come to the synthesis of my own model of how consciousness relates to the systems which possess it.

Russell famously wrote that physical entities are described by their causal relationship with other entities. In other words, a description of an entity's entire causal impact on the world constitutes its complete physical description. However, such dispositional relationships (such as charge and mass) are likely to be grounded by the intrinsic nature of these entities; a nature which is in principle inaccessible through physics. For instance, two particles may share a precise causal pattern in the world, identifiable as electrons. However, they may not be identical *intrinsically,* a state of affairs which we would have no way to verify. Already this has echoes of the problem of consciousness, and indeed has been suggested as a solution to it; if the intrinsic characters of entities are either phenomenal themselves or can *constitute* phenomenal properties (i.e. are protophenomenal) then there is an inherent relationship between the physical and the phenomenal which is entirely compatible with an otherwise physical ontology. For instance, microphysical causal closure is preserved, as it is not with theories such as epiphenomenalism, and the laws of physics remain unaltered. Whether such a model is called physicalist or dualist is a matter of semantic debate, but it does seem to be at the very least a *property* dualist theory.

Panprotopsychism has been argued to have problems associated both with panpsychism and physicalism, and in the chapters to follow I shall examine whether or not it can be rescued from these criticisms, ultimately arguing that it can.

**Conclusion**

In this chapter, I have attempted to argue that, at least initially, a physicalist approach to an explanation of the existence and nature of phenomenal consciousness is unsatisfying. In the following chapter, I shall examine panpsychism, and attempt to argue that while it has several advantages over physicalism, an emergentist theory is far more desirable from a scientific point of view.

# CHAPTER TWO

# PANPSYCHIST MODELS

In the previous chapter, I argued that physicalist models of consciousness have certain problems, and that it seems desirable to solve these while remaining as true to a physicalist picture as possible. In the next chapter, I shall evaluate physicalism and its problems more thoroughly. However, because panpsychism more obviously suits the criteria for a solution which I stipulated previously, I shall examine that first. Although there are several kinds of panpsychist theory, it is important to emphasise here that I am only discussing panpsychism in terms of phenomenal consciousness; for psychological states, I am content with a purely physical model.

I shall begin by examining some of the advantages that panpsychism has over physicalism, and then its disadvantages, specifically those which are applicable to panprotopsychism. Finally, I shall discuss ways in which these problems can be easier to solve with a panprotopsychist model than with panpsychism. My conclusion will seem to motivate physicalism, however in the next chapter I shall attempt to apply a similar analysis to physicalism.

## The appeal of panpsychism

The most obvious benefit of adopting a panpsychist approach to consciousness is that one immediately eliminates the problem of emergence. In particular, true ontological emergence, where higher-order features are truly novel, as opposed to merely being epistemologically surprising. As physicalists assume non-functional matter to be non-conscious, they must account for the obvious consciousness of human beings, which is not an easy task if one wishes to do so

with a physicalist model. However, a panpsychist does not have to explain how simple matter becomes conscious in a particular functional arrangement, since he will assume that all matter is conscious, because all fundamental entities are conscious. This is not the end of the story for the panpsychist; at this point the 'combination problem' is apparent, which I shall discuss this later in the chapter.

Some advantages include that a panpsychist does not *have to* attribute a causal role to consciousness (though it could), and is capable of admitting that consciousness is surprising, and that a world with no phenomenal consciousness is conceivable. As I shall briefly explore in the next chapter, physicalism's denial of the conceivability of a zombie world is one of its most significant problems, because it seems counterintuitive at best and question-begging at worst.

Nagel (1979) puts forward four premises which seem to indicate that a panpsychist model of consciousness is viable. Of course, Nagel was writing about mental properties in general, including those which I would accept to be reducible to physical properties (memory, report, access *etc*.), however I believe that several of his arguments are applicable to the case of phenomenal consciousness. I shall now examine some of his arguments.

One of the principles of computing is that of 'multiple realisability'; that two sets of apparatus may perform the same calculations as long as they are functionally equivalent, regardless of their material composition. In fact, this principle is taken by some (including Chalmers, 1996) to include the functions of the mind; an arrangement of silicon chips which is functionally identical to a human brain should be capable of precisely a brain's function. Nagel makes a similar point, in that a living organism can be composed of any matter at all, be it carbon

molecules from a distant star, or iron from the Earth's crust. He concludes that because any matter can constitute a living, conscious organism, the matter itself must possess mental properties such that they can produce *our* mental properties in proper combination. He calls this "a kind of mental chemistry." (*ibid.,* 182)

Nagel's second and third premises are that consciousness is not a physical property, nor implied by the physical properties, of a system. However, it is a property of the possessing organism, rather than a substance in its own right (as a Cartesian would have it) or a property of some entity such as a soul.

Finally, Nagel argues that no system demonstrates truly (ontologically) emergent properties; all properties of such a system are merely the aggregated properties of its constituents. It may be that we cannot know how the constituents will combine without first combining them, giving the final system epistemologically emergent properties, but there are still no new properties to the whole system which are not combinations of its constituent properties.

Nagel goes on to say that if we take mental properties to be non-physical (and here, he is talking about psychological as opposed to phenomenal properties, but if we take him to be talking about phenomenal properties his arguments thus far hold), we cannot do other than admit the basic constituents of a living organism to have non-physical properties also. Therefore, we must conclude that panpsychism is true. I believe that Nagel may be partially correct; I am inclined to believe it to be true that if a system has non-physical properties, its constituents must also. However, the system's non-physical properties being

phenomenal does not seem to necessitate the constituents having phenomenal properties.

Nagel argues that because there are no ontologically emergent properties of a system, any system with property Q must have some component or components that also have property Q. Yet I am not convinced that the lack of truly ontologically emergent properties gives us sufficient motivation to be panpsychists. Water is liquid at standard temperature and pressure, so let this be property W. Neither oxygen nor hydrogen, water's only constituents, exhibit property W, but this is trivial; knowing everything we do about hydrogen and oxygen, we could predict how water will behave. Property W is not ontologically emergent, and may not even be sufficiently surprising to be epistemologically emergent. Nonetheless, the W property does not appear until the 'system' is in place, and does not descend to any of its constituents. Therefore, it does not seem to injure my case to admit that phenomenal consciousness might not be ontologically emergent; even if it is not, I do not have to admit that it is a property which extends to my constituent components.

In short, the intrinsic properties which many panpsychists espouse do not themselves need to be phenomenal in order for the properties of overall systems to be so; it is here that panpsychism and panprotopsychism fundamentally differ. I shall expand upon this in the final chapter, when I discuss how intrinsic non-phenomenal properties relate to a system's consciousness.

**Problems with panpsychism**

A panprotopsychist theory can share many of the problems of panpsychism, and seems to lack the latter's most significant advantage: avoiding the problem of

emergence. A panprotopsychist model still has to account for how consciousness emerges. I shall now explore several problems shared commonly by panpsychism and panprotopsychism. I shall try to show how in most cases a panprotopsychist theory is less susceptible to the problems, and finally I shall argue that panprotopsychism has one clear advantage over panpsychism: it can account for the one example of a system we can be sure is usually conscious, but that loses consciousness while remaining functional.

One of the most apparent problems with panpsychism is that it lacks even the weak empirical link between physical and mental properties that a broadly functionalist approach espouses. Most emergentists would argue that there is likely to be a correlation between functional and phenomenal states, and that one can infer conscious properties to exist in functional systems at least as complex as human brains (although different theorists are more or less generous as to this point, and one cannot strongly assert an entailment between functional systems and consciousness, based solely on our own functionality and consciousless). However, there is none but the flimsiest evidence that simple entities have anything which might be called behaviour. In other words, the neural correlates which exist in humans and most animals have not been observed in inanimate objects.

A major difficulty for panpsychism is the combination problem. Although panpsychism is not emergentist in that it does not claim phenomenal consciousness *per se* to be emergent from physical arrangements, it still must be the case that conscious experience grows more complex as functional complexity increases. A human mind would experience much more vivid qualia than a thermostat, and so much more than a fundamental entity that the consciousness

of an electron must be incomprehensibly basic. The combination problem thus arises, and has two aspects: why, if all matter is conscious, is there not a more general diffusion of consciousness; and how is it that the consciousnesses of more basic entities become subsumed into the consciousness of the whole?

The latter is more of a mystery than a fundamental problem, however the former certainly represents a problem for panpsychism. Descartes famously held the mind to be non-physical, there being no coherent sense in which it could be said to have a location in space. However, it seems to be perfectly sensible to assert that my consciousness is located in my body, as opposed to outside it. I might even be more specific and point to my nervous system as its ultimate location. However, if panpsychism is true, then why does the consciousness of my brain not merge with the consciousness of the molecules within my skull, skin and so on? Why, when molecules from my food become part of me, is my experience not supplemented?

An answer to this may come from what I have said previously about the simplicity of the qualia of simpler entities. In fact, I might well be experiencing the collective qualia of every molecule in my body, but my senses and the processing which enables them to work are flooding my consciousness with far more vivid qualia than I would be experiencing as an inert body (a panpsychist must surely concede that a corpse has consciousness, although to what extent they very likely differ). In fact, I shall make a similar argument later on when I argue that panprotopsychism can avoid the combination problem. I do not believe that panpsychism can avoid the problem, because in panpsychism the only aspect of consciousness which is altered by increased functional complexity is its intensity, rather than its presence. Thus, I believe it is much more difficult

for a panpsychist to dismiss the combination problem than it is for a panprotopsychist.

