MONASH University

The Promise of Panpsychism: Understanding Integrated Information Theory as a panpsychist theory of mind.

Henry Dobson
Bachelor of Arts (Honours)

A thesis submitted for the degree of Master of Arts at
Monash University in 2016
Monash University

Copyright notice

Notice 1

Abstract

In contemporary philosophy of mind many philosophers regard physicalism to be the most promising theory of mind. In an attempt to defend physicalism against arguments such as *the knowledge argument, the conceivability argument,* and *the explanatory gap argument*, some philosophers have *broadened* the view. But in doing so they modify physicalism is such a way that it runs the risk of becoming more a form of panpsychism. If modifying physicalism has this effect then we have good reasons for taking panpsychism seriously. In my thesis I examine three contemporary versions of panpsychism, namely *pan-experientialism, pan-phenomenalism* and *pan-protopsychism*. All panpsychist theories must deal with what is called the 'combination problem'. There are many different combination problems and I argue that both pan-experientialism and pan-phenomenalism fail to resolve their respective combination problems. And because pan-protopsychism is yet to be developed into a philosophical theory, it is yet to be seen whether it can do better than other forms of panpsychism. I turn my attention to a more recent theory of consciousness which is said to have panpsychist implications, namely *Integrated Information Theory (IIT)*. I demonstrate that IIT can be thought of as a version of pan-protopsychism. I then identify two combination problems that confront IIT as a pan-protopsychist theory and present solutions to these problems. My purpose in writing this thesis is to present a philosophical investigation into how IIT is to be understood as a panpsychist theory of mind. I argue that it has the potential to resolve the relevant combination problems and I therefore believe that as a panpsychist theory IIT is the most promising theory of mind to date.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Acknowledgements.

I acknowledge and would like to thank my primary supervisor, Dr. Monima Chadha, for her continuous support, guidance and sage advice throughout the writing of this thesis. I greatly appreciate all your help and assistance over the past two years.

I also acknowledge and would like to thank my secondary supervisor, Dr. Jakob Hohwy, for his support and assistance, particularly in respect to the more technical aspects of this thesis.

I must also thank all my family and friends who have continued to love and support me throughout my time writing this thesis, especially my Mum and Dad – thank you.

The Promise of Panpsychism: Understanding Integrated Information Theory as a panpsychist theory of mind.

Thesis by Henry Dobson, Monash University

Table of Contents

Thesis Introduction

In this thesis I investigate the philosophical aspects and implications of *Integrated Information Theory (IIT)* and I argue that IIT is best understood as a panpsychist theory rather than a physicalist theory of mind.

In contemporary philosophy of mind many philosophers believe the most promising theory of mind is *physicalism*; the traditional thesis that *everything is physical* or the more contemporary thesis that *everything supervenes on the physical*. While the world around us – *i.e.* the universe and everything in it – certainly appears physical, not all philosophers accept the view that *everything is physical*, especially philosophers of mind. It is in philosophy of mind therefore that we find the most critical objections and challenging arguments against physicalism.

My aim in this thesis is to demonstrate why I believe panpsychism is a more promising theory of mind than physicalism. In order to demonstrate this I will begin in chapter one by presenting a general analysis of physicalism, detailing how the doctrine has been understood in the traditional forms of *reductive* and *non-reductive physicalism* and in the more contemporary form of *supervenience physicalism*. Following this I present what I think are three critical arguments against physicalism, the first of which is Frank Jackson's *knowledge argument,* secondly David Chalmers' *conceivability argument* involving philosophical zombies, and thirdly Joseph Levine's *explanatory gap argument*. I believe that these three arguments when taken together raise serious philosophical concerns that ultimately leave physicalism in a good deal of doubt. In light of these three arguments against physicalism I therefore believe that we have very strong reasons for considering alternative philosophies of mind, particularly panpsychism.

Why panpsychism? One reason why I think panpsychism is a promising theory has a lot to do with how some philosophers have attempted to *broaden* physicalism for the purpose of defending it against the three arguments as

mentioned above. One such example is given by Daniel Stoljar (2001) in his paper *Two Conceptions of the Physical*. In this paper Stoljar presents what he calls *object-based physicalism* (otherwise called *o-physicalism*). Developed as a defence against Frank Jackson's knowledge argument, o-physicalism is a broader conception of the physical in that it views qualitative properties (qualia) as being *categorical* properties. Viewed this way, qualia are taken to be intrinsic to all paradigmatic physical objects. The implication here is that some fundamental or base properties are true mental properties, specifically qualia. Stoljar is aware, however, that by viewing qualia as fundamental properties o-physicalism is at risk of collapsing into either neutral monism or panpsychism. To reject the claim that o-phyicalism is a form of panpsychism Stoljar argues that qualia (as categorical properties) *supervene* on some combination of other non-qualitative/physical categorical properties, thus making o-physicalism a form of supervenience physicalism. That qualia supervene on other non-qualitative/physical properties is, in my view, a tentative metaphysical assumption. I argue that the only way for Stoljar to save o-physicalism from collapsing into panpsychism is for this supervenience relation to be necessarily true. I am critical of this supervenience relation in o-physicalism because in my view categorical properties are more likely to have *equal status* with one another, in which case qualia are inter-dependent of other categorical properties and are therefore equally fundamental with all categorical properties. If equal status is true in respect to the ontological organisation of categorical properties then Stoljar's o-physicalism appears to collapse into panpsychism.

Also in chapter one I consider David Chalmers' (1996) *naturalistic dualism*, which he develops in light of the conceivability argument against physicalism and the possibility of philosophical zombies. Though dualist in name, Chalmers claims that his naturalistic dualism is more a form of monism but one that is *broader* than materialistic monism. Chalmers provides little detail as to how his dualism is to be understood as a form of monism, but I argue that if his naturalist dualism is similar to Stoljar's o-physicalism in the sense of viewing qualia as categorical properties, then his naturalistic dualism/monism might also collapse into panpsychism. Thus, in light of Stoljar's *o-physicalism* and Chalmers

naturalistic dualism (monism), such attempts to broaden physicalism appear to run very close to panpsychism. I conclude chapter one by arguing that if broadening physicalism has this effect then we have very good reasons for taking panpsychism seriously.

In chapter two I turn my attention to contemporary panpsychism and I examine three recent panpsychist theories of mind. These are: *pan-experientialism* as developed by Galen Strawson (2006); *pan-phenomenalism* as developed by Sam Coleman (2012, 2013); and *pan-protopsychism* as described by David Chalmers (2013a). Chapter two is therefore divided into three sections, with each section devoted to one of the aforementioned versions of panpsychism. Within each section I will explain how each respective version characterises the essential and fundamental nature of mental properties.

Furthermore, each version of panpsychism must deal with what is called the *combination problem for panpsychism*. Broadly speaking, the combination problem involves having to explain how certain micro-mental properties at the fundamental level are able to combine into more complex macro-mental properties. Strawson's *pan-experientialism* faces one of the most challenging combination problems called the *subject-summing problem*. This is also referred to as the micro-subject/macro-subject gap. Of these three versions of panpsychism, only pan-experientialism faces the subject-summing problem for the reason that it is the only version of panpsychism that posits micro-subjects at the fundamental level. As I will show, there are strong arguments demonstrating that subjects cannot combine and therefore the subject-summing problem cannot be resolved. Two arguments that I will focus on are Philip Goff's (2009) epistemological argument and Sam Coleman's (2013) metaphysical argument. Because pan-experientialism fails to resolve the subject-summing problem I argue that it is a philosophically implausible version of panpsychism.

Pan-phenomenalism and pan-protopsychism both face different combination problems. Because neither version posits the existence of fundamental micro-subjects the problem they face therefore is that of explaining how non-subjective

mental properties can combine into subjective properties. This is otherwise referred to as the *non-subject/subject gap.* For pan-phenomenalism, this involves having to explain how non-subjective phenomenal properties are able to combine so as to yield subjective conscious experience. Sam Coleman presents one solution to this gap by developing a form of functional representationalism; this sees the phenomenal properties representing themselves within a central/perceptual domain, this being the domain of subjective experience. I argue that Coleman's solution fails to successfully bridge the non-subject/subject gap for the following two reasons: firstly, it presumes the existence of this central/experiential subjective domain, and secondly, phenomenal representation alone is not metaphysically sufficient for yielding subjective experience. However, by tightening the association between phenomenal representation and the subjective domain, I think pan-phenomenalism can be strengthened such that it can resolve the non-subject/subject gap more convincingly.

David Chalmers has recently developed pan-protopsychism into what he calls *pan-protophenomenalism*. This version holds that fundamental physical entities have protophenomenal properties, and when arranged in the right structure protophenomenal properties *a priori* entail phenomenal properties. Pan-protopsychism therefore runs up against the following two combination problems: i) the proto-phenomenal/phenomenal gap, and ii) the non-subject/subject gap. It is worth noting that Chalmers (2013a, 2013b) does not aim to provide solutions to the combination problems facing pan-protopsychism. Instead, his papers focus more on presenting the logical framework relevant to each version of panpsychism, as well as highlighting the specific combination problems that confront each version as a result. He also suggests what he think is the best strategy for resolving these problems but does not present any solution as such. Chalmers explains that in order for pan-protopsychism to be a successful theory it must first and foremostly deal with the protophenomenal/phenomenal gap. My aim in this thesis is to present two philosophical solutions for pan-protopsychism, which I will present in chapter four where I focus on and develop IIT into a form of pan-protopsychism.

In chapters three and four I turn my attention directly to IIT as a theory of consciousness. In chapter three I focus on the following two things. Firstly, I introduce IIT by presenting a basic framework of the theory, in which I articulate the phenomenological axioms of consciousness – *i.e. intrinsic existence, composition, information, integration, exclusion* – and how these axioms correlate with the physical postulates as stated in IIT. Through this basic framework I will highlight the important philosophical aspects and implications of the theory, such as how IIT identifies consciousness as integrated information, the notion of *qualia-space*, and the distinction between small-phi 'φ' and high-phi 'Φ' complexes.

After presenting this basic framework I then consider IIT as a physicalist theory of mind. IIT is empirically rigorous and has been developed in accordance with the latest neuroscience and psychological evidence regarding brain activity. Due to this IIT is by first approximation a physicalist theory of mind. By viewing IIT as a physicalist theory I look closely at how IIT stands up to and deals with the arguments against physicalism as featured in chapter one. This will firstly involve examining IIT as a *realisation* theory of mind, which will include an analysis of what IIT considers to be a "minimally conscious" photodiode. Secondly, and somewhat surprisingly, IIT claims that it is possible to build "zombie" systems. These zombie systems are not philosophical zombies, however. I will compare and contrast philosophical zombies with what I will call *IIT zombies* and then examine how these two zombies differ from one another. I also argue that even if IIT is true philosophical zombies are still conceivable, which means that IIT is susceptible to the conceivability argument. I present a conceivability argument against IIT and then consider how Tononi might respond to it. Thirdly, I will look at Tononi's response to the knowledge argument and how he appeals to IIT in respect to what we can know about consciousness. I also consider whether or not Tononi's response is considered successful in defending IIT against the knowledge argument. Because IIT fails to conform to the standard physicalist theories of behaviouralism, dualism, classical identity theory, functionalism and computationalism, I therefore conclude

chapter three by claiming that IIT cannot be considered a physicalist theory of mind.