The other aspect of the problem remains, however; my nervous system is comprised of many functional systems, yet I seem to experience a constant, integrated stream of experience. This is sometimes referred to (such as by Revonsuo, 1999 and Blackmore, 2003) as the 'binding problem'. The term is also used for a problem in the study of visual perception, and it is a very similar problem: how do individual elements of perception (or in our case, qualia) come together to form what seems to be a 'Cartesian theatre' – a unified, smooth experience rather like a film? Unless one wishes to argue that there is indeed a homunculus within one's mind for whom an elaborate sensory theatre is being orchestrated, there is immediately a problem with regards to how experience comes together. After all, in neurological terms, the brain is divided into quite differentiated areas of processing.

I believe that this is what Chalmers would call an easy problem, because it does not seem to be a problem of phenomenal consciousness, but a problem of psychological consciousness. In other words, it seems clear that the reason we have a unified, smooth experience is that we have neurological mechanisms which bring all of the disparate processes together in some way. I believe that a standard psychological explanation will suffice here, given the relationship between the phenomenal and the physical which I discussed in the first chapter. I shall now review, briefly, a few prominent neurological and cognitive theories of the unity of consciousness. Again, what I am aiming to show is that a panprotopsychist theory can use such models to solve the combination problem, but a panpsychist theory cannot (at least, it cannot as easily), because the latter

cannot make an appeal to the functionality of a system without itself becoming emergentist.

Certainly, we know from neuropsychology that the information our brains receive is filtered so that insignificant data are excluded from our perceptions. It has also been demonstrated that our perceptions can be altered as we are experiencing them (for instance, several optical illusions are broken by the addition of some piece of information). Dennett (1991) proposed something called the 'multiple drafts' model, wherein some sort of neurological process 'oversees' the various perceptual inputs and constructs a dominant perception according to various criteria (if something makes sense based on past experience, for instance). This idea leads to an argument for the evolution of consciousness (and here, I refer to psychological rather than phenomenal consciousness). There may be no Cartesian Theatre, however the illusion that our minds are whole and unified is a strong and very pervasive one. This seems to have an evolutionary advantage, since the brain has to control all of our motor functions. A brain that could not correctly integrate all of its inputs would result in a body which was not properly under control. Thus, the mechanism which gives the illusion of the Cartesian Theatre seems to be necessary if psychological order is to be maintained.

Crick and Koch (1990) argued that the synchronicity of neural firing in the visual cortices of cats (so-called 'gamma oscillations' of 35-75 Hz) had something to do with the binding of visual information into a single stream of visual consciousness. The idea is that when all of the neurons which are processing the information derived from the perception of a single object are firing in synchronicity, the object 'comes together' in the brain. Crick (1994, 245) called this "the neural correlate of visual awareness." A mechanism within the brain

(suggested by Crick and Koch to be the thalamus) selects which features are to be bound together, and then binds them by synchronising the firing rates of the relevant neurons.

What relevance are such theories to the combination problem? I believe that they demonstrate something crucial, namely that consciousness seems to come about only with the correct arrangement of functional systems. However, this argument may be equally applicable to panpsychism; the reason I am not experiencing the consciousness of my individual neurons may simply be that the 'editor' (to use Dennett's newspaper analogy) which intercedes between the higher and lower levels of consciousness does not exist. That is to say, the system which unifies my psychological (and therefore phenomenal) consciousness by linking all of the different functional areas of the brain, may not happen to have an equivalent which is able to 'feed' the consciousness of my brain's constituents (*i.e.* neurons) into the consciousness of the larger systems. To distinguish panprotopsychism from panpsychism, there must be a situation in which the two disagree as to whether a system is conscious. There must exist a functional system which does not produce consciousness, because that would demonstrate that an emergentist theory was correct, while casting serious doubt over a panpsychist theory. Unfortunately, in most cases, it is impossible to show conclusively that a system is not conscious. However, I believe there to be one case for which we can make strong, evidenced arguments.

**Unconsciousness**

As previously mentioned, it is not enough for me to say that because I am both functional and conscious, I can assert that all conscious entities must be

functional. This idea has some *prima facie* appeal, but it is too weak to do the job I would have it do: to show that consciousness requires function.

If panpsychism is correct, then basic, non-functional entities are conscious, and their consciousness is merely *amplified* by being in a functional system. If there were a functional system which did not exhibit consciousness, it would severely weaken the claim that all functional systems were conscious. If not even all functional systems are conscious, then grave doubt is cast over the notion that any non-functional entity could be. Previously, I discussed Chalmers' idea that one cannot in principle say that even the most basic functional system is non-conscious, and that this can lead not only to counterintuitive conclusions, but ultimately to panpsychism. I believe that I can demonstrate one case of a complex functional system which we have no reason to believe to be phenomenally conscious: the sleeping human brain.

Of course, a brain in deep (or slow-wave) sleep is still functioning, and is still a more complex functional system than Chalmers' thermostat. However, it exhibits none of the standard psychological properties which we might label as conscious. There is no sense of time, nor of one's own body, one has no awareness of self or surroundings, and one does not generate memories (which is not to say that memories are not in some way processed during this time, merely that no new ones seem to be generated). In fact, compared to when one is awake, there is very little brain activity during slow-wave sleep; neurons exhibit 'delta waves', meaning that they only exhibit measurable electrical activity up to four times per second (compare to the 30-100 cycles per second demonstrated during gamma activity, such as when integrating sensory data from two or more senses (Kisley and Cornwell, 2006)). It should be noted that electroencephalographs only

measure synchronised activity by large clusters of neurons; individual neurons may still be firing far more often.

The part of the brain which is chiefly responsible for the regulation of circadian cycles, the suprachiasmatic nucleus, remains active throughout the sleep-waking cycle (Mistlberger, 2005), so that it can regulate the stage of sleep which the brain occupies. Furthermore, one can still detect sound, albeit at a much higher threshold than when awake (Bonnet and Johnson, 1978). This indicates that the brain is still functional, and in a state analogous to an electronic device that is 'on stand by'. The brain is at its most basically functional during this state, and it is likely that it does not produce consciousness.

It may seem as though to assert strongly that one is not phenomenally conscious during deep sleep is to say more than the evidence allows. It is true that, as with any matter of phenomenal consciousness, there is no empirical way to be sure. However, I believe it is possible to be reasonably confident in the matter. I believe that I am not a zombie because not only am I experiencing qualia at this very moment, I remember having done so in the past (which is to say, I can re-create certain sensory conditions at will, and they accord with what I believe has happened to me in the past, so I presume that I was conscious at the time of these events). Furthermore, I demonstrate certain kinds of measurable brain activity when I am experiencing these qualia. In other words, I am conscious by the standard psychological definitions of the term. As I have argued before, the conditions which we label as 'psychological consciousness' are the neural correlates of the phenomenal consciousness which I cannot demonstrate to an external observer. They are, in short, the only reason I have to assume that anyone besides myself is capable of experiencing qualia. It is not conclusive

evidence, but it makes enormous intuitive sense; I have no reason to think that I alone am conscious among zombies.

However, on a nightly basis, I undergo neurological conditions which do not constitute any form of psychological consciousness. In fact, my brain activity is very slow indeed (normally, only infants exhibit wakefulness while in delta-wave activity (Taylor and Rutter, 2002)). Furthermore, unlike the periods of sleep which I do remember (*i.e.* the periods of rapid eye movement which indicate dreaming), I have no memory at all of these periods of deep sleep. Without the ability to remember having experienced qualia, and knowing that I was not conscious in any psychological sense, I do not believe I can presume to assert myself as having been phenomenally conscious. Furthermore, I believe that I can, being the *only* authority on my own consciousness, rule out my consciousness during deep sleep.

If I am indeed not phenomenally conscious while my brain is still a functional processor, then I do not think panpsychism can hold. For, if I assert that my qualia were merely less complex in a slow-wave state, while lacking even the first-person evidence which suffices for my belief in my own consciousness when awake, I am making a baseless assertion. I have, quite literally, no more reason to believe that I am conscious during slow-wave sleep than I have to believe that an electron or a quark is conscious.

This is no problem for a panprotopsychist; a panprotopsychist can say that phenomenal consciousness only comes about as a property of functional systems *of a certain complexity.* In fact, this even gives us an idea of where to draw the line on Chalmers' sliding scale; if a psychologically conscious human is

phenomenally conscious, and a partially-conscious human (*e.g.* one in REM-sleep) has less vivid qualia, and an unconscious human experiences no qualia, we have measurable standards for comparison with other functional systems. There would be, in short, no need to assert the consciousness of anything simpler than a human brain in the delta state, because we would already have ruled out the consciousness of such a brain.

This is not a refutation of panpsychism; a panpsychist may argue that because the unifying system responsible for the over-arching consciousness we experience is not active during unconsciousness, that the unified consciousness is not present, although the consciousnesses of individual units would still be present. However, such a position would seem akin to making excuses. Let us consider the three relevant positions with regards to this example. First, the physicalist approach does not need to evoke consciousness in lesser brain systems, so can accommodate unconsciousness with no difficulty. My position, as we shall see in the final chapter, can accommodate unconsciousness because the intrinsic character of the slow-wave brain is not such that it features consciousness. The panpsychist response is similar to my own, however it seems to run afoul of Occam's razor, by including an entity (phenomenal consciousness) beyond its necessity. Once again, panpsychism is the least intuitive and most elaborate among a number of models. It is far easier to accept physicalism, or a panprotopsychist approach such as my own, than to continue to add counter-intuitive excuses for every problem which panpsychism faces.

Another piece of evidence which seems to support the idea that a functionally complex system need not be phenomenally conscious is that there is undoubtedly a great deal of unconscious processing within the brain. That is, a

lot of processing which occurs without there being anything it is like for these processes to occur. A rather obvious example is that of brain output; I cannot attend to any sensation of commanding my limbs to move. There is nothing it is like for my brain to control my body, yet it quite obviously happens. In the final chapter, I shall attempt to account for the difference between phenomenally conscious brain processes and those which are not accompanied by qualia.

**Conclusion**

I believe that my final theory will have something in common with panpsychism; that is, a recognition that consciousness is non-physical, and a need to show how consciousness is extant in light of these non-physical properties. However, I do not believe that the notions of fundamentality and ubiquity central to panpsychism hold, nor that they need to in order to avoid the problems of physicalism.

I have argued that while both panpsychism and panprotopsychism can avoid the combination problem, a panprotopsychist theory can do so far more straightforwardly with appeals to cognitive psychological models of informational binding. The ability of a panprotopsychist theory to deny phenomenal consciousness to systems and entities which are not functional in the right way is a strength, because it can avoid intuitive problems such as the phenomenal thermostat and the absurd notion of conscious electrons. Furthermore, I have argued that the most important difference between panpsychism and panprotopsychism is that the latter can accommodate non-conscious functional systems while the latter cannot accommodate any non-conscious system; for this, I have appealed to the one example from science of a system whose consciousness depends upon its functional state.

# CHAPTER THREE

# PHYSICALIST MODELS

In the first chapter, I argued that physicalist attempts to bridge the epistemic gap between the physical and the phenomenal were unsatisfactory. In the previous chapter, I argued that panpsychism is not a viable alternative, and made arguments which seemed to lead us towards a physicalist conclusion. Now, I intend to alter the mood of the dissertation once again, by presenting arguments against physicalism. My purpose here is to show that there is a middle ground between panpsychism and physicalism which can appeal to proponents of both kinds of model. A panprotopsychist approach may be simpler and more appealing to a certain kind of physicalist than the elaborate counter-arguments which physicalists often employ. Throughout this chapter, I shall refer to *a posteriori* physicalism, since the *a priori* variety (as discussed in the first chapter) constitutes more of a denial that there is a problem, rather than a solution to it.

I shall begin with an analysis of physicalism in general, and its advantages over the main alternatives: substance dualism and panpsychism. I shall then discuss its problems, namely those deriving from the knowledge and conceivability arguments. Finally, as in the panpsychism chapter, I shall try to show that panprotopsychism, properly interpreted, can avoid these problems far more easily than standard physicalism.

## Conceptions of the physical

If, as I mentioned in the first chapter, I am content to preserve a physicalist picture as it applies to every other phenomenon in nature, why is it that it does

not apply to consciousness? In order to answer this question I shall examine two pertinent conceptions of what it means for a property to be considered physical. I intend to argue that of the two conceptions of the physical I shall here examine, one cannot account for consciousness, and the other, which is far more likely to be able to do so, is not in fact a conception of the physical at all, at least in the standard sense.

Central to my forthcoming argument is the notion that phenomenal consciousness is subjective, which is to say, accessible only in the first person. All that is accessible from biology and psychology are the neural, chemical and behavioural correlates of consciousness. These are all, of course, physical in the sense that they supervene upon physical entities and properties. Therefore, for a theory to be able to account for consciousness, it must be able to account for the subjective. If we are to avoid substance dualism, we must be able to place the subjective into our ontology as either fundamental (as a panpsychist) or part of some other aspect of the natural world, physical or otherwise. In the final chapter of this dissertation, I shall appeal to the notion of the intrinsic character of physical entities. As I am about to attempt to show, this notion can form a part of physicalism insofar as physicalism can admit the existence of intrinsic properties. Ultimately, however, such an admission significantly blunts the claim that all properties are physical.

Stoljar (2001) argued that there are two possible conceptions of the physical. The aim of his paper was to show the inherent error of the conceivability argument against physicalism: that the first premise talks of one conception of the physical, and the second premise talks of the other, and therefore that the conclusion (that physicalism is false) cannot be reached from their conjunction.

The first conception is the theory-based conception of the physical. A property is physical under this conception if is a property contained within physical theory, or supervenient upon such a property. Charge is a physical property under this conception, because charge is one of the properties of fundamental entities as described by physical theory. Chemical properties such as electronegativity are also theory-physical, being supervenient upon fundamental physical properties. The second is the object-based conception, under which a property is considered physical if is required for a complete description of *all* properties of a paradigmatic physical object. The composite property of being a stone is thus a physical property under this conception.

By 'all properties', Stoljar is referring to the intrinsic properties of physical objects as well as the properties contained in physical theory. Physical theory, after all, contains only such properties as are relational or dispositional. Therefore, there are more object-physical properties than theory-physical properties. The intrinsic properties would not be accessible from the outside (*i.e.* to science), and the only allusion to their existence is the physical (*i.e.* relational and dispositional) properties which may supervene upon them. If we were to find consciousness, it would surely be among those object-physical properties, for consciousness is inaccessible from the outside. Thus, if we accept the subjective, first-person nature of phenomenal consciousness, it is clear that only this conception of the physical can accommodate it.

Is the object conception physical? This is primarily a matter of semantics, and whether a particular physicalist wishes to argue that object-physical properties are themselves physical will depend upon his attachment to calling himself a

physicalist. I, for instance, am not content to define object-physical properties as physical if they do not concern an entity's dispositional relationship towards other entities. In short, physics concerns causality, and intrinsic properties have no causal effect. Furthermore, any attempt to force all properties in the Universe into the category 'physical' seems suspicious, for it makes physicalism both undeniably true, and trivially so; without a concise definition of what is physical, the term 'physicalism' loses any useful meaning.

Since I shall be utilising the term frequently from this point, it would be useful to clarify what I mean by 'intrinsic'. Of course, it is the case that in a certain (everyday) sense of the term, an electron is *intrinsically* negatively charged. This sense seems to be synonymous with 'by definition', and is not the sense in which I employ the word. Another aspect of the common definition of 'intrinsic' is the notion that an object *necessarily* possesses such a property. Needless to say, this is not part of my use of the word. By 'intrinsic', I mean those properties which describe what an entity *is,* independent of its dispositional properties, i.e. charge, mass and so on. An electron has certain physical properties, but it is likely that there is also a character to it which is intrinsic, *i.e.* the thing in itself. Two electrons could be physically identical, but intrinsically quite different. Characterising intrinsic properties is exceptionally difficult, since the only properties we can conceptualise are physical ones. Even if one is a physicalist, however, one can see that this is very much akin to the relationship between neural activity and phenomenal consciousness: we never have access to consciousness, it seems to be 'inside' the brain in some way. In an analogous way, the intrinsic nature of an electron is 'inside' the electron, and can in principle never be accessed, since for equipment to detect something requires interaction, and interaction is by definition physical.

**The appeal of physicalism**

The most obvious advantage of physicalism over dualism and panpsychism is that it makes a certain amount of intuitive sense. Given that physicalism is paradigmatic in all areas of science, it is easy to see how one might naturally presume its truth in the case of consciousness. Substance dualism held a similar appeal in Descartes' time, when the idea of there being immaterial souls of some kind was commonplace as a result of the religious influence on Western culture.

The chief advantage of physicalism over substance dualism is that it does not have to account for interaction between mind and body. The precise nature of mental causality remains a mystery, but it is generally taken that the decision-making and cognitive processes are physical, so there is no need to account for a non-physical entity somehow interacting with a physical one.

An intuitive advantage of physicalism over panpsychism is that it removes the somewhat uncomfortable notion of simple entities having an inner mental life. Of course, the problem of emergence remains, and whether or not anything can be truly ontologically emergent is an ongoing debate. It seems clear that if a theory can account for consciousness without requiring the existence of ontological emergence, such a theory would have an advantage over the alternative, simply because it would be less metaphysically elaborate. Such a theory could emerge from the physicalist paradigm, as physicalism *per se* does not require that consciousness be truly emergent (some versions do not even hold that it is epistemologically emergent, being simply obvious from the totality of physical facts about the world).

**Causal closure**

Another argument often employed is one from the principle of causal closure. It is argued (such as by Yablo, 1992) that, since mental events cause physical events, and the physical is causally closed, the mental must itself be physical. This is obviously advantageous to those who (like me) want to retain a physicalist ontology. However, the argument may be of little or no relevance to the matter of phenomenal consciousness.