What I find most interesting about IIT is that it claims to have panpsychist implications. In chapter four I focus directly on understanding IIT as a panpsychist theory of mind. I begin this chapter by highlighting exactly what motivates IIT towards panpsychism. These panpsychist implications can be found in the following three claims: that consciousness *is* integrated information; that integrated information is a *fundamental quantity* (as fundamental as mass, charge and energy); and that integrated information is an *intrinsic property* of physical systems. Stating that integrated information is both a fundamental and intrinsic property implies that consciousness itself is a fundamental and intrinsic feature of the world. Thus, IIT does have panpsychist implications.

In chapter four I begin by comparing and contrasting IIT against the three versions of panpsychism as featured in chapter two. Given that IIT identifies and describes consciousness as integrated information, I explain why IIT does not conform to either pan-experientialism or pan-phenomenalism. Because the information postulate in IIT states that information is both *causal* and *intrinsic*, I argue that IIT conforms best to pan-protopsychism. To explain this in more detail I appeal to Adam Barrett's (2014) *Field Integrated Information Hypothesis (FIIH)*. In his FIIH Barrett claims that truly intrinsic information exists only within fundamental fields. I give a detailed exposition of Barrett's FIIH to explain how intrinsic information can be seen as having the *potential* to yield phenomenal consciousness. I take issue with Barrett's definition of "potential consciousness" and I present an alternative definition regarding this notion. I believe that my definition better suits IIT in respect to understanding it as a form of pan-protopsychism.

As a form of pan-protopsychism IIT confronts the following two combination problems: i) the protophenomenal/phenomenal gap, and ii) the non-subject/subject gap. In the final two sections of chapter four I present two solutions to these combination problems. To resolve the

protophenomenal/phenomenal gap I appeal to Barrett's FIIH to demonstrate why fundamental fields qualify as protophenomenal properties. I then explain how the intrinsic information in fundamental fields can be seen to *a priori* entail phenomenal consciousness within the *qualia-space* of a phi-complex with Φ^{MAX}.

In the final section of chapter four I develop IIT into a form of *emergent panpsychism* as a means of resolving the non-subject/subject gap. This involves analysing how IIT distinguishes between small-phi 'φ' and high-phi 'Φ' complexes. Central to this distinction is how small-phi complexes are said to be non-subjective in nature, whereas high-phi complexes are said to be subjective in nature. Hence IIT faces a non-subject/subject gap between small-phi and high-phi complexes. I attempt to resolve this gap by demonstrating how subjectivity can be seen to emerge as an ontologically novel macro property in high-phi complexes only.

The panpsychist implications in IIT have received very little philosophical attention in recent literature. My purpose in writing this thesis is to present a philosophical investigation into how we can better understand IIT as a panpsychist theory of mind. By demonstrating that it can successfully resolve the combination problems relevant to it, my ultimate aim in this thesis is to show that IIT succeeds as a panpsychist theory of mind, thereby making panpsychism a promising theory of mind.

CHAPTER ONE

1.0 – Introduction

In contemporary philosophy of mind the prevailing theory is *physicalism*, the traditional thesis that *everything is physical*, or the more contemporary thesis that *everything supervenes on the physical*. The latter thesis is otherwise known as supervenience physicalism.

Many philosophers regard physicalism as the most promising theory of mind, however it is not without controversy. As Daniel Stoljar writes, "while physicalism is a thesis we have overwhelming reason to believe, believing it without qualification is no easy matter. For physicalism is on the face of it incompatible with, or at least in some tension with, various claims that are central to ordinary or common sense views about humans and what they are like, views which in various ways are presumed also in the sciences."[1] The claims that Stoljar refers to are rather straightforward acknowledgments about some of the most striking features of what it is to be human: features such as being *conscious*, thinking *rationally*, having *emotions* and *feelings* in the form of wants, beliefs, desires, fears and so on. There are other mental capacities such as the ability to *understand* and *comprehend* abstract mathematical truths such as 2+3=5. Everyone is intimately familiar with these features and there is no denying how essential and important they are in constituting human mentality. Yet understanding how the mind is incorporated within the physical body still remains a mystery, and while much of what we do know about the world is largely, if not fully understood in terms of its physical nature, the nature of mind and consciousness continue to escape physical explanation. It is in philosophy of mind therefore that we find the most challenging and indeed persuasive objections against physicalism.

These objections traditionally target the general physicalist methodologies of i) *reduction*, that of reducing mental properties to physical properties; and ii)

[1] Stoljar, D. (2010), p.13.

supervenience, the view that mental properties supervene on physical properties. It is widely accepted that these two philosophical methods – reduction and supervenience – go hand in hand, and that to some extent reduction involves or entails supervenience. To see this, if a set of properties (*A*-properties) reduces to another more basic set of properties (*B*-properties) then it is said that *A*-properties supervene on *B*-properties. There is, however, disagreement over whether physicalists need to be reductionists, and whether supervenience always involves reduction (*i.e. A*-properties might supervene on *B*-properties but that *A*-properties do not ontologically reduce to *B*-properties). To see how these arguments and objections work against physicalism it is necessary to consider both *reductive and non-reductive physicalism* and *supervenience physicalism* respectively. In this chapter I will consider both these versions of physicalism and afterwards I will examine three critical arguments against physicalism, namely the knowledge argument, the conceivability argument, and the explanatory gap argument. In light of these arguments my purpose in this chapter will be to show that we have strong reasons for taking other philosophical views more seriously, particularly panpsychism.

1.1 – Reductive and Non-Reductive physicalism.

Reductionism has been developed in a variety of forms (See Searle 1992 for five different senses of 'reduction'). According to Searle, the most important form of reduction is *ontological reduction*.[2] This form of reductionism holds that certain things in the world are, to use Searle's phrase, *nothing but* certain sorts of other things. One philosophical movement to which ontological reduction applies is that of *logical behaviourism*.

Developed in large part by Gilbert Ryle (1949) behaviourism claims that, contrary to dualism, mental states are not something ontologically distinct from the physical body and behaviour. Rather, behaviourism proposes that mental-conduct concepts such as 'being in pain' can be logically assimilated with behavioural-conduct concepts, such that mental states can be fully captured in

[2] Searle, J. (1992), p.113.

terms of physical behaviour and dispositions to behave. According to behaviourism, there is nothing *mental* over and above physical behaviour: the mental is the physical, specifically physical behaviour and dispositions to behave.

One uncomfortable consequence of behaviourism, which I take to be a definite failing of the theory, is that it appears to *eliminate* mental states, and by extension eliminates the mind altogether. As Campbell writes, "Behaviourist theory has no place for mental objects. Sometimes men are in pain, but this does not mean that there are things called "pains" which they have, feel, or are in."[3] But this of course seems intuitively wrong due to the simply fact that we do feel things – people do *feel* pains! The objection that follows from this is that behaviourism fails to provide a phenomenological account of the feelings and sensations associated with our every-day behaviours. On conceptual analysis, pain, says the behaviourist, is nothing more than pain-behaviour or the disposition to behave in such a way. However, nothing about our behaviour or dispositions to behave can explain why it is that, as a matter of fact, such behaviours are accompanied by certain feelings and sensations. Moreover, human behaviour can be very misleading: people can act and behave as though they are in pain when in fact they are not, and in some more strange cases, some pains can be a form of pleasure. Pain-behaviour does not therefore always and directly correlate to mental states such as *being in pain*. Thus, behaviourism as a form of reductionism fails to adequately explain the phenomenological nature of mental states. To account for mental feelings and sensations J.J.C. Smart proposed another reductionist view in the form of mind/brain identity theory.

Smart proposed the thesis that "sensations are brain processes".[4] Similar to the attempt made by Ryle to capture mental states in terms of physical behaviour and dispositions, Smart proposed that when it comes to understanding the nature of mental sensations there is no need to assert anything "over and above" the physical activity of the brain itself. According to Smart, there is no need to *correlate* brain states with anything else, namely immaterial mental states.

[3] Campbell, K. (1970), p. 61.
[4] Smart, J.J.C. (1956), p. 144.

Instead, mental states *are* just brain states. Smart appealed to Occam's razor to argue that mind/brain identity theory is ontologically parsimonious when compared to other theories of mind, particularly those that assert irreducible or non-physical mental properties. However, not all philosophers agree that appealing to notions such as "parsimony" provides a theory with any preferential weight. Jaegwon Kim is particularly critical of such philosophical manoeuvres, writing that:

> Reductionisms, we tend to feel, attempt to impose on us a monolithic, straight-jacketed view of the subject matter, the kind of cleansed and tidy picture that appeals to those obsessed with orderliness and discipline. Perhaps, this impression has something to do with the reductionists' ritual incantations of slogans like "parsimony", "simplicity", "economy", and "unity", all of them virtues of a rather puritanical sort… In fact, the word "reductionism" seems by now to have acquired a negative, faintly disreputable flavour – at least in philosophy of mind. (1993, p. 266)

This monolithic, straight-jacketed view as Kim describes it is, I think, perhaps best exemplified by *psycho-neural identity theory*. This identity theory claims that mental states are identical with neural states of the brain. Given Smart's thesis that mental sensations are brain processes, the psycho-neural identity claim that follows from this is that the mental sensation of pain *is* the physical process of, say, C-fibres firing: in short, pain *is* C-fibres firing. While it is true that pain involves the firing of C-fibres for human beings, one issue with this identity theory is that it appears *chauvinistic*; it restricts pain only to those creatures who possess C-fibres. The consequence of this is that all other organisms without C-fibres cannot *ex hypothesi* feel pain. Again, this appears intuitively wrong since we all to readily ascribe pain to other animals and organisms, all of whom are in various ways physiologically unlike us. Moreover, it is conceivable that alien creatures, whose physical constitution is entirely different from our own (*i.e.* without C-fibres), would also be capable of feeling pain. The possibility of this illustrates how mental states such as pain can be multiply realised, and this leads

to another major argument against psycho-neural identity theory, this being the *multiple realization argument*.

The multiple realisation argument asserts that particular types of mental states such as pain can be multiply realised and experienced in many non-identical physical systems. For example, an alien organism with a silicon brain, one that is completely different from our own carbon-based brain, might nonetheless feel pain in much the same way that we do. In this case the physical basis of pain is *not* a neurological state like C-fibres but is instead something quite different; for an alien, pain might be identical to Silicon fibres (or what might be called S-fibres) or it might be something radically different altogether. What the multiple realisation argument plausibly highlights is that pain cannot be reduced to and identified with one specific neuro-physiological state such as C-fibre stimulation. Instead, there could well be multiple physical states that *realise* pain, including all other mental states for that matter. The multiple realisation argument had a critical impact in philosophy of mind in the late 1960's, so much so that physical reductionism, including identity theory, declined in popularity.