As argued in the first chapter of this dissertation, it is not obviously the case that phenomenal consciousness has a causal role to play, and there seem to be strong arguments *against* its being part of any causal chain. Yet there is more to the mental than the phenomenal; the observable biomechanical chain of events that leads me to avoid painful stimuli can be explained, as I have argued previously, without referring to phenomenal consciousness. The events in this chain are 'mental' in that they affect an organism's behaviour, so according to Yablo they would indeed have to be classed as physical. Yet if the phenomenal properties of the system are not part of its function and therefore causal chain, we are not forced to admit that they must be physical.

The strengths of physicalism seem to revolve around its intuitive appeal and its ability to conserve the nature of the Universe as we currently see it. The argument from causal closure not only preserves the physical status of mental events, but does so by evoking a principle of physics with which people are generally very comfortable. The argument seems to hold if by 'mental' we are referring to psychological mechanisms such as decision and the command of limbs. However, for the purposes of this dissertation, the only relevant aspect of

the mental is the phenomenal, and it is not clear that phenomenal consciousness has a causal role.

**Jackson's knowledge argument**

It is a fact doubted by very few (albeit one which can only be verified in the first person) that consciousness exists in this world, and it is also a fact that there is no agreed-upon physicalist explanation for how this phenomenon occurs. Therefore, there is an epistemic gap. Whether or not there is an ontological gap I shall explore with two arguments; the knowledge and conceivability arguments.

Jackson's (1982) knowledge argument can be used as an example to illustrate a problem with physicalism which I believe a panprotopsychist theory does not have. The scenario, in short, is that a physicist, Mary, is housed in a completely colourless room in which she learns every physical fact about visual perception (in a world where all such facts are known). When she has completed her training, and knows every physical fact about visual perception, she steps out of her room and sees a red flower. The question that is then posed is: does Mary learn something new when she sees the flower? If she does, then the physical facts about perception do not constitute *all* of the facts about perception, and physicalism is false.

There are several physicalist arguments which attempt to show either the invalidity of the argument or its inability to cause distress to physicalism. The former sort, for instance, might argue that Mary could not have known all physical facts if she did not know what red looked like. This, it seems, is question-begging, because such an argument must take as a premise that phenomenal redness is a physical property. The two other dominant positions in

response to the knowledge argument are known as the ability hypothesis and the acquaintance hypothesis. I shall outline both of these here, and shall argue that ultimately neither is as intuitively appealing as accepting the knowledge argument and moving forward with a theory which is not entirely physicalist. I shall focus on the arguments made by David Lewis in particular, as they highlight what I believe to be a fundamental problem with physicalist counter-arguments to the knowledge argument.

The ability hypothesis (defended primarily by Lewis, 1988 and Nemirow, 2007) is that Mary does not learn any new fact when she sees colour for the first time. Instead, she gains the ability to recognise that colour. In other words, in learning what red is like, Mary in fact learns how to visualise and recognise red. The alternative, according to Lewis, is to accept that learning what red looks like constitutes propositional knowledge, and that because this is incompatible with physicalism, the ability hypothesis is preferable.

On the one hand, it seems as though I might sympathise with this; I am attempting to argue for a theory that is as compatible with physicalism as possible. However, to argue that a theory should be rejected because it is incompatible with physicalism presupposes that physicalism is preferable to every alternative, which is what the knowledge argument disputes. Lewis further argues that the ability hypothesis is not only compatible with physicalism, but explains everything that could be explained by claiming Mary's knowledge to be factual. I believe this not to be the case, for by definition the ability hypothesis does not explain how Mary gains propositional knowledge, as it seems clear that she does.

When Mary is learning everything physical about colour perception, for example, she will learn everything that happens in the realm of physics, chemistry and biology which results in red sensations. She would certainly know the names of a great deal of colours. She would not, however, be able to map any quale to a colour name. Lewis himself gives the analogy of an x-y co-ordinate graph; being given all of the facts about a point's location on the x-axis tells us nothing about its place on the y-axis. Only when Mary exits the room, and is probably told by someone that this flower is red and that its stem is green, can she learn *that* this is what red looks like. No new ability seems to have been acquired.

Indeed, a major problem with the ability hypothesis is that abilities seem to be able to vary greatly while experiences remain the same. Suppose that Mary does not have the ability to visualise colours. This does not mean that, when she experiences green, she cannot identify it as green. When she is no longer having the perception, she cannot remember what the experience was like. She has not gained any ability by looking at the colour, but as long as she is doing so, she has some sort of knowledge: knowledge *that* green looks a particular way.

Lewis agrees that no amount of lessons in the psychophysics of perception will grant Mary the knowledge of what it is like to see red, however he emphasises that no lessons of any kind can provide this knowledge. Hypothetically, if Mary were to learn every *non-physical* fact about something, she still would not be able to deduce the *experience* of seeing red. At this point, in order to avoid the precise kind of question-begging that I have only lately accused certain physicalists of doing, it seems possible only to say: talk of physical and non-physical lessons is clearly unhelpful. Whether one is a physicalist, a dualist, a panpsychist, or what have you, to say that Mary learns everything *simpliciter*

about colour perception but never learns what red looks like is clearly nonsensical. Lewis and I might both conclude that the only way to have a thorough education on the subject is to experience every colour. According to Lewis, then, Mary's education is not incomplete because it is a physicalist education; it is incomplete because its teaching methods are insufficient.

I believe the point remains, however, that Mary learns something new upon seeing the colour red for the first time. It is significant that while physical facts can be conveyed by one person to another, phenomenal ones cannot. I do not think that Lewis' casual reference to Mary potentially learning non-physical facts holds much weight; such things seem to be as inherently difficult to discuss substantially as qualia. Since we have no access to anything non-physical (as previously discussed, the only realistic candidates for non-physicality are intrinsic properties, and we have no access to these), it is hard to see how Mary could possibly learn them from lessons. This is not, in itself, an argument for qualia being non-physical, but it highlights that phenomenal facts and 'non-physical facts', whatever these might be, have one thing in common: neither can be conveyed, in stark contrast to physical facts. That there can be facts outside of physical facts is the point of the knowledge argument.

Ultimately, the ability hypothesis is unconvincing. The distinction between the knowledge that Mary gains and the ability that she gains is far too strong for the hypothesis to be sustainable.

As Lewis hinted, there seems to be no way other than acquaintance by which Mary can come to know what red is like. Some, such as Conee (1994), believe that acquaintance is an entirely separate kind of knowledge which is neither

propositional nor an ability. Mary, then, *knows* everything about colour perception before she exits her room, but is not *acquainted* with it. She knows everything about the physical mechanisms behind the perception of the colour red, but she has not had intimate access to it, and this is what she gains when she sees it for the first time.

There is a large and obvious flaw in this hypothesis. I certainly agree that acquaintance is the only way in which knowledge about consciousness can be obtained. What is less clear, however, is how a physicalist can account for this. I can, as a non-physicalist, easily accommodate facts which do not pertain to the physical realm and as such cannot be accessed by physical means (including teaching, which is a physical method of information distribution). It is not at all clear that a physicalist can account for this. Merely admitting that this is the case, as Lewis does, does not constitute an *account* of why such facts are incommunicable.

There does not seem to be a physicalist counter-argument to the Mary case which is less elaborate and counter-intuitive than the initially uncomfortable conclusion that physicalism is far from certainly true. Not according with intuition is not a major problem normally, but since the knowledge argument is an argument which appeals to intuition, it seems problematic that physicalist solutions are, more often than not, less intuitive than the admission of non-physicalism. In the final chapter of this dissertation, I shall return to the knowledge argument and argue that my own theory can accommodate Mary's situation without compromising too much of a physicalist attitude. It is my hope that physicalists will be able to accept my position, reconciling the knowledge argument with a naturalist model.

The *a posteriori* physicalist view has intuitive appeal in that it preserves a physicalist view of the world while still acknowledging the evident existence of consciousness. In the first chapter I discussed the way some compare knowledge about consciousness with knowledge about water; once one has enough empirical knowledge of water, one can know that it is identical with $H_2O$. However, there seems to be a disanalogy between water and consciousness. It is impossible for us to conceive of a world identical to our own with regards to $H_2O$ but which differs with regards to water. To modify Jackson's thought-experiment, Mary would know everything about water once she knew everything physically about $H_2O$ (except what it is like to drink it and so on, but that is hardly the sort of objection that a physicalist would raise). A parallel thought experiment wherein Mary learns every physical fact about water, steps outside of her laboratory and discovers that water is $H_2O$ does not make even *prima facie* sense. Of course, for a physicalist to claim that the original version does not make sense for a similar reason is to assume that there is a physical explanation for consciousness. The $H_2O$-water identity has been empirically confirmed; the consciousness-function identity has not.

It seems that problems with this form of physicalism can be revealed by the following argument involving intensions. Where the extension of a term is its truth value (i.e. its presence or absence in the world), intension refers to possible worlds in which a statement might be true. The primary intension of an expression is whatever fulfils the role of that expression in a particular centred world. For instance, the primary intension of 'water' would be something along the lines of 'the clear, drinkable liquid in rivers'. The secondary intension is fixed in all counterfactual worlds by its use in an actual world, *ergo* the secondary

intension of 'water' is $H_2O$. These intensions have different epistemic functions, so that it is *primarily* or 1-conceivable that there is a world in which water is not $H_2O$, where the clear, drinkable liquid in rivers is something else in a particular centred world. It is not 2-conceivable, because the secondary intension makes us consider other worlds as counterfactual, so the term 'water' is fixed by our use for it to refer to '$H_2O$'.