As reductive physicalism declined in popularity throughout the 1970's, non-reductive physicalism became the more promising alternative. Non-reductive physicalism holds that mental properties are indeed physical but need not be reduced to the more basic physical properties as specified by the physical sciences. An important distinction between reductive and non-reductive physicalists is that the latter tend to be mental realists; non-reductionists believe in the reality of the mind and that mental state are genuine properties of physical systems.

Although non-reductive physicalism denies that mental states need to reduce to physical states, it nevertheless asserts that mental states are closely and intimately related to physical states, so much so that mental states are said to *supervene* on physical states. Non-reductive physicalism is therefore generally understood in the form of supervenience physicalism.

1.2 – Supervenience physicalism

Whereas traditional physicalism is the thesis that *everything is physical*, supervenience physicalism is the thesis that everything *supervenes* on the physical. The notion of supervenience plays an important technical role in analytic philosophy because it is used as a method for providing explanatory power or for demonstrating logical entailments within a given conceptual framework. Since physicalism holds that everything in the universe is physical, and physical as a matter of fact, supervenience physicalism holds that the most fundamental physical facts determine all other higher-order facts, *e.g.* biological facts, technological facts, sociological facts, etc. A general definition of supervenience is given by David Chalmers who writes that, "supervenience is a relation between two sets of properties: *B* properties – intuitively, the *high-level* properties – and *A*-properties, which are the more basic, *low-level* properties."[5] Chalmers provides the following template for defining supervenience:

> B-properties *supervene* on A-properties if no two possible situations are identical with respect to their A-properties while differing in their B-properties. (*Ibid*)

Because supervenience plays an important role in philosophy of mind I therefore think it is worth explaining here the various ways in which it works, for this will provide some background detail and clarity for when I come to examining the arguments against physicalism further on in this chapter.

Chalmers distinguishes between more precise notions of supervenience: these being *local* and *global supervenience*, and *logical* and *natural supervenience.* In respect to local and global supervenience, the notion of local supervenience relates to the properties of *individual* objects, such that any two possible objects with the same *A*-properties will possess the same *B*-properties. For example, *length* supervenes locally on physical objects: any two objects with identical physical properties will be necessarily identical in length. In contrast, global

[5] Chalmers, D. (1996), p. 33.

supervenience holds that *B*-facts supervene *globally* on *A*-facts when the *A*-facts are facts about the entire *world*, or be it the whole universe. What this means is that for all worlds identical in terms of their *A*-properties, their *B*-properties will all be identical too.

In respect to *logical* and *natural* supervenience, logical supervenience (also known as conceptual supervenience) holds that *B*-properties supervene *logically* on *A*-properties when there are no two logically possible situations where the *A*-properties are identical but the *B*-properties are different. Logical supervenience will become important in chapter three when I focus on how IIT might defend itself against the conceivability argument and the possibility of what I will call *IIT zombies*. One way to think about logical supervenience is in respect to the concept of aunt-hood. The property of being an aunt supervenes *inter alia* on the property of being female but not on being male. This means that there is no possible world in which there exist *male aunts*, for the idea of a male aunt is contradictory because the concept of aunt-hood applies only to females and not males. Thus, given the concepts of male, female, uncle and aunt, male aunts are not logically possible people simply because aunt-hood supervenes on females only.

Natural supervenience differs from logical supervenience insofar as two sets of properties are seen as being perfectly *correlated* in the natural, empirical world. Chalmers uses a mole of gas to exemplify this, stating that, "it is empirically impossible that two distinct moles of gas could have the same temperature and volume, but different pressure. It follows that the pressure of a mole of gas supervenes on its temperature and volume in a certain sense."[6] An important difference between logical and natural supervenience is that while natural supervenience is constrained by the laws of nature, logical supervenience is not. The distinction between logical and natural supervenience is most important in philosophy of mind, for it is reasonable to think, as many philosophers do, that mentality should supervene on the physiology of the brain. This leads to one

[6] Chalmers, D. (1996), p. 36.

more version of supervenience that I want to mention, this being the notion of *psychological supervenience*.

Ardent believers in physicalism maintain the view that mental states supervene on the physical states or processes of the brain. All psychological properties (mental state, events, processes, etc.) are therefore said to supervene upon physical properties (physical states, events, processes, etc.). Jaegwon Kim states that psychological supervenience is *asymmetric*: "the physical determines the psychological, but the psychological does not determine the physical."[7] To say that the physical determines the mental is also to say that the physical *causes* the mental or that psychological states are causally determined by physical states. This is true to a large extent, *e.g.* the physical state of the environment will influence (*i.e.* causally determine) how one feels within that environment. But as we shall see further on in this chapter, there are strong philosophical arguments against psychological supervenience, namely the knowledge argument and the conceivability argument.

One of the distinctive features about non-reductionism is that mental properties are taken to be causally relevant and efficacious. That is, mental properties are seen to play a causal role within the overall functioning of physical systems. As a precursor to *functionalism*, David Lewis developed a causal theory of mind that accepts psycho-neural identity theory but defines mental states in terms of causation; *i.e.* cause and effect. On Lewis' view:

> The definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states posses the definitive characteristics of experience, they must be the experiences. (1966:16)

[7] Kim, J. (1993), p. 176.

Central to Lewis' identity theory is the recognition that conscious experiences and mental states are in fact *real* and that their reality is a matter of physical causation. On Lewis' view, mental states do not reduce to any physical property *per se* but rather, mental states and experiences are understood in terms of their causal role, which according to Lewis "is expressible by a finite set of conditions that specify its typical causes and its typical effects under various circumstance."[8] Lewis' causal theory rules out other philosophical theories of mind such as epiphenomenalism, parallelist dualism and behaviourism because these theories all deny the causal efficacy of mental states and experiences.

I want to briefly consider the doctrine of functionalism here for it will become more relevant further on in respect to the 'explanatory gap' argument against physicalism. Functionalism holds that mental states depend not on the physical constitution of a given system *per se* but are instead defined by the causal role they play within the overall functioning of a system. According to functionalism, the mental state of *pain* for instance can be defined by its functional role within a system's overall causal network involving sensory input, motor output and other psychological states (whether conscious or unconscious). A functionalist explanation of pain might be as follows: pain *is* a state that is typically caused by physical injury to the body; a state that produces beliefs and desires about those causes (*e.g.* that such pain-inflicting causes should be avoided); and has the effect of producing pain-instantiated behaviours such as yelling or crying. Similar to Smart's identity theory, pain, says the functionalist, is not something "over and above" all the physical processes and cognitive on-goings in the brain but rather, pain *is* an effect of all the physical processes that constitute the overall functioning of the brain. Functionalism is regarded by many philosophers today as one of the most promising theories of mind, but as we shall see further on, there are objections against it.

So far in this chapter I have considered how physicalism has been defined in the form of i) reductive and non-reductive physicalism, and ii) supervenience physicalism. I have also briefly considered the notions of identity theory and

[8] Lewis, D. (1966), p. 17.

functionalism because these notions are raised within and challenged by particular arguments against physicalism. I will now turn my attention towards these arguments, the first of which is the knowledge argument against physicalism.

1. 3 – The Knowledge Argument against physicalism

The aim of the knowledge argument is to refute physicalism. There are many versions of the knowledge argument but Frank Jackson's version is certainly one of the most well known. This is because, as Stoljar and Nagasawa explain, "Jackson's argument provides a much better illustration of the knowledge intuition than other thought experiments."[9] Jackson firstly articulates the knowledge intuition in the following passage:

> Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky. (1982, p. 127)

What the knowledge intuition intends on showing is that knowledge of the physical kind is not sufficient for knowledge of the phenomenal kind. That is, physical knowledge does not logically entail knowledge about phenomenal consciousness. The knowledge intuition therefore presents us with a *prima facie* modal truth: that it is possible to know everything physical about the world while not knowing any phenomenal truths, these being truths about specific conscious experiences and what it is like to have such experiences.

[9] Ludlow, P., Nagasawa, Y., & Stoljar, D. (2004), p. 9.

Many philosophers, mainly physicalists, have rejected the knowledge intuition on the grounds that it does not present conclusive evidence against physicalism; their general defence is that the intuition does not in actual fact strike upon nor reveal any fundamental truth (or falsity) about the physical world, thus claiming to save physicalism from the knowledge argument. To turn the intuition into a more substantial argument Jackson introduces us to Mary, a brilliant neuroscientist who has lived her entire life in a black and white room and who has learnt everything through black and white books and a black and white television. Put simply, Mary has only ever seen the world in black and white. Having specialised in neuroscience, Mary knows all there is to know about the neuro-physiology of the brain. Mary knows how the optical system works when certain wave-lengths of light stimulate the retina and subsequently that different words such as 'red' and 'blue' are ascribed to particular objects when specific neural networks are activated as a result of, say, looking up at the sky or at ripe tomatoes. For all intents and purposes, Mary knows all there is to know about the world in terms of its physical nature. Given this scenario, Jackson then considers what would happen to Mary when released from her black and white room for the first time. Jackson asks: "Will [Mary] *learn* anything or not? … It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that the previous knowledge was incomplete. But she had *all* the physical information. *Egro* there is more to have than that, and physicalism is false."[10] Jackson's argument can be set out as follows:

> 1. Prior to her release, Mary knows all the physical facts about the world and other people.
>
> 2. Prior to her release, Mary does *not* know all there is about the world and other people because after her release she *learns* new facts, namely phenomenological facts about conscious experience.
>
> C. Therefore, there are other facts about the world and other people (including Mary), specifically phenomenological facts, which cannot be known by all the physical facts alone, thus physicalism is false.

[10] Jackson, F. (1982), p. 130.

One way to see why physicalism is false in this case is that if physicalism *were* true, then phenomenal facts and qualia should have a physical basis of some kind from which they can be known. And if someone such as Mary knows all the physical facts then these facts should entail knowledge about phenomenal facts and qualia. But as Mary's situation reveals, "qualia are left out of the physical story", the implication being that there are certain facts, namely phenomenal facts, that cannot be *a priori* deduced from all the physical facts alone.[11] It is from Mary's lack of phenomenal knowledge that Jackson infers physicalism's falsity: phenomenal facts appear to be non-physical. The knowledge argument has had a significant impact on physicalism and physicalists have had a hard time refuting the argument.

One analysis of physicalism that I think is particularly important to consider in relation to the knowledge argument is Daniel Stoljar's (2001) paper *Two Conceptions of the Physical*. In this paper Stoljar considers how the contemporary debate around physicalism involves four central theses, these being:

(1) If physicalism is true, then a priori physicalism is true.

(2) A priori physicalism is false.

(3) If physicalism is false, epiphenomenalism is true.

(4) Epiphenomenalism is false.

Stoljar explains that these four theses are inconsistent: the first two theses taken together entail physicalism's falsity, and the second two taken together entail that physicalism is true. While all of the above (1-4) theses are plausibly true, there are many arguments that assert each individual thesis as false, and defending physicalism usually involves rejecting at least one of these four central theses.[12]

However, Stoljar proposes that one need not have to reject any of the four theses in order to defend physicalism. Another possibility, according to Stoljar, "is to

[11] Jackson, F. (1982), p. 130.