Assuming one endorses the contentious position, as I do, that conceivability entails possibility, if it is conceivable that a world may exist which is physically identical to our own but has no phenomenal properties, then there is some world in which the primary intension of 'all physical facts but no phenomenal facts' is true. If this is the case, then some world must instantiate this fact, and thus physicalism (which holds that in all worlds, the totality of physical facts comprises the totality of all facts) is false. Note that the primary and secondary intensions of consciousness are identical; an experience cannot *feel* conscious and not therefore *be* an instance of consciousness.

An *a posteriori* physicalist may respond that the case of consciousness is unique, and that the psychophysical dependence must hold in all worlds, unlike other cases in which a statement S being conceivable entails some world instantiating S. However, it is arguable that such a position is motivated only by the desire to preserve physicalism. Another possible response is to claim the existence of intrinsic properties such as those mentioned previously.

**The conceivability argument**

A far deeper debate concerns the conceivability argument (Chalmers, 1996). This requires us to imagine that there is a world physically identical to our own but

which lacks phenomenal properties. The residents of this world, known as zombies, are capable of thinking and behaving as we do, however there is nothing it is like for them to do so.

The success of this argument depends upon there being an entailment from the conceivability of a proposition to its possibility; if there is no such entailment then the argument, while valid, does not pose a problem for physicalism. It would be impossible to discuss the whole breadth and depth of the conceivability-possibility debate here. However, I shall provide a very rudimentary argument for the entailment. I simply appeal (as Chalmers does) to logical possibility. If something is conceivable then it must be logically possible. If it is logically possible, then the world could have turned out that way but did not because of contingent natural laws. Therefore, the world could have turned out to be zombie-world.

As with the knowledge argument, the conceivability argument is designed to refute physicalism by showing it to be unable to accommodate something which is obviously the case. Once I have presented this argument, I will be in a position me to present my theory, and in the next chapter, I shall do so and show how well it bears the conceivability argument.

I have already mentioned *a priori* physicalism, which denies the first premise of the conceivability argument on the grounds that the proposition that there could be two worlds which are physically identical but not identical *simpliciter* is contradictory. My response to this argument is that its first premise seems to be that physical identity entails identity *simpliciter*, which is what physicalism holds to be true anyway. Thus, such an argument is question-begging.

A more suitable argument against the first premise might be to suggest that we cannot conceive of a physically identical world because our battery of physical properties may be incomplete. Dennett, for instance, claims that when trying to conceive of zombies, we may "underestimate the task of conception" (1998, 172). If it is possible that we do not currently know all of the physical properties or entities, then we can be said to have an incomplete notion of the 'physical' nature of an entity. Therefore, we may not have a clear enough understanding of what it means for two of anything to be physically identical. Thus, even if we are only asked to imagine two people rather than two worlds (itself an enormous task), we have insufficient data for a full picture. I do not believe that this argument is an especially powerful indictment of the conceivability of zombies. I cannot conceive of every property of a coffee mug, since if the battery of physical properties currently understood is incomplete, then the coffee mug could easily have physical properties of which I cannot conceive. Nonetheless, I can conceive of its relevant details; a vessel which holds liquid, has a certain weight, shape and appearance. Thus, I believe it can be said of me that I am capable of conceiving of it. When trying to imagine a zombie, we are asked to imagine someone with all of the normal human physical, anatomical and psychological characteristics, only without phenomenal consciousness. For obvious reasons, it is impossible to imagine what it is like to be such a person (because there *is* nothing it is like), so we imagine this being from a third-person perspective. That is not difficult; it is only a presumption that every other human is phenomenally conscious, but it is not something I take into consideration when imagining or interacting with people. Thus, to imagine a zombie, there is no reason why we must do anything more strenuous than imagine a person standing before us. Furthermore, we are not being asked to imagine the zombie

so much as the consequences of the person *being* a zombie, i.e. none. We are being asked to imagine that there is no outward difference between ourselves and our zombie counterparts. This is very easy indeed.

What route is available to a physicalist who does not deny either of the premises but only the conclusion; that physicalism is false? It might be tempting at this stage for a physicalist to say that while physicalism might not be true in every world, it is true in *this* one. I do not believe that this argument carries much weight, however, because physicalism is surely only a significant hypothesis if it holds the physical to be supreme *in every possible world.*

There does not seem to be much in the way of substantive counter-arguments to the conceivability argument. At least, none which do not lead to *a priori* physicalism, which is a distinctly undesirable class of theory. In the next and final chapter, I shall show that my theory is not weakened by the conceivability argument, although it may seem to be at first.

**Conclusion**

I have attempted to show that the two most significant anti-physicalist arguments, the knowledge and conceivability arguments, are sufficiently damaging to physicalism to warrant a new theory. The theory needs to be able to survive these arguments, or at least not be susceptible to any particularly substantial issues (such as those suffered by physicalism). Any new theory should also take into account the problems with panpsychism which I discussed in the previous chapter.

The most crucial aspect of physicalism which I believe should be preserved is the *general* physicalist picture, as it applies to every other field of science. Apart from consciousness, we have no reason to suspect the existence of anything that does not fit into such a model of the world. If we wish to create a theory which is compatible with a physicalist picture of everything else, then it needs to be based upon a conception of the physical which can accommodate the non-physical; such a theory should be based upon the notion of intrinsic character.

# CHAPTER FOUR

# INTRINSIC NATURALISM

Throughout this dissertation, I have used the term 'panprotopsychism' to describe the theory to which all others have been compared; the theory which I have promised to describe in this chapter. However, while the final theory proposed here will be panprotopsychist as it is defined, I do not believe that the term is precise. The theory I am going to discuss here is more properly called 'intrinsic naturalism'. It will share aspects of panpsychism, physicalism, and neutral monism. My aim is to arrive at a model which is compatible with the scientific spirit of physicalism, but which does not claim consciousness itself to be physically reducible. It can be thought of as belonging broadly to the category of property dualism.

The use of the term 'naturalism' in the name of this theory is justified in that the theory is perfectly compatible with the findings of science. Further, it shows consciousness to be dependent upon *natural* properties, not strange 'additional' or unscientific properties such as in a substance dualism.

## Intrinsic properties

Physics studies and explains the behaviour of entities in the Universe. It is fair to say that an entity's physical properties, therefore, are its dispositions to behave in certain ways under certain conditions. For instance, electrons have the physical characteristics of a particular charge, mass, spin and so forth. Each of these characteristics is an electron's disposition to act in particular ways under particular conditions.

These dispositional properties are likely to be supervenient upon intrinsic properties (Strawson (2008) argues that the two may even be identical, on the grounds that they are indistinguishable to us; I do not endorse this view, but I do concur that the dispositional and categorical properties of an entity are surely related). Intrinsic properties are properties which cannot be accessed through our senses, or through science. This is because, by definition, the only properties which science can examine are those which react with the world in some way; the dispositional or physical properties. The intrinsic nature of an electron remains mysterious, perhaps forever. I shall refer to the totality of an entity's intrinsic properties as 'what it is to be' the entity, or its 'intrinsic character'.

As an illustrative example, let us consider a being who can occupy any form it wishes, to the point where it is indistinguishable by any means at our disposal from the object which it mimics. If it chooses to take the form of a brick, it will demonstrate all of the properties of a brick, keeping its shapeshifter properties shielded. The brick properties are analogous to physical properties, being all that can be accessed by human science. No matter what we attempt, the object responds as a brick would respond, revealing nothing of its secret nature. It is objectively not a brick (or more accurately, there is more to it than its brick properties), yet we have no way of discovering this.

This notion of an objective character inaccessible by empirical means has its roots in Plato's cave allegory, as well as Kant's concept of a 'noumenal realm' in which the objective characteristics of entities can be found. In both of these ideas, a clear distinction is made between the sorts of properties found in standard empirical enquiry, and properties which are in some way 'internal' to entities, and cannot be accessed from the outside. These are the properties

which I call ‘intrinsic’. One can identify an object with its physical character (as we do in everyday discourse), or its intrinsic character (as I intend to, for the most part).

As I noted in the previous chapter, this distinction between the physical properties of objects and their intrinsic nature is one which has been made by Daniel Stoljar (2001). He classifies his model, however, as an *a priori* physicalist view of consciousness. This, I believe, is a mistake. The intrinsic properties which Stoljar hold to be a subset of ‘o-physical’ properties are not physical in any standard sense. They are, by definition, non-causal. An entity’s physical properties are the only ones to which we can ever access, because our methods are physical. Therefore, any theory which holds intrinsic properties to exist is acknowledging the existence of non-physical properties, and cannot legitimately claim to be physicalist.

Intrinsic properties are, as Stoljar argues, properties required by a complete description of an object, however they are not required for a physical description of the object. When I consider two worlds which are physically identical, I only consider what Stoljar would call the t-physical properties: the properties which are physical in the standard (dispositional and relational) sense. Stoljar, however, believes that physical identity also entails intrinsic identity. Thus, if consciousness is found among the intrinsic properties of the Universe, one cannot have two worlds which are physically identical and intrinsically (thus phenomenally) distinct. As I shall argue later in this chapter, I believe that the relationship between the intrinsic and the physical can be argued to be such that two worlds can differ intrinsically while still being physically identical.

**Perceivers**

The difference between 'what it *is* to be' and 'what it is *like* to be' is, I believe, inexpressibly enormous when discussing, say, an electron; it is, however, only slight when discussing a living human brain. Both entities have an intrinsic character, so why have I denied electrons to be conscious in previous sections of this dissertation? In the first chapter, I mentioned that it was likely that consciousness could only belong to a functional system. When I discussed panpsychism, and gave the example of a brain in a state of slow-wave unconsciousness, I introduced the idea that consciousness can only be a property of functional systems *of a particular kind.* Now I shall be yet more specific: I believe that consciousness can only be a property of perceivers. Further, that consciousness is part of the intrinsic character of such entities.

Let a perceiver be any system which takes in information, the system altering itself to accommodate and process the information, resulting in the creation of a representation of the external world. Clearly, a human brain is the paradigmatic example. I do not believe there is any *prima facie* reason to suggest, for instance, that a computer is not a perceiver. However, for simplicity's sake, and to avoid controversy, the brain is by far the more straightforward example.

An electron is not a perceiver. It reacts instantly to external conditions according to its dispositions to do so. It takes in no information, and is not capable of altering itself to accommodate or process the information. How could it? An electron is an elemental entity, having no component parts. What it is to be an electron is, quite literally, to be among the simplest entities in the Universe.

Let me briefly return to Chalmers' example of a thermostat, which is a functional system. It is true, as Chalmers argues, that thermostats hold 'information' in the strictest sense of the term. However, the information is not processed in any sense; while I might concede that the bimetallic strip is somehow 'representing' the ambient temperature by its curvature (and I doubt that I would concede even that), there is certainly not even the most rudimentary kind of information *processing* taking place. The difference between this system and a human brain is that the thermostat is only *passive* in its representation (if indeed the curvature of metal can be said to be a representation of temperature). A human brain constructs representations of the external world, in a series of processes which have been demonstrated to add information to plug gaps in sensory input (this is how optical illusions come to be effective). There is not even the most rudimentary sense in which any kind of construction is occurring within the bimetallic strip inside the thermostat.

The human brain, being a perceiver, is a system which takes in information, accommodates and manipulates that information, and produces some kind of output. It therefore has the intrinsic character *of* such a system. It is my intention to argue that phenomenal consciousness is part of this intrinsic character, and comes about as a result of the processing of particular kinds of information. My account has *prima facie* similarities with *a priori* physicalism, in that I seem to be identifying consciousness with representation. What differentiates my account from a Type-A account is that I am identifying experience with the intrinsic character of systems which create representations, not merely the systems or representations from a functional point of view.

We have already seen that there are kinds of information which the brain processes unconsciously. Here I have claimed that perceptual information is the sort that gives rise to consciousness, but what is the essential difference between perceptual input (which is conscious) and the processing and output (which is not)? It is difficult to articulate such an idea, and appeals to intuition would be circular: perceptual information is different because it seems different, but it is the fact that it seems like *anything* that is the puzzle. Perhaps an appropriate articulation would involve an appeal to representation; we have phenomenal experiences of perceptual input because the result of that input is the creation of an image (including sound, smell etc.), and consciousness is part of the intrinsic character of a system that creates such an image. Outputs such as the control of limbs, however, do not involve the creation of a representation.

How does this position deal with an *a priori* physicalist position which holds consciousness to be identical to our representations of the world? The representation is the result of physical processes, however it does not *produce* consciousness, nor *is* it consciousness. Phenomenal consciousness is an intrinsic property of systems which construct and hold such representations.

What sort of system does this? I have already argued against 'mechanical' thermostats, but there are electronic thermostats which operate with chips that detect the external temperature (by mechanical means as in a basic thermostat, using 'thermistors', which change their electrical resistance as temperature alters) and then transform that information in a representation, usually on a liquid crystal display that can be read. It is a very basic system compared to a brain, but I can see no reason not to qualify it as a perceiver: it absorbs information, which has an effect on the structure of the system, and manipulates

that information to produce an output, in this case changing what is displayed on a screen.

In short: Intrinsic naturalism holds phenomenal consciousness to be an intrinsic property of any system which creates and holds complex representational information. This happens whenever we are awake or dreaming; the 'Cartesian theatre' is very detailed and constantly updated.

**Physical-Intrinsic supervenience and the conceivability argument**

The exact relationship which my theory requires between the physical and intrinsic is a complex one, but I do not believe there are any essential problems with it. There must also be a relationship between the physical properties of systems and their intrinsic properties, otherwise a non-functional collection of the same entities would be conscious. Therefore, the physical (functional) nature of a system must be a factor in its intrinsic nature. Immediately a problem seems to appear; essentially I seem to be proposing is that the intrinsic properties of systems (henceforth $I_S$) are supervenient upon the physical, which itself supervenes upon the intrinsic properties of elemental entities ($I_E$).

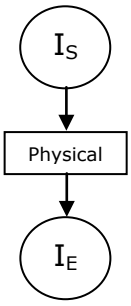


*Fig. 4.1 - Supervenience*
The arrows indicate supervenience.

If this were the case, then it might make intrinsic naturalism vulnerable to the conceivability argument, since no two worlds could be physically identical without also being $I_S$-identical. In zombie-world, the functional systems are all

dispositionally identical; why should the intrinsic properties be different? After all, zombies are perceivers and thus there is something it *is* to be them, even when they are processing sensory representations. How can one say that there is not something it is *like* to be them if one argues that in non-zombies the intrinsic character contains qualia?

I believe that such an argument may arise from a confusion. To argue as one might above, that zombie worlds are impossible due to the intrinsic being identical with the phenomenal, one has to alter the proposed relationships so that the physical can modify the intrinsic. This is not what is happening. Instead, the component parts of any system (ultimately the elementary particles) retain their intrinsic characters. The system itself has an intrinsic character, *i.e.* what it is to be a brain, and it is this intrinsic character of which phenomenal consciousness is a part.

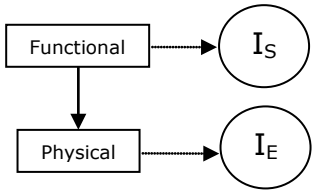


*Fig. 4.2 – Intrinsic character as properties*
Here, X⟶Y means that Y is a property of X.

What I am saying is that instead of a supervenience relationship, we should think of the $I_E$ as properties of physical entities, and $I_S$ as properties of functional systems. Thus, we begin to see a very clear property dualism emerge. At the most basic level are the intrinsic and physical properties of fundamental entities. Then we build systems, and we have the intrinsic and physical properties of those, the latter of which we can label 'functional'. Finally, we are left with the puzzle of the relationship between the two levels of intrinsic property. I believe that we can satisfy ourselves as to this point by reminding ourselves of the relationship between the physical and the intrinsic: the physical properties of an entity are the dispositional properties of that entity, and this entity also has an

intrinsic character. Therefore, if a system has physical properties above those of its constituent parts (and it must or we would not build systems), then it too must have an intrinsic nature: the intrinsic nature of the system as a whole. There is something it *is* for an object to be a computer, or a cat, that is distinct from non-functional entities such as a deactivated computer or dead cat.

What is most important to note is that it is conceivable, and therefore possible that zombie world could exist under intrinsic naturalism. It is a very difficult thing to conceptualise, since it would entail a world identical to ours with respect to its physical properties which differs with regards to intrinsic ones. Chalmers claims that our world could easily have turned out to be zombie-world, however in intrinsic naturalism, this is not the case. Copying all of the physical properties of one world into another requires a compatible intrinsic character. This *could* vary wildly, but we have no reason to speculate that it should do so, nor that the battery of intrinsic properties are infinitely co-variable (physical properties themselves are not; not all combinations of physical properties will create a functional universe). Instead, we might create a great many physical duplicates before one is created whose intrinsic properties do not happen to include qualia. Vulnerability to the conceivability argument is only entailed by denying the possibility of zombies, and intrinsic naturalism is not obliged to do this. Unlike Chalmers, however, I am forced to admit that zombie worlds are probably an extreme rarity. The hypothetical God would have had a lot more work to do to make this world a zombie world.

What of the principle which I outlined at the beginning of this chapter, wherein the intrinsic underlies the physical in a supervenience relationship between the two? In figure 4.2 I showed that we should think of the intrinsic properties of

physical entities as just those: properties. This does not preclude the idea of the intrinsic being the most fundamental level of reality; it is only a different way of conceptualising the same idea. I could just as relevantly have drawn the diagram showing the intrinsic nature of entities having physical properties (as in *Fig*. 4.3), but because physical properties are the ones with which we are most familiar, it was easier to communicate the relationships in the way it was done.