[12] Put briefly: A posteriori physicalists reject (1); A priori physicalists reject (2); interactionist dualists reject (3); and epiphenomenalists reject (4).

argue for some kind of ambiguity. If we can discern an ambiguity in (1-4) then it would seem possible to believe *all* the theses that make up the debate, rather than rejecting one of them."[13] I want to focus here on Stoljar's two conceptions of the physical and how these two conceptions are used for defending physicalism against the knowledge argument.

Stoljar's first conception of the physical is called *the theory-based conception*, on which "a physical property is a property which *either* is the sort of property that physical theory tells us about *or* else is a property that metaphysically (or logically) supervenes on the sort of property that physical theory tells us about".[14] Put simply, on this conception, properties such as having mass, charge, spin, etc., or the property of being a rock, a tree, a planet, etc., are physical properties when such properties are known and described by physical theory. Physical properties of this sort are therefore referred to as *t-physical properties*.

The second conception is called *the object-based conception*, on which "a physical property is a property which *either* is the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents *or* else is a property that metaphysically (or logically) supervenes on the sort of property that is required by a complete account of the intrinsic nature of paradigmatic physical objects." Put simply, on this conception, because rocks, trees, planets, etc., are considered paradigmatic physical objects, the property of being a rock, for instance, is a physical property. Importantly, if a physical property (*e.g.* having mass) is required for giving a *complete account of the intrinsic nature* of physical objects and their constituents, then such properties (*e.g.* mass, charge, spin) are physical properties. Physical properties of this sort are referred to as *o-physical properties*.

What distinguishes these two conceptions of the physical is that the first one concerns *theories* in the sense of what a particular theorem actually *tells us about* in respect to the world. The second conception, on the other hand, concerns

[13] Stoljar, D. (2001), p. 255.
[14] *Ibid*. p. 256.

objects in the sense of having a complete account of both the dispositional (extrinsic) and categorical (intrinsic) properties of physical objects and all their constituents. One might think that both these conceptions of the physical are co-extensive, such that t-physical properties are ultimately identical to o-physical properties and vice-versa, but Stoljar claims otherwise: "some o-physicals are not t-physicals".[15]

The reason why some o-physicals are not t-physicals is because "physical theory tells us only about the *dispositional* properties of physical objects and so does not tell us about the *categorical* properties, if any, that they have."[16] The distinction between dispositional and categorical properties is central to how Stoljar differentiates between t-physical properties and o-physical properties, and there are two important theses that motivate Stoljar's two conceptions of the physical.

The first thesis is that physical theory only *finds* or can tell us about dispositional and relational properties of objects in the world. The consequence of this is that physical theory is silent on the intrinsic nature of physical objects and is therefore unable to tell us about any categorical properties or the grounding relations that fully constitute physical objects. The second thesis is that "dispositional properties of physical objects *do* require categorical grounds, i.e. for all dispositional properties, there must be a non-dispositional property, or non-dispositional properties, such that instantiation of the latter is metaphysically sufficient of the former."[17] While it is fair to think that physical theory might at some point reveal the link between dispositional and categorical properties, Stoljar explains that physical theory cannot be epistemically certain of any references to categorical properties. This is because physical theory is limited to discovering dispositional and relational properties only and therefore has no epistemic access to any categorical properties that physical objects might possess. On the other hand, the object-based conception of the physical will however include an account of the intrinsic nature of physical objects, and so this

[15] Stoljar, D. (2001), p. 258.
[16] *Ibid.* p. 258.
[17] *Ibid.* p. 259.

conception will have epistemic access to the connection between dispositional and categorical properties. Stoljar's two conceptions of the physical are therefore epistemically distinct: whereas the *theory-based conception* says nothing about the intrinsic/categorical properties of physical objects, *the object-based conception* does say something about the intrinsic/categorical properties of physical objects.

These two conceptions of the physical lead to what Stoljar calls *t-physicalism* as defined by the theory-based conception, and *o-physicalism* as defined by the object-based conception. With these two forms of physicalism thus defined, Stoljar claims that while the two conceptions *do not* affect reasons for believing in either thesis (1) and (4), these two conceptions *do* affect reasons for believing thesis (2) the claim that a priori physicalism is false, and thesis (3) the claim that if physicalism is false, epiphenomenalism is true. I now want to focus on how Stoljar applies his two conceptions of the physical to thesis (2) because it is the most relevant conception of the physical when considering the knowledge argument. With Stoljar's two conceptions of the physical in mind, thesis (2) can be written in the following two ways:

(2-t) A prior t-physicalism is false.
(2-o) A priori o-physicalism is false.

Stoljar claims that (2-t) remains true in light of Frank Jackson's knowledge argument involving Mary. (2-t) is true because, as Jackson's argument illustrates, even with complete physical theory in mind, there are facts about the world which Mary remains ignorant of, namely phenomenological facts – facts about what it is like to have certain qualitative experiences. This is because physical theory – or t-physicalism – only tells Mary about dispositional and relational properties and nothing about categorical properties. Thus, Jackson's knowledge argument gives us good reason to believe (2-t). But does the knowledge argument give us good reason to believe (2-o), the thesis that a priori *o-physicalism* is false? According to Stoljar the answer is, No.

The reason why Stoljar thinks this is because o-physicalism contains facts about categorical properties which physical theory (t-physicalism) knows nothing about. Stoljar therefore argues that, "the object-based conception provides a way in which we might defeat the knowledge argument and at the same time concede the central intuition that motivates it."[18] In light of Stoljar's two conceptions of the physical, physicalists can now reject (2-t) on the grounds that t-physicalism lacks a certain class of properties, these being categorical properties, but accept (2-o) because o-physicalism contains knowledge about dispositional *and* categorical properties. *O-physicalism* therefore allows physicalists to argue that Mary does not and cannot know anything about phenomenal properties for the reason that she is working within the epistemic confines of physical theory, or t-physicalism. If, however, Mary were to instead possess o-physicalism her knowledge would be *broader* in such a way that it includes facts about the categorical grounds relevant to phenomenal consciousness.

While I think that Stoljar's conception of o-physicalism works successfully as a strategy for defending physicalism against the knowledge argument, it appears to me that o-physicalism runs very close to panpsychism. Stoljar is in fact aware of this, and he does consider the objection that o-physicalism runs the risk of collapsing into either neutral monism or panpsychism. This is because o-physicalism permits the categorical (or intrinsic) properties of paradigmatic physical objects as being definitive mental properties, specifically qualia. I want to examine here how Stoljar deals with the objection that o-physicalism appears to be more a form of panpsychism, and I will argue that his defence rests upon a very tentative metaphysical assumption.[19]

On the question of whether o-physicalism is a form of panpsychism Stoljar writes that in such a case "the categorical properties which underlie dispositional t-physical properties are in every case qualia".[20] In Stoljar's view, there are two motivations for panpsychism: Firstly, "that (dispositional) t-physical properties

[18] Stoljar, D. (2001), p. 264.
[19] Stoljar also defend o-physicalism against being a form of neutral monism, but it is his defence against panpsychism that is of more relevance to my own thesis.
[20] *Ibid*. p. 272.

exhaust the class of physical properties", the implication being that categorical properties are therefore non-physical mental properties; and secondly, due to the way in which qualia appear to be categorical properties of subjective conscious experience, "we seem to glean our concept of a categorical property from our concept of qualia".[21] Stoljar therefore claims that the first motivation for panpsychism can be rejected by anyone who accepts the object-based conception of the physical, which is to accept that categorical properties are definite physical properties. In my view, this response is fair enough. However, what I have an issue with is how Stoljar rejects the second motivation for panpsychism. Stoljar states that "even if one derives one's concept of a categorical property from one's concept of qualia… it does not follow that all categorical properties are qualitative properties." The implication here is that there can be a range of categorical properties that are also physical and non-qualitative. Stoljar goes on to claim that, "If o-physicalism is right, then some of these [physical and non-qualitative] properties will in combination be the supervenience base for qualitative properties."[22] That qualitative properties supervene on o-physical/non-qualitative categorical properties is, I believe, a very tentative metaphysical assumption. In my view this supervenience claim is the only way for Stoljar to save o-physicalism from collapsing into panpsychism. To remain true to supervenience physicalism o-physicalism needs it to be the case that *all* qualitative categorical properties supervene on physical and non-qualitative categorical properties. However, I do not find this supervenience claim convincing. I find it hard to see why qualitative properties would supervene on non-qualitative properties when the latter properties are categorically distinct from the former. I think the panpsychist can rightly argue that all categorical properties have (to use Strawson's 2006 term) *equal status* – that is, all categorical properties exist on equal ground at the fundamental level and are therefore inter-dependent of one another. I am more inclined here to believe the panpsychist view over Stoljar's supervenience o-physicalism. Even if qualia do supervene on other categorical properties, they are nonetheless fundamental properties simply by virtue of being categorical and intrinsic to

[21] Stoljar, D. (2001), p. 272.
[22] *Ibid*. p. 272-273.

paradigmatic physical objects – no matter how small or large the physical object, qualia are there no matter what.

My criticism here of o-physicalism centres on the supervenience claim that some combination of physical and non-qualitative categorical properties form the supervenience base for qualitative properties, which implies that o-physicalism is a form of supervenience physicalism. As much as I find o-physicalism an appealing conception of physicalism, I think it is fair to argue that o-physicalism is certainly at risk of collapsing into panpsychism, but only if qualitative categorical properties are to some extent inter-dependent of all other categorical properties. Even if qualia do supervene on other categorical properties as Stoljar claims they do, qualia still appear to be fundamental properties, thereby making o-physicalism run very close to panpsychism.

I now want to turn my attention to the next argument against physicalism, this being the conceivability argument. One of the most well-known conceivability arguments involves the idea of a philosophical zombie. Since supervenience physicalism claims that consciousness supervenes on the physical, the zombie argument presents a challenging counter-example against supervenience physicalism.

1.4 – The conceivability argument against physicalism

The idea of a philosophical zombie is quite straightforward: a zombie is someone who is physically identical to me – a minimal physical duplicate – except the zombie is completely lacking phenomenal consciousness. Not only is my zombie twin physically identical to me but he is also *functionally* and *psychologically* identical to me as well. As Chalmers explains, my zombie twin is functionally identical to me in that, "he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behaviour resulting".[23] He will also be psychologically identical to me in that his behaviour has the same

[23] Chalmers, D. (1996), p. 95

causal and explanatory basis as my own. It is important to note here that according to the *psychological* concept of mind as defined by Chalmers, "it matters little whether a mental state has a conscious quality or not. What matters is the role it plays in a cognitive economy." [24] Chalmers contrasts the psychological concept of mind with the *phenomenal* concept of mind, on which a mental state is one that is consciously experienced. With these two concepts of mind thus defined, zombies are physically, functionally and psychologically identical to me yet the zombie has no phenomenal *feel* and no qualitative *experience*, hence there is nothing it is like to be a zombie – all is dark on the inside.

While most philosophers do not believe that philosophical zombies actually exist, many concede that they are conceivable and therefore possible in some way or another. As Chalmers explains, when it comes to proposing the possibility of zombies, "it certainly seems that a coherent situation is described", coherent because the situation involves no obvious contradiction of terms.[25] It follows then that if zombies are conceivable, whereby a complete physical replica of oneself can lack phenomenal consciousness, then physicalism is false. The proposed falsity of physicalism therefore turns on the notion of logical possibility: that it is logically possible for all the physical properties to be instantiated while phenomenal properties remain absent or at least fail to be instantiated.

The conceivability argument is formalised by Chalmers as follows: let P be the conjunction of all physical truths (micro and macro) about the world, and let Q be any arbitrary phenomenal truth.

(1) $P\&\neg Q$ is conceivable.
(2) If $P\&\neg Q$ is conceivable, then $P\&\neg Q$ is metaphysically possible.
(3) If $P\&\neg Q$ is metaphysically possible, then physicalism is false.
(C) Physicalism is false.

[24] Chalmers, D. (1996), p. 11.
[25] *Ibid*. p. 96.

I think it is worth unpacking this argument to explain it in more detail. Premise (1) is conceivable because it cannot be ruled out *a priori*. The zombie argument illustrates this, for there is nothing contradictory about the claim that a complete physical replica of myself is, or at least could be lacking phenomenal consciousness. Premise (2) is supported by general reasoning about modal arguments and possible worlds: if a certain state of affairs is plausibly conceivable, then such a state of affairs is metaphysically possible. That conceivability entails possibility is highly contentious, but I have not the space here to consider the many arguments against premise (2). Premise (3) is supported by the reasoning that if $P\&\neg Q$ is metaphysically possible, then P (physical properties) do not metaphysically necessitate Q (phenomenal properties), meaning that phenomenal properties are not grounded in physical properties and therefore do not supervene on physical properties. The conclusion then is that supervenience physicalism is false.

There are many arguments against the conceivability of philosophical zombies. Premise (1) has been rejected on the grounds that zombies are *inconceivable*. I agree with Daniel Dennett who argues that the conceivability of zombie involves a "mistake of the imagination" and that "this mis-imagination undoubtedly provides ill-gotten support for the wide spread conviction among the populous… that there is an important different between zombies and conscious beings."[26]

Like Dennett, I also believe that zombies so conceived are incoherent entities. In my view, the *phenomenological* concept of mind cannot be so easily divorced from the *psychological* concept of mind such that these two concepts of mind can be held mutually exclusive. On the contrary, I strongly believe that phenomenality is an inextricable aspect of our psychology, so much so that the phenomenological concept and the psychological concept are mutually *inclusive* due to the causal role that qualia play within our psychology and overall cognitive economy. Because I view the phenomenological concept of mind as being inextricable from the psychological concept of mind, the notion of a

[26] Dennett, D. (1995), p, 325.

philosophical zombie is, I believe, an incoherent one. The incoherency, in my view, is based not in the logically possibility of zombies so conceived but rather in the ontological mystery surrounding the placement of qualia within the world; when qualia are given a rightful place in the world, I think philosophical zombies will disappear.

Premise (2) of the conceivability argument has been rejected on the grounds that since conceivability is an epistemic notion and possibility is a metaphysical (modal) notion, premise (2) falls apart because it presumes that conceivability is a reliable *guide* to possibility (See Yablo 1993). The criticism here is that our ability to conceive of things such as zombies is always intellectually informed by what we actually know about the world, knowledge that is of course incomplete and by no means infallible. That conceivability is an unreliable guide to possibility challenges the burden of proof that is said to lie with those who reject the logical possibility of zombies (cf. Chalmers 1996, p. 96). I think it is worth recalling how Descartes, by way of clear and distinct perceptions, conceived of his mind as being completely separate from his own physical body. Descartes' line of thinking reaches a conclusion not all that dissimilar to the philosophical zombie argument: that the mind (in Descartes view) and certain mental properties (phenomenal properties in the case of zombies) are metaphysically distinct from all things physical. In my view, Cartesian Dualism and philosophical zombies suffer from the same mistake, namely a mistake of the imagination, one that appears *prima facie* possible, but is actually metaphysically impossible mainly for reasons to do with causation between the mind and the body.

Although few philosophers nowadays believe in Cartesian Dualism, the many arguments against physicalism lead Chalmers to endorse what he calls *naturalist dualism*. Though dualist by name, Chalmers goes on to claim that his naturalist dualism is more a form of monism, qualifying further that "if a variety of monism is true, it cannot be a *materialist* monism. It must be something broader".[27] Chalmers provides little detail about how his naturalistic dualism is to be understood more as a form of monism, but if Chalmers' naturalist dualism takes

[27] Chalmers, D. (1996), p. 129.

into account intrinsic categorical properties in the same way that Stoljar's o-physicalism does (itself a broader form of physicalism) then Chalmers' naturalist dualism (monism) also appears to run very close to panpsychism as well. I therefore believe that any attempt to broaden physicalism by way of including qualia as intrinsic/categorical properties runs the risk of collapsing into panpsychism. All of this depends, however, on the ontologically organisation of categorical properties as fundamental properties. But if broadening physicalism has this effect, *i.e.* possibly collapsing into panpsychism, then I think that we have very good reason for taking panpsychism seriously.

I will now turn my attention to the third argument against physicalism, this being the explanatory gap argument.

1.5 – The explanatory gap argument

So far I have examined two arguments against physicalism: firstly, the knowledge argument, which presents an epistemic gap between physical facts and phenomenological facts, and secondly, the conceivability argument, which presents a metaphysical gap between physical and phenomenal properties. The third argument against physicalism that I will focus on here is what Joseph Levine (1983) has labelled as the *explanatory gap* – that mentality cannot be explained physically.

Levine's explanatory gap argument extends on Kripke's modal argument against physicalism by turning it into an epistemological argument. To see why physicalism faces an explanatory gap it is first necessary to have an understanding of Kripke's modal argument against physicalism.

In *Naming and Necessity* Kripke (1980) calls into question the claims made by psychophysical identity theory, on which certain types of mental states are claimed to be identical to certain types of physical states. Kripke challenges the following identity claim: that the mental state of *pain* is identical to the physical state of C-fibre stimulation; in short, that pain *is* C-fibre stimulation. Kripke

claims that identity theories of this kind (*i.e*. type-type identities) must be *necessarily* true, if true at all. This is to say that there cannot be a possible world in which the identity claim fails to be true. Kripke refers to the scientific identification of heat with molecular motion (*i.e.* heat *is* molecular motion) as an example of a necessarily true identity claim. The reason why this claim is considered necessary is because the two notions of 'heat' and 'molecular motion' are essential to one another – *i.e.* there cannot be one without the other. Central to Kripke's modal argument against identity theory is the notion of a *rigid designator*, the details of which I think are important to explain here, if only briefly.

The identity claim that 'heat is molecular motion' is to say that there are two types of phenomena in the world: one going by the name of 'heat' and the other being the motion of molecules. For these two phenomena to be identical it is required that both phenomena exist together and are essentially one another. The word heat is therefore said to rigidly designate the very phenomenon of heat, and likewise 'molecular motion' rigidly designates the phenomena known as molecular motion. Thus, let *A* be the rigid designator for heat, and *B* the rigid designator for molecular motion. Identity theory holds that *A* = *B*; that heat *is* the motion of molecules. According to Kripke, "If *A* = *B*, then the identity of *A* with *B* is necessary, and any essential property of one must be an essential property of the other."[28] If the said identity claim is necessarily true, the implication is that the phenomenon of heat cannot exist apart from the phenomenon of molecular motion: where there is heat there is by necessity molecular motion and *vice versa*.

Now what about the identity claim that pain *is* C-fibre stimulation? Kripke writes that the two notions of 'pain' and 'C-fibre stimulation' are both rigid designators for the reason that "if something is a pain it is essentially so, and it seems absurd to suppose that pain could have been some phenomenon other than the one it is. The same hold for the term 'C-fibre stimulation'… Thus the identity of pain with

[28] Kripke, S. (1980), p. 148.

the stimulation of C-fibres, if true, must be *necessary*."[29] That pain *is* C-fibre stimulation implies that the property of pain must be an essential property of C-fibre stimulation such that there cannot be pain without C-fibre stimulation and *vice versa*. But as Kripke argues, it certainly appears possible for there to be C-fibre stimulation without there being any associated pain, which is to say that the phenomenon of pain can exist apart from the phenomenon of C-fibre stimulation, thereby rendering pain a non-essential property of C-fibre stimulation. The identity claim that 'pain *is* C-fibre stimulation' is therefore said to be *contingent* and not necessary, the reason being that it is possible for the two phenomena to exist apart from one another. According to Kripke, psychophysical identity claims must be necessarily true, if true at all. But it appears that the identity between pain and C-fibre stimulation is not necessary but contingent instead.

Joseph Levine extends on Kripke's modal argument to present an epistemological argument against physicalism. Given Kripke's argument that the identity between pain and C-fibre stimulation is apparently contingent and not necessary, Levine explains that one plausible alternative to save physicalism is to adopt a functionalist view of pain. For functionalism, pain is not identical to a physical state such as C-fibre stimulation but is instead identical to a certain functional state. Thus, the following identity claim can be made for functionalism: To be in pain is to be in state F; where F stands for whichever functional state is responsible for realising pain. If this functional identity claim is true, then it follows that pain must be an essential aspect of state F such that pain is always instantiated whenever state F is instantiated. However, Levine argues that it certainly seems possible for state F to be instantiated without there being any pain associated with it. Pain therefore does not appear to be an essential aspect of the functional state F.

To see why physicalism faces an explanatory gap let us reconsider the three identity claims mentioned so far, with these being:

(1) Pain is C-fibre stimulation

(2) Heat is molecular motion

[29] Kripke, S. (1980), p. 148-149.

(3) Pain is to be in functional state F

In respect to these three identity claims Levine writes that statement (2) "expresses an identity that is *fully explanatory*, with nothing crucial left out. On the other hand, statements (1) and (3) do seem to leave something crucial unexplained, there is a "gap" in the explanatory import of these statements."[30] The reason why Levine claims that statement 2 is *fully explanatory* has to do with the fact that our knowledge of chemistry and physics "makes it intelligible" how the motion of molecules is essential to and causally responsible for the phenomenon of heat. In other words, the phenomenon of heat can be understood – fully explained – by what we know about the motion of molecules.