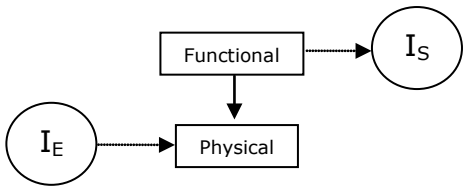


*Fig. 4.3 – Physical character as properties*
The same relationship as in *4.2,* seen in a different light. Here, I am equating the basic entity with its intrinsic properties, whereas in *3.2* I had equated it with its physical properties.

The intrinsic properties of systems, on the other hand, *do not* have this relationship with the functional properties of systems. This is one of the most difficult ideas to understand, however I believe that it holds. The relationship between the functional and intrinsic properties of *systems* sees the intrinsic dependent upon the functional (though only insofar that an intrinsic property is a property of the system). I shall characterise this relationship very shortly.

**Intrinsic naturalism and physicalism**

A crucial question now is that of intrinsic naturalism's compatibility with physicalism. This has been my objective from the beginning. Of course, by 'compatibility' I do not mean that one can still be a physicalist while holding intrinsic naturalism to be the case. I merely mean that one can retain a physicalist attitude insofar as physicalism is scientifically grounded (in all cases but consciousness), and the dominant paradigm for the explanation of all other natural phenomena.

The first way in which intrinsic naturalism is compatible with physicalism is that it does not conflict with existing notions of the physical. As previously discussed, we can preserve the causal closure of the physical realm while still acknowledging consciousness and other intrinsic properties of entities and systems. The idea that the physical is only the external manifestation of an objective world which we can never in principle access is by no means a new one, having been discussed most famously by Plato and Kant. Generally, the idea evokes something of a "ho, hum" reaction since if this objective reality cannot be accessed in principle then it is of little but academic relevance. In the case of consciousness, however, it is extremely relevant that there is a level of reality more fundamental than, and inaccessible to, the physical. The argument that there is such a level is valid and somewhat unfalsifiable, and this validity lends a certain strength to the argument that consciousness itself is non-physical. In the first chapter we saw that *a posteriori* physicalists are, for the most part, uncomfortable with consciousness, and their attempts to fit it into a physicalist picture of the world reflects this. If, however, one accepts the validity and possibility of the idea of an objective reality underlying the physical, then it does not seem a massive leap to attach consciousness to that reality. Note my use of the word 'objective' to describe the realm to which consciousness belongs. I mean objective in the sense that it is truly *there*, and cannot in principle be an illusion (the constituents of consciousness may be non-veridical, but consciousness itself certainly exists). Of course, it is still subjective in the sense of being accessible only by its possessor and incommunicable to anyone else.

One issue seems to remain *re* causal closure. How can the causal closure of the physical remain intact if I seem to be arguing that changes in the physical nature of a system can result in alterations to its intrinsic character? The simple answer

is merely to reply that the intrinsic character is a property of the system. When talking functionally, we may talk of a single system (such as the brain) processing multiple perceptual data and producing any number of images from any number of inputs. Physically, there is only the very slightest difference between my brain when viewing a red flower and my brain when viewing a blue one, so because we tend to label multiple things which are incrementally different as being the same in some capacity, we consider it the same system. Intrinsically, however, these systems are *different*. It is less that the physical change has brought about a *change* in the intrinsic character of the perceptual system, and more that an intrinsically new system now exists. In terms of personal identity, this means that one's intrinsic self is constantly being replaced by an incrementally (but in intrinsic terms, substantially) different self.

This may not satisfy everyone with regards to causal closure. Nevertheless, it is difficult to see why anyone would be so attached to the notion of the intrinsic character of systems that they would flinch at these characters being fleeting and often completely unique. One might argue that the difference between my consciousness now and when I move my head slightly to the right is trivial, and that the idea that this constitutes an altogether new system is overly elaborate. To this, I simply reply that in the intrinsic realm, any difference between one system and another makes the two systems distinct. In functional terms, and for reasons of convention, we tend to think of a computer displaying one image and then displaying another as the same system, but intrinsically the two are not the same. In short, the intrinsic realm is one in which two entities are either identical or distinct, and the notion of two non-identical systems being in some abstract way 'the same' is a nonsense.

Does intrinsic naturalism explain anything besides consciousness? Of course it does not. The presence or absence of intrinsic properties of entities on a more fundamental level than the physical does not impinge upon what we already know about the physical nature of the world, except to add a trivial bit of information about the underlying nature of entities. Since intrinsic naturalism does not alter the nature of the physical, it is perfectly compatible with science. In fact, as should be obvious by now, intrinsic naturalism does not require science to alter its model of anything. The intrinsic qualities of entities are never directly relevant to science, except in the instance of consciousness.

Contrast this with some versions of panpsychism which hold primitive entities to demonstrate certain psychological behaviours. Physicalism cannot co-exist with these theories, and they are not well grounded in scientific evidence. Similarly, substance dualism is not compatible with physicalism because it often requires us to reject the causal closure of the physical.

Compatibility with physicalism is extremely important, as it is rightly the dominant paradigm for the explanation of every phenomenon in nature. A naturalist theory is therefore highly preferable to one which would tamper with existing physical and scientific principles.

**Relationship to similar theories**

Intrinsic naturalism is a property dualist theory, however it also fits into a few other categories. I shall now explore the extent to which it is subject to the criticisms of these other theories.

The category into which I am primarily placing intrinsic naturalism is that of what Chalmers (2002) calls type-F monism or, more specifically, panprotopsychism. The properties which are protophenomenal in this instance are the intrinsic properties of fundamental entities. I have already argued that, unlike in panprotopsychism generally, these properties do not *combine* to produce consciousness, but instead consciousness is *among* the intrinsic properties of certain systems, the systems themselves having been formed through the combination of their constituents. All entities possess intrinsic properties, if not consciousness itself. This puts intrinsic naturalism in the same class of theory (according to Chalmers) as panpsychism, although without the latter's most significant disadvantages.

Another similar theory, or class of theories, to intrinsic naturalism is neutral monism. Though there are many variants, the position common to all neutral monist positions is that the world is comprised of only one substance. However, while a physicalist (or even an idealist) would agree with this, a neutral monist claims that this one substance is neither mental nor physical: it is *neutral*. This adjective is generally applied to the intrinsic character of the world, because most neutral monists would not wish to argue that there are *not* physical and mental properties of the world; merely mental and physical properties of a neutral substance.

It is easy to see the similarity between intrinsic naturalism and neutral monism; both theories argue that the world is inherently non-physical. Intrinsic naturalism also holds that the world is not intrinsically phenomenal, however since it holds that the phenomenal is contained within the battery of intrinsic properties while the physical is not, one could claim that intrinsic naturalism borders on

Berkeleianism. This, in fact, has been one of the criticisms of neutral monism, made by several philosophers (Lenin, 1970; Popper and Eccles, 1977; Nagel, 2000). In defence of intrinsic naturalism from the charge that it is idealist, I shall once again explain the precise relationships implied in intrinsic naturalism. Phenomenal consciousness is part of the intrinsic character of systems. The functional aspects of a system are supervenient upon the physical properties of its component entities, which are themselves supervenient upon the intrinsic properties of these entities (see *Fig. 4.2*). For intrinsic naturalism to be idealist, the physical (and mechanical, and functional *etc*) would have to be supervenient upon consciousness. This is clearly not the case. Another criticism of neutral monism is that it does not accommodate the mental at all, especially with regards to the subjective character of the mind. This criticism obviously does not apply to intrinsic naturalism, since consciousness (and with it subjectivity) arises not at the fundamental level but on a higher (but still intrinsic) level.

I have previously argued that consciousness itself does not have a causal role, or indeed any part in a causal chain. Yet I am also claiming a tie between the functionality of a system and its consciousness, so there seems to be some tension between these ideas. It seems then that intrinsic naturalism resembles parallelism, wherein the phenomenal and the physical coincide without there being any interaction between the two. Parallelism is usually only held by those who believe that the coincidental parallel between the phenomenal and the physical was set up by a supreme being of some sort. Clearly this is not a belief which accords with the scientific attitude of intrinsic naturalism, yet what intrinsic naturalism seems to argue for *is* a parallel between the physical and the phenomenal. Surely this seems unusual, for if intrinsic naturalism were a panprotopsychist theory such as that espoused by Chalmers, then the

phenomenal would be dependent to some degree upon the physical. Instead, the relationship between the functional and the phenomenal in intrinsic naturalism seems to show them running in parallel. In fact, the functional and the phenomenal *are* related more closely than this, and this is because the phenomenal aspects of a system belong to the intrinsic character of the system, and the system also has a functional aspect (see *Fig. 4.4*). These aspects are only separable in conceptual terms. One cannot change the intrinsic character of an object and still claim it to be the same object (my zombie twin and I are not identical *simpliciter*, in other words).

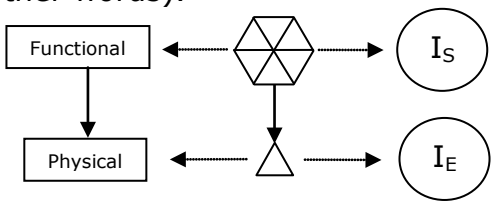


*Fig. 4.4 – Physical and intrinsic as part of a whole*
The same relationship as in *4.2* and *4.3*, shown with the system (hexagon) and its components (triangle) as the focal points of the diagram. This emphasises that both the physical and intrinsic properties are merely part of a whole.