The same might be said of course about statements (1) and (3), that both statements explain a lot about the phenomenon of pain. I think an explanation might go something like this: that pain is a state that plays a certain causal role for practical biological purposes (e.g. pain causes us to avoid potential harms and hazards, pain informs us of dangerous situations, pain-avoidance mechanisms assist in our survival etc.). And statement (3) explains the mechanisms that underlie the causal processing of these functions. Indeed, such an explanation is broadly correct, at least in principle. But is our concept of pain *fully explained* but our functionalist understanding of it? According to Levine, No it is not:

> There is more to our concept of pain than just its causal role, there is it qualitative character, how it feels; and what is left unexplained by the discovery of C-fibres firing is *why pain should feel the way it does!* For there seems to be nothing about C-fibres firing which makes it naturally "fit" the phenomenal properties of pain, any more than it would fit some other set of phenomenal properties.[31]

I think Levine successfully exposes a major problem for physicalism, which is that our physical and functional concepts do not *fully explain* why it is that certain physical states such as C-fibre stimulation should *feel* the way they do. It

[30] Levine, J. (1983), p. 357.
[31] *Ibid.* p. 357.

therefore seems that there is something more to the nature of phenomenal states than just their physical counter-parts and functional mechanisms. In summary, Kripke argues that for type-type identity claims to be true, they need to be necessarily true, if true at all. However, the identity claims of (1) and (3) appear contingent. That is, both claims could be false, at least in other possible worlds. According to Levine, in order to save physicalism, physicalists need to dissolve the felt contingency for claims (1) and (3) thereby presenting them as necessary. But since our physical and functional concepts cannot explain why it is that certain states *feel* the way they do, physicalism faces an explanatory gap between physical and phenomenal states.

I think that the explanatory gap argument as presented by Levine is a major challenge for physicalism. I want to consider just briefly how the physicalist might *further (but not fully) explain* why it is that the firing of C-fibres should *feel* the way it does; that is, why the firing of C-fibres *is* painful. To explain why C-fibre stimulation *is* painful I think physicalism needs to take more seriously the *intrinsic* and/or *categorical* nature of physical states. By holding the phenomenal property of *pain* as an intrinsic property of physical states, the reason why C-fibre stimulation feels painful is because the intrinsic nature of such a state just so happens to be that of painfulness.[32] This might appear slightly trivial but I think it is at lease a step in the right direction. Moreover, besides appealing to some kind of dualism, I cannot see any alternative strategy available to physicalism other than to formally include intrinsic/categorical properties into the doctrine. I therefore believe that to make some advancement regarding the explanatory gap physicalism needs to be broadened to include intrinsic/categorical properties. But as I have demonstrated so far, to formally

[32] Levine raises another issue by asking why pain and not some other phenomenal state is associated with C-fibres? This remains an open question, but I suspect that the structural properties of physical states might play a role in correlating specific phenomenal states with specific physical states, just as we see in the correlation between C-fibre stimulation and pain. This is simply my own philosophical speculation, however.

include such properties appears to modify physicalism more into a form of panpsychism.

1. 6 – Conclusion

Since many philosophers regard physicalism as the most promising philosophical theory of mind, my aim in this chapter has been to examine how physicalism has been traditionally understood in the forms of reductive and non-reductive physicalism, and in the more contemporary form of supervenience physicalism. After outlining these differing forms of physicalism, I turned my attention to what I think are the most challenging philosophical arguments against physicalism, these being the knowledge argument, the conceivability argument and the explanatory gap argument.

I believe that these three arguments against physicalism when taken together certainly call into question the viability of the doctrine and they definitely leave physicalism in a good deal of doubt. In this chapter I have considered ways in which physicalism might resolve these arguments. One option is to adopt Stoljar's broader conception of physicalism, namely o-physicalism, which sees qualia as categorical properties supervening on physical/non-qualitative properties. I argued against Stoljar's view by claiming that it is possible that all categorical properties have *equal status* and are therefore interdependent of one another. This means that some fundamental properties of the world are definite mental properties, specifically qualia. If my view is correct then the consequence is that o-physicalism appears to collapse into panpsychism. All of this depends, however, on the ontological organisation of intrinsic/categorical properties at the fundamental level.

Due to the conceivability of philosophical zombies Chalmers proposes what he calls naturalistic dualism, which in his view is more a form of monism, but one that is broader than materialistic monism. I argued that if naturalistic dualism is to include intrinsic/categorical properties then it appears very much like Stoljar's o-physicalism, itself a broader form of physicalism. And if this is the case

then Chalmers' naturalistic dualism (monism) also runs very close to panpsychism.

And in light of Levine's explanatory gap argument, I claimed that one way for physicalism to explain the correlation between physical states and phenomenal states is to take seriously the notion of intrinsic/categorical properties. Indeed, I think this is the only viable option available to physicalism as a monistic doctrine; if there is nothing *over and above* the physical, then there must be something within and therefore intrinsic to the physical. This therefore requires broadening physicalism to formally include intrinsic/categorical properties.

Given these three arguments against physicalism I therefore believe that panpsychism is a promising alternative to physicalism, and could possibly aid or advance physicalism so as to become a broader philosophical doctrine. Although the two doctrines of physicalism and panpsychism may seem philosophically opposed to one another, I think that these two doctrines are to some extent philosophically compatible. And if panpsychism is able to assist in broadening physicalism for the purpose of overcoming the many arguments against it, then it worth investigating how panpsychism might do this. There are many versions of panpsychism but there are three contemporary versions that I think are worth investigating. These three versions of panpsychism will become my focus in the next chapter.

CHAPTER TWO

2.0 – Introduction

Panpsychism is traditionally defined as the thesis that *everything has a mind* or that *everything is conscious*. Generally speaking, panpsychism views *psyche* or mentality as being a universal phenomenon, which is to say that the fundamental nature of the universe and everything in it has or involves both physical and mental properties, either in some dualistic format or by way of some underlying monistic substratum.

Panpsychism is not only a highly speculative philosophical view of the world, it is also highly controversial, mainly because of the way in which it appears that philosophers are simply abstracting certain properties of ourselves – mental properties in this case – only to implant them into everything from atoms to plants, stars and galaxies, right up to the universe itself, such that *everything has a mind*. Panpsychism is more sophisticated than this however. There are many versions of panpsychism and each one differs in terms how they view the universal nature of mind and consciousness. It is interesting to note that panpsychism has been a widely held philosophical world-view throughout history dating back to Plato and Aristotle, both of whom held world-views that can be considered quasi-panpsychist.[33]

Panpsychism can range between weak and strong versions. Weak versions of panpsychism view only a small number of primitive mental properties in the world. For instance, Aristotle viewed the cosmos has having two essential properties: firstly, *ether*, which was a self-moving entity that powered the rotation of the heavens, and secondly *pneuma*, which is taken to be the principle of 'psychic action' that functioned as the "faculty of all kinds of souls" and is thus the "principle soul".[34] Strong versions of panpsychism hold that all fundamental physical properties are mental properties or mental states in some form or

[33] For a fascinating historical account of panpsychism, see Skrbina (2005).
[34] Skrbina, D. (2009). p. 7.

another. The strongest version of panpsychism is *pure* panpsychism, which views everything from atoms to single-celled organisms, right up to tables, chairs, thermostats, trees, plants and the universe itself as all having minds of their own. It is fair to say that pure panpsychism no longer carries any credibility simply because no one believes that tables and chairs or any other inanimate object for that matter have minds of their own. From the point of view of modern science only *living* organisms and biological entities are seen as capable of being conscious and as having some degree of mentality whether it be primitive or complex.

In this chapter I will focus on three contemporary versions of panpsychism. The first version is *pan-experientialism*, which holds conscious *experience* as being the intrinsic nature of material reality. On pan-experientialism, conscious experience is intrinsic to all material reality and is therefore seen as the essential nature of the universe – *everything* is or has an experience of some kind. I will focus specifically on Galen Strawson's version of pan-experientialism, which he also calls *realistic physicalism*.[35]

The second version of panpsychism that I will focus on is *pan-phenomenalism* as developed by Sam Coleman. This version views phenomenal properties as being external, mind-independent properties. Pan-experientialism and pan-phenomenalism differ significantly in respect to how experiential properties and phenomenal properties are understood in relation to subjectivity. Put briefly, pan-experientialism holds that conscious experience necessarily requires a 'subjects-of-experience', *i.e.* there cannot be an experience without a subject-of-expereince, whereas pan-phenomenalism claims that phenomenal properties do not contain subjects and are therefore non-subjective properties. This ontological distinction regarding subjectivity is crucial for assessing how these two versions of panpsychism deal with particular combination problems, namely the *micro-subject/macro-subject gap* and *the non-subject/subject gap* respectively.

[35] See Strawson, G. (2006).

The third version of panpsychism that I will focus on is *pan-protopsychism* as defined by David Chalmers. Pan-protopsychism is the view that the most fundamental physical entities have proto-phenomenal properties. These proto-mental properties are therefore metaphysically responsible for grounding all conscious phenomenal experience. Pan-protopsychism therefore confronts a more unique combination problem: that of explaining how protophenomenal properties *a priori* entail phenomenal properties. This is referred to as the protophenomenal/phenomenal gap.

My aim in this chapter is to examine how all three of the aforementioned versions of panpsychism deal with their respective combination problems. By focusing on these three versions of panpsychism I will show that although each version has its own respective merits, neither pan-experientialism nor pan-phenomenalism successfully bridge their combination problems, while pan-protopsychism has no current solution because it has yet to be developed in to an actual theory. I will argue that in their current forms these three versions of panpsychism remain either philosophically problematic or else require further philosophical development if they are to successfully resolve their combination problems.

2.1 – Pan-experientialism and *the micro-subject/macro-subject gap*

Pan-experientialism is the view that material reality is extrinsically physical and intrinsically *experiential*. On this view the intrinsic nature of matter is not simply *void* or inert but rather, matter is seen as being energised in such a way that it is experientially *active* and therefore characteristically experiential. The most comprehensive account of pan-experientialism has been developed by Galen Strawson, who argues that material reality consists of both "experiential and non-experiential phenomena."[36] According to Strawson, these two phenomena -

[36] Strawson refers to his version of panpsychism as *Realistic Monism* in which he argues that *physicalism entails panpsychism* when the 'experiential phenomena' of consciousness are given equal status in terms of their reality in the same way as the 'non-experiential phenomena' of strictly physical properties are held as being real properties of material reality. See Strawson (2006).

the *experiential* and the *non-experiential* – are fundamentally equal to each other, such that "experiential [mental] phenomena do not depend on non-experiential physical phenomena".[37] Strawson accepts the materialist premise that *all reality is physical*, but then goes on to argue that physical reality "is, in some fundamental sense, [made up of] only one kind of [monistic] stuff in the universe".[38] What Strawson proposes therefore is that physical reality is fundamentally singular in terms of its substantial nature, which is *energy*, and that the duality of this singular universal energy is to be understood as being both 'experiential/mental' and 'non-experiential/physical'.

For Strawson the motivation for pan-experientialism comes from one of the most obvious facts about our lives, which is that human beings are undoubtedly conscious beings. Moreover, Strawson claims that conscious experience is "the most certainly known thing there is" – to deny conscious experience is to deny one of the most obvious facts about ourselves and the essential nature of our lives.[39] I think Strawson is absolutely correct on this point. The epistemic certainty of conscious experience is hard to refute. Not even deep scepticism can succeed in denying the existence of consciousness for the simple reason that it is only in virtue of being conscious that we can ever come to doubt reality in the first place. Physical reality may well be illusory, but whatever the illusion of reality might be we nevertheless experience it most explicitly in terms of its phenomenality, that is, the way in which we feel and sense it through our every-day conscious experiences.

By acknowledging the certainty of conscious experience, and that human beings are indeed conscious organisms, Strawson goes on to claim that the experiential nature of consciousness is itself an *a priori* feature of physical reality, which is to say that conscious experience is a fundamental characteristic of the physical world. Pan-experientialism therefore rejects traditional physicalism and the view that phenomenal/mental facts supervene upon physical facts. According to

[37] Strawson, G. (1994), p.73-74.
[38] *Ibid.* p. 187
[39] *Ibid.* p. 226.

pan-experientialism, phenomenal/mental facts exist alongside physical facts such that the physical and the mental exist in unison.

Although I find pan-experientialism to be an intuitive world-view, I think the theory becomes philosophically problematic in one respect, this being the ontological economy of subjects-of-experience in the world. In my view, pan-experientialism entails a grossly over-populated world of subjects-of-expereince. The reason for this has to do with how Strawson defines conscious experience as being necessarily subjective. Strawson writes that:

> There cannot be experience without a subject of experience simply because experience is necessarily experience *for* – for someone-or-something. Experience necessarily involves experiential 'what-it-is-likeness', and experiential what-it-is-likeness is necessarily what-it-is-likeness *for* someone-or-something. Whatever the correct account of the substantial *nature* of this experiencing something, its *existence* cannot be denied.[40]

Because conscious experience necessarily requires a subject-of-expereince, Strawson defines a subject as a SESMET: '**S**ubjects of **E**xperience that are **S**ingle **M**ental **T**hings'.[41] For Strawson, SESMETS are metaphysically necessary entities for grounding conscious experience. This is because, as Strawson states, there cannot be conscious experience without there being a *subject* (or SESMET) for which the conscious experience belongs. I definitely agree with Strawson on this point: I think it is true to say that all conscious experience is *subjective* and that it is impossible for there to be an *unexperienced* experience. I believe that all conscious experience, whether human or otherwise, is necessarily experienced in a subjective manner, that is, by someone or something. While I certainly believe this to be true, my issue with Strawson's view relates to how pan-experientialism entails a grossly over-populated ontological economy of subjects/SESMETS in the world. I will elaborate on this further on, but for now I want further explicate Strawson's world-view.

---

[40] Strawson, G. (2006), p. 6.
[41] Strawson, G. (1999), p. 118.

Strawson endorses a version of panpsychism that he calls *Equal-Status Fundamental Duality Monism* or ESFD Monism for short:

> *Equal-Status Fundamental-Duality monism*
> Reality is substantially single. All reality is experiential and all reality is non-experiential. Experiential and non-experiential being exists in such a way that neither can be said to be based in or realized by or in any way asymmetrically dependent on the other (etc.)[42]

What ESFD monism holds is that the classical duality of 'mind and body' is reducible to a more fundamental duality in the metaphysical form of 'experiential and non-experiential' phenomena. Both the experiential and the non-experiential have *equal status* in terms of their ontological reality: the mind reduces only to the experiential/mental and the body reduces only to the non-experiential/physical. Both the experiential and the non-experiential are fundamentally unified in terms of being the two most primitive properties of the singular universal substance of *energy*. This is because in Strawson's view: "Energy is experientiality; that is its intrinsic nature."[43] If this is true, and given Strawson's claim that conscious experience is necessarily an experience *for – for-someone-or-something*, then this logically entails that all energy and energetic physical states include subjective experiences. That is, all energetic states in their physical form (*i.e.* particles, atoms, molecules, objects, etc.) require micro-subjects in order to ground the experience that is intrinsic to all the physical states of the world from the fundamental level upwards.

The metaphysical consequence of ESFD monism as defined by Strawson is that wherever there is experience there must also be a subject that is having the

[42] Strawson, G. (2006) p. 241.
[43] Strawson further states that: All that exists is substance [or] substances. All subjects (insofar as they are properly conceived as plural) are substances. All substances… are subjects. Equally, all substance is experientiality. Equally, all substance is energy, for substance is essentially active. The fundamental definition of 'substance' is not Aristotle's or Descartes', but Leibniz's: to be a substance is *to act*." *Ibid*.

experience, and if experience is a fundamental, universal phenomenon, then by logical necessity there must also be fundamental micro-subjects to which all fundamental experience belongs. From the viewpoint of physics the foundation of reality is understood in terms of elementary particles such as preons, leptons, bosons, and other field quanta at the quantum level. On pan-experientialism, one implication is that all fundamental field quanta are in and of themselves micro-subjects by virtue of being microstates of energy. Hence there are subjects at every level of reality, from the quantum level right up the classical, biological level.

I want to point out here that it is impossible to say what such micro experience might be like, however we can speculate. Presumably micro-subjects have a specific experience in the form of their own unique 'what-it-is-likeness', which means that they contain certain phenomenal qualities or perhaps a single phenomenal quality such as 'redness', in which case a micro-subject is simply having a *pure red* experience.[44] These micro-subjects therefore stand as the fundamental placeholders for all other (macro) conscious experience; nothing more is fundamentally required for generating conscious experience other than these micro-subjects at the fundamental level. The broader view then is that these micro-subjects are what underlie and thus metaphysically necessitate the macro-subjective experience as had (at least) by human beings.

The philosophical problem that confronts pan-experientialism here is that of explaining how these micro-subjects *combine* so as to yield a more complex macro subject (a macro subject being one that has a phenomenally rich and complex subjective experience). This combination problem is known as the *micro-subject/macro-subject gap* or the *subject-summing problem*.

A preliminary concern for the subject-summing problem, which I think is important to note, is that there is little in the way of a philosophical consensus

[44] It is impossible for us to know exactly what a micro-subject is experiencing, and any theoretical postulation of micro-subjectivity is done simply for the purpose of critiquing panpsychism.

regarding what the ontological status of a subject *is* exactly. David Chalmers goes so far as to say that "it is not obvious that we are Subjects. There is no introspective datum that we are Subjects, and it is not obvious that there are strong theoretical arguments to that effect."[45] Nevertheless, everyone is familiar with their own subjective experience of the world, and there are some obvious phenomenological facts about the nature of subjective experience which I think need to be put on the table for the purpose of providing a basic definition of what a subjective experience involves.

I think that subjective experience can be generally understood in the following way: Subjective experience is *unified* – all sensory experiences converge into a single unified perspective and experience of the world around us (See Bayne 2010); Subjective experience is *private/internal* – it is an exclusively private and internal phenomenon knowable only to the subject/person; and subjective experience is phenomenologically complex – it contains multiple phenomenal properties that pertain to the *qualitative* features of conscious experience such as specific colours, sounds, tastes, textures, smells, etc. A subject can therefore be understood as a fully unified, private/internal, phenomenal conscious state/entity/being.

Returning to the *subject-summing problem*, pan-experientialism holds that conscious experience is the essential nature of material reality. This entails that there must be fundamental micro-subjects that ground conscious experience at the fundamental level, and that at least some of these subjects can be said to constitute the macro-subjective experience as had (at least) by human beings. To clarify this further, a *micro*-subject in its most primitive form can be said as having simply *one* experience of say the colour 'red' or perhaps a particular 'audio' experience or any other specific phenomenal property that can be consciously apprehended. Thus, a micro-subject is quite literally a micro mind whose consciousness is the complete sum of its experiences. On the other hand, a *macro*-subject is one that can experience (in a unified manner) two or more phenomenal properties simultaneously, much in the same way that we can both

[45] Chalmers, D. (2013) p.24.

see and *hear* simultaneously. The standard five senses thus provide human beings with a highly complex, phenomenally rich experience of the world.

Pan-experientialism therefore holds that these fundamental micro-subjects are able to combine so as to yield macro-subjects. This is otherwise referred to as the micro-subject/macro-subject gap. As I will now show, a number of arguments have recently been put forward contending that the subject-summing problem is philosophically problematic and difficult to resolve.

The first argument that I want to consider here regarding the subject-summing problem is Philip Goff's epistemological argument. This argument focuses on the *a priori* route taken by panpsychists like Strawson who claims that experience is an *a priori* feature of the world. Goff states the panpsychist *a priori* view as follows:

A priori route: The physical facts plus micro-experience (of a certain kind) *a priori* entail o-consciousness.[46]

By 'o-consciousness' Goff means, "the kind of conscious experience that mirrors the overall behavioural function of the organism".[47] That is, 'o-consciousness' is the completely unified macro phenomenal state-of-mind as had by an organism such as a human being. To demonstrate why micro-experience fails to yield o-consciousness, Goff introduces what he calls a *micro-experiential zombie*. A micro-experiential zombie is a complete physical duplicate of an organism that also contains micro-experiential subjects, e.g. Strawsonian SESMETS. To see the mental life of this zombie in more detail, imagine that it has been stabbed in the leg with a knife. In this case the zombie feels pain but it does so only at the micro-level – *only* micro-subjects are in pain – all the while the zombie does not feel pain at the macro level of o-consciousess. According to Goff, the zombie still functions in such a way that the act of *being stabbed* causes it to run around

[46] Goff, P. (2009), p.294.
[47] *Ibid.* p. 290.

screaming just like anybody else, yet the zombie is actually pain-free at the level of o-consciousness – the zombie is only in pain at the micro-level.

The reason why the micro-experiential zombie does not feel pain at the level of o-consciousness is because micro-pain is experienced only by a micro-subject. Goff therefore argues that the existence of a micro-subject in pain does not *a priori* entail any o-conscious experience of pain. I want to elaborate on Goff's argument for a moment. Let us assume that the micro-experiential zombie, after having been stabbed in the leg, is now running around a smoky bonfire with loud music in the background. Within the zombie there are now three micro-subjects, each of which is individually instantiating one of the respective experiences of feeling pain, smelling smoke, and hearing music. Recall that o-consciousness is that of having a *unified phenomenally complex experience* of the world around us. If the zombie were a "normal" organism (e.g. a human being) it would be having a fully unified, phenomenally complex subjective experience of *being in pain while smelling smoke and hearing music* – but this is exactly what the micro-experiential zombie is lacking.

What Goff's *a priori* argument highlights is that the existence of micro-experience along with micro-subjects does not bring us any closer to understanding how macro 'o-conscious' experience is generated: there is nothing about micro-experience and the existence of micro-subjects that *a priori* entails the existence of o-consciousness. This leads Goff to introduce the following principle that he calls the: *No Summing of Subjects* principle or *NSS* for short.

> NSS: It is never the case that the existence of a number (one or more) of subjects of experience with certain phenomenal character *a priori* entails the existence of some other subject of experience.[48]

I believe that NSS is true and that it stands as a solid metaphysical principle. Goff does however consider one possible panpsychist manoeuvre, which involves pushing o-consciousness down to the micro-level such that "the only way

[48] Goff, P. (2009), p. 302

microexperiential facts could entail the existence of o-experience is if one (or more) of the physical ultimates constituting an organism itself instantiated the o-experience of the organism."[49] The problem here is that this manoeuvre appears overly demanding, mainly because of the way in which the functional states of the whole body now have some causal connection to the intrinsic nature of simply one or more physical ultimates. That one or more physical ultimates can ground the complex functional states of o-consciousness is a large task, and in my view, is *prima facie* implausible simply because the functional states of the whole organism obviously involve not only the physical ultimtes but more complex physical mechanisms as well, *e.g.* neural networks.

I think that Goff's epistemological argument successfully demonstrates that positing micro-expereience at the fundamental level fails to explain *a priori* the manifestation of macro-subjective/o-consciousness; the conclusion is that no combination of micro-subjects *a priori* entials the existence of a more complex, macro subject with o-consciousenss. Hence, combining micro-subjects does not bring us any closer to understanding macro-subjective experience. Sam Coleman also argues that the *subject-summing problem* is not only epistemologically problematic as Goff demonstrates, but that the very notion of subject-combination is metaphysically impossible. I think Coleman's argument reveals an important metaphysical problem for pan-experientialism, which I will now explain.

Coleman illustrates his argument by using a water molecule as a template for detailing the combinatorial features of natural physical objects. A water molecule is the assembly of two hydrogen atoms and one oxygen atom. When hydrogen and oxygen atoms combine into water "they bond *covalently*, sharing electrons. The oxygen atom completes its outer shell by borrowing an electron from each hydrogen. Thus the three atoms are *deformed*, intrinsically modified, by participating in the combination of water. Yet, importantly, all three atoms continue to exist once combination is achieved."[50] This last sentence is important

[49] Goff, P. (2009), p. 303
[50] Coleman, S. (2013), p. 30.

to note: that *all three atoms continue to exist once combination is achieved*. On pan-experientialism, all physical ultimates have experiential properties, and as we saw earlier, experiential properties necessarily belong to and require a subject-of-experience. By taking each atom of a water molecule as a micro-physical entity, we can suppose that pan-experientialism holds the hydrogen atoms as having some sort of 'hydrogen-experience' and the oxygen atom as having some kind of 'oxygen-experience'. Furthermore, when the two hydrogen atoms combine with an oxygen atom their combination yields a water molecule with, supposedly, 'water-experience'. Thus, a single water molecule, according to pan-experientialism, has the atomic structure of three fully physical atoms *plus* the three micro-subjects each having the respective hydrogen/oxygen experiences. This logically entails that the water molecule as a *whole entity* must contain a molecular subject given how the experience of a water molecule is presumed to be phenomenally distinct from the more primitive 'hydrogen/oxygen' experiences that belong to each atom respectively. But as Coleman demonstrates, the combination of micro-subjects cannot metaphysically constitute a higher-order subject such as a molecular subject.

The process of physical combination is one whereby atoms become molecules, molecules become cells and cells become organs, all of which ultimately constitute the organic bodies of human beings with macro-subjective experience. Pan-experientialism holds that the macro-subjective experience as had by human beings is generated by all the underlying micro-subjects down to, in this case, the atomic level. But as Coleman points out, given that all the atoms composing the body *continue to exist once combination is achieved*, then so too must all the micro-subjects that belong to all of the combined physical properties that form the human body, *i.e.* all the micro-subjects within atoms, molecules, cells, organs, etc. For pan-experientialism, these micro-subjects are said to metaphysically ground all other forms of conscious experience and they are also required for generating macro-subjective experience, or o-consciousness as Goff

describes it. But as Coleman argues, " this cannot work [for panpsychism], simply because points of view cannot combine."[51]

Coleman's argument builds on the fact that since physical ultimates survive and continue to exist after combining into, say, a molecule, then so too should all the micro-subjects continue to survive and exist after combination. Since pan-experientialism holds that just as all the physical constituents of the world survive when combined, then it must be the case that all the micro-subjects survive as well. Thus, subjects do not cease to exist after combination – they live on! Coleman therefore makes the point that, "If one subject is left where formerly we had two, this means at least one subject has gone out of existence, which is *not combination but a fight to the death.*"[52]

Continuing with Coleman's template of a water molecule, the subjective framework of a water molecule can be seen as one in which there are at least three atomic subjects (two hydrogen and one oxygen subject respectively) and one molecular subject. In this case there is a total quantity of *four* subjects inhabiting a water molecule, each one with their own unique subjective *point-of-view*. Coleman's central metaphysical argument is this: "There's really no such thing as a 'quantity' of point-of-view-edness, or perspective. Points of view are binary entities: they exist either wholly or not at all."[53] I completely agree with Coleman on this point. Since subjects are taken to be individual entities, to claim that a combination of micro-subjects can yield a more macro-subjective entity is metaphysically implausible simple because there is no such thing as a "quantity" of perspectives. I firmly believe that subjects are ontologically singular entities and, in light of the arguments given by Goff and Coleman, that the subject-summing problem cannot be resolved. The consequence of this is that Strawson's pan-experientialism becomes philosophically unviable because it needs subject-combination in order to explain macro-subjective/o-conscious experience. Coleman has in fact proposed another version of panpsychism that he believes is

[51] Coleman, S. (2013), p. 32.
[52] *Ibid.* (My italics)
[53] *Ibid.* p. 32

a more viable alternative to pan-experientialism for the reason that it avoids the micro-subject/macro-subject gap. Coleman calls his version of panpsychism *pan-phenomenalism*. I will now turn my attention to Coleman's pan-phenomenalism as an alternative version of panpsychism and in doing so I will look at how it must deal with another combination problem – the non-subject/subject gap.

2.2 – Pan-phenomenalism and the non-subject/subject gap

Pan-phenomenalism is a weaker version of panpsychism for the reason that it rejects the existence of micro-subjects (*e.g.* Strawsonian SESMETS) at the fundamental level. The consequence of this is that by denying the existence of fundamental micro-subjects pan-phenomenalism faces the problem of having to explain how subjective experience arises from non-subjective phenomenal properties. This problem is otherwise referred to as the *non-subject/subject gap*.

I want to begin by looking at how Coleman rejects two traditional panpsychist assumptions. These two assumptions are:

1. Phenomenal ultimates are themselves subjects of experience.
2. Phenomenal assembly can only be aggregative.

Coleman states that, "the First Assumption is what makes the Second Assumption plausible."[54] The first assumption holds that the physical ultimates of material reality are in and of themselves subjects of experience, or SESMETS as Strawson calls them. Having demonstrated in the previous section that micro-subjects cannot combine, viz., the NSS principle, Coleman's pan-phenomenalism holds that the physical ultimates of the material world are *not* fully-fledged subjects but rather, the ultimates as simply phenomenal properties. This means that the fundamental physical ultimates are themselves phenomenal properties. The major claim here is that, unlike Strawson's view, phenomenal properties can exist *unexperienced*. The key difference here in comparison to pan-

[54] Coleman, S. (2012) p. 144.

experientialism is that phenomenal properties do not need to be metaphysically grounded in a subject-of-experience.

To reject the first panpsychist assumption as stated above Coleman begins by differentiating phenomenal ultimates from the more common conception of a sense-datum. Coleman explains that a sense-datum, *e.g.* an externally located sensory item such as the redness of a red London bus, is typically understood *as a sensory item whose existence is restricted to the episode of being experienced by the subject in question*. Coleman claims that there are two important disanalogies between sense data and instances of phenomenal qualities. The first is that phenomenal qualities are not *objects* in the way that sense-data are considered to be. Redness, for instance, can be thought of as an item of sensory information that a neurological system interprets and presents as some shade of the colour red. Coleman argues that phenomenal properties should not be considered in such a manner, and I will explain the reason why further on. The second disanalogy is that phenomenal ultimates are not to be construed as *perceptual objects* as on the 'act-object model' of perceptual experience. What the act-object model holds is that a perceived item such as a particular colour does not exist beyond it being observed by a subject. The act-object model holds that the act of observation fully captures and positively exhausts the ontological being of the item observed. But as Coleman explains, there is something wrong with this current hypothesis that phenomenal qualities cannot exist beyond the act of observation.

The problem as Coleman sees it has to do with phenomenal ultimates being assimilated with the kind of objects as described by the sense-data theory of perception. According to the sense-data theory, "an item will either *derive its existence from* figuring in such awareness, or else its existence *will consist in*, will take the form of, figuring in such awareness."[55] Coleman argues that neither of these possibilities is coherent. The first option, that an item derives its existence from figuring in awareness, is incoherent because for the item in question to figure in the awareness of a given subject of experience logically requires the

[55] Coleman, S. (2012), p.151.

item to exist *prior to* the awareness of the subject of experience. This means that the item itself must exist outside the subject's field of awareness. The act of figuring in awareness by the item then occurs upon entering the subject's field of awareness. Hence, the phenomenal item must exist *prior to* the act of perceptual awareness and therefore outside the subject's overall field of awareness.

To explain the incoherence of the second option, that an item's existence will consist in or take the form of figuring in such awareness, is a harder task. What this option says is that the precise mode of being of the item in question is generated by its figuring in an episode of awareness. This results in phenomenal properties being *essentially presentational* items. The problem here is that this conception of phenomenal qualities violates our current conception of what awareness involves. "Awareness is a relation between two items: *a* is aware of *b*. But this presupposes that *b* has a logically possible existence outside of *a*'s being aware of it (and so outside of awareness in general)."[56] According to Coleman, the incoherence of this option can be seen in how consciousness is constituted. If the phenomenal item was something that existed simply within the field of awareness and not outside of it, then the item itself must be a constituting and modifying element of the actual act of awareness instead of being an object that enters into the awareness of the subject. This means that 'phenomenal red' would be characterised as a *manner* of being aware. Rather than being aware of 'redness', the subject would be in a state of *being-aware-redly.* The colour red thus becomes nothing more than a modification of the subject's awareness rather than something that enters into the conscious field of the subject. The aim of adverbialism (*redly*), as Coleman puts it, is to do away with the existence of phenomenal properties outside the domain of the awareness. However, Coleman wants to deny this hypothesis, and he does so by arguing that instead of constituting consciousness *per se*, phenomenal properties can be seen as constituting the intrinsic nature of the fundamental ultimates that constitute the actual physical world.

[56] Coleman, S. (2012), p. 151.

For phenomenal properties to be fundamental, Coleman claims that pan-phenomenalism must adhere to some form of the 'Independence Argument'. This argument holds that phenomenal ultimates do not obtain their existence simply by virtue of being experienced by a subject. This requires that phenomenal qualities must have a logically possible life outside the field of subjective conscious awareness. Coleman therefore states that there cannot be a property, a real feature of reality, that gains its existence simply via awareness. I certainly agree with Coleman on this point: For if it were true that the existence of a property is generated simply by awareness, then this would imply that consciousness itself has ontological control over bringing certain entities into reality, meaning that consciousness in effect creates certain properties, namely phenomenal properties. I am doubtful that consciousness operates like this. I therefore support Coleman's argument that the existence of a phenomenal property cannot derive from nor consist simply in someone's experience of it. Phenomenal properties, according to Coleman, *must* (logically) be capable of existing *un*experienced, and so the first panpsychist premise, that phenomenal properties cannot exist unexperienced, is arguably false.

I think that pan-phenomenalism is intuitively plausible and I agree with Coleman's argument that we can reject the premise that every phenomenal ultimate is a subject-of-experience. By ridding subjectivity from the phenomenal ultimates pan-phenomenalism therefore avoids the micro-subject/macro-subject gap. The consequence of this, however, is that pan-phenomenalism confronts another combination problem, this being the *non-subject/subject gap