**The knowledge argument**

How does intrinsic naturalism fare against the knowledge argument? This is more straightforward than my response the conceivability argument. Simply, Mary does learn something new, and this is no problem for intrinsic naturalism. In this model, Mary becomes, for the first time, a system which perceives red flowers. Therefore, she learns part of what it is to be such a system, namely, what it *feels like*. In contrast to physicalism, intrinsic naturalism holds as a central tenet that there are facts about the world separate to (and more fundamental than) physical facts. Thus, there is no contradiction in Mary knowing every physical fact about perception but lacking some other fact. In a zombie world, Mary would not learn this fact (though she might believe that she had). In short, the intrinsic naturalist position supports the conclusion of the knowledge argument, that physicalism must be rejected.

**The combination problem**

We seem to come up against a new combination problem, because I have argued that systems have an intrinsic character which simply comes about magically, quite separate from the intrinsic character of protons and so forth. To this, I merely reply that there is nothing magical about it. Protons have a different set of physical dispositions to quarks, even though the former is entirely comprised of the latter. There are epistemologically emergent facts in physics, and likewise in the realm of intrinsic characters. Three quarks combine to form a proton, which is held together by forces, and can be considered a particle. It can then combine with other similar particles to form an atom, which can be considered a particle (in that it is a discrete object). This can continue up to molecules, cells, organs, organisms and maybe beyond to planets and so on (the Universe itself is a particle, in that it is a discrete object and can be considered a single entity); these composite entities are discrete 'particles', regardless of being made of other particles. They have their own physical characteristics which come about through the combination of the characteristics of their constituents. Similarly, there is something it is to be a human, a cell, a molecule and so on all the way down to the quarks and leptons which ultimately comprise matter.

At this point, it should be emphasised that the possession of 'overlapping' properties is far less plausible when applied to consciousness itself (as in panpsychism). If one accepts the link for which I have argued between functional systems (specifically perceivers) and consciousness, then it is clear that my consciousness is not further attributable to my neurons. If one does not, then this implies panpsychism or a far more generous emergentism than mine, and one encounters the problems I have discussed throughout this dissertation.

**Intrinsic naturalism – the final picture**

In this chapter, I have shown aspects of the intrinsic naturalist position with regards to the relationships contained therein. This picture may seem a little unclear due to the emphasis placed upon different aspects of it at different stages of the chapter, so here I shall present the final picture as clearly as I am able.

The most basic entities in the Universe have both physical and intrinsic properties. The physical properties are those with which we are familiar, since it is these which determine how an entity will act in the world, and how it will interact with other entities. Beneath this dispositional level, however, there is the intrinsic character of each entity. This is what the entity *is*, and it cannot be accessed by science; by definition, detection is a form of interaction, and interaction is always physical.

Not only fundamental entities have an intrinsic character. Nucleons, atoms, molecules, cells and so on all have an intrinsic character which is separate from the intrinsic characters of their constituent parts. These intrinsic characters interact via the medium of the physical, which invokes another intrinsic character; that of the system itself.

Some systems process certain kinds of information in a particular way, making them perceivers. The intrinsic character of a perceiver includes phenomenal consciousness, which is only a part of the intrinsic character of the system. The qualia which comprise consciousness depend upon the physical changes brought about in the system.

# CONCLUSION

The purpose of this dissertation has been to put forward a theory which can account for phenomenal consciousness as a non-physical feature of the world while still preserving the useful features of a physicalist model.

I attempted to defend the assertion made by David Chalmers that consciousness is a surprising feature of the world when seen from a physicalist point of view. There have been several historical examples of phenomena which have been thought of as surprising, yet none have continued to be do as our physical knowledge has expanded. One might be tempted to suggest that consciousness will follow this pattern, however I have argued that it will not, and that a non-physicalist theory must be employed to accommodate a phenomenon that clearly exists.

That having been said, a physicalist picture *does* hold for everything else in the Universe (that we have discovered so far), or at least there do not seem to be any problems with it doing so in principle. Physicalist principles also stand at the centre of scientific and naturalist thinking about the world, so have enormous value when discussing issues such as morality, biology, causality and more exotic issues such as spirituality. Thus, if we cannot preserve physicalism when discussing consciousness, I believe it is desirable to preserve as much of this attitude as possible (lest we invite the 'magical thinking' which can lead to substance dualism). My aim, therefore, has been to extract the best of physicalism.

After showing initially that physicalism seems unsatisfactory, I explored pansychism. While panpsychism has its merits, it has crucial problems such as its tendency to contradict or impinge upon established physical principles. Next, I discussed physicalism in more detail, exploring the particular advantages and disadvantages of a physicalist position. I concluded that what seems to motivate a physicalist (including a desire for a science-friendly position on consciousness) can be preserved even in a non-physicalist theory.

In the final chapter, I argued for my model: intrinsic naturalism. This model includes features from physicalism, neutral monism, and panprotopsychism. It appeals to the intrinsic character of fundamental entities and systems. I argued that consciousness is a component of the intrinsic character of certain (perceptual) systems, and that although a system's physical (and functional) characteristics supervene upon the physical characteristics of the component entities, and that these in turn are supervenient upon the entities' intrinsic characters, nothing supervenes upon consciousness. Thus, intrinsic naturalism is *not* a form of idealism.

I showed that intrinsic naturalism could withstand the knowledge and conceivability arguments, as well as the combination problem as applied to panpsychism. It can also accommodate certain phenomenal states (or lacks thereof) which I have argued panpsychism to be unable to do, at least without elaborate arguments.

## REFERENCES

Blackmore, S. (2003) *Consciousness: An Introduction.* London: Hodder Education

Bonnet, M.H. & Johnson, L.C. (1978) Relationship of arousal threshold to sleep stage distribution and subjective estimates of depth and quality of sleep. *Sleep,* 1: 161-8

Brogaard, B. (2010) Strong representationalism and centred content. *Philosophical Studies,* 151: 373-392

Chalmers, D.J. (1996) *The Conscious Mind*. Oxford: Oxford University Press

Chalmers, D.J. (2002) Consciousness and its Place in Nature.
In: Chalmers, D. (Ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: Oxford University Press

Chalmers, D.J. (2004) The Representational Character of Experience.
In: Leiter, B. (Ed.) *The Future for Philosophy*. Oxford: Oxford University Press

Chalmers, D.J. (2007) Phenomenal Concepts and the Explanatory Gap.
In: Alter, T. & Walter, S. (Eds.) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* New York: Oxford University Press

Conee, E. (1994) Phenomenal Knowledge. *Australasian Journal of Philosophy,* 72: 136–50

Crick, F. (1994) *The Astonishing Hypothesis.* New York: Scribners

Crick, F. & Koch, C. (1990) Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences,* 2: 263–75

Dennett, D.C. (1991) *Consciousness Explained*. London: Penguin

Dennett, D.C. (1998) *Brainchildren: Essays on Designing Minds.* Cambridge, MA: MIT Press

Fair, D. (1979). Causation and the Flow of Energy. *Erkenntnis* 14: 219-250.

Jackson, F. (1982) Epiphenomenal Qualia. *The Philosophical Quarterly*, 32: 127-136

Kisley, M.A. & Cornwell, Z.M. (2006) Gamma and Beta Neural Activity Evoked During a Sensory Gating Paradigm: Effects of Auditory, Somatosensory and Cross-Modal Stimulation. *Clinical Neurophysiology,* 117: 2549–63

Lenin, V.I. (1970) *Materialism and Empirio-criticism: Critical Comments on a Reactionary Philosophy*. (Reprint.) New York: International Publishers

Lewis, D. (1988) What Experience Teaches. *Proceedings of the Russellian Society*, 13: 29–57

Mistlberger, R.E. (2005) Circadian regulation of sleep in mammals: role of the suprachiasmatic nucleus. *Brain Research Reviews,* 49: 429–54

Nagel, T. (1979) Panpsychism.
In Nagel T (Ed), Mortal Questions. Cambridge: Cambridge University Press

Nagel T (2000) "The Psychophysical Nexus"
In: Boghossian, P. & Peacocke, C. (Eds.) *New Essays on the a Priori.* Oxford: Oxford University Press, 434-72

Nemirow, L. (2007) So This is What It's Like: A Defense of the Ability Hypothesis
In: Alter, T. & Walter, S. (Eds.) *Phenomenal Concepts and Phenomenal Knowledge. New Essays on Consciousness and Physicalism.* Oxford: Oxford University Press, 32-51

Popper, K.R., Eccles, J. (Eds., 1977) *The Self and Its Brain*. New York: Spinger-Verlag

Revonsuo, A. (1999) Binding and the phenomenal unity of consciousness. *Consciousness and Cognition,* 8: 173-85

Russell, B. (1927) *The analysis of matter*. London: Kegan Paul

Taylor, E. & Rutter, M. (2002) *Child and adolescent psychiatry*. Oxford: Blackwell Science

Stoljar, D. (2001) The Conceivability Argument and Two Conceptions of the Physical. *Philosophical Perspectives*, 15: 393–413

Strawson, G. (2008) The identity of the categorical and the dispositional. *Analysis,* 68: 271-82

Yablo, S. (1992) Mental Causation. *The Philosophical Review*, 101: 245–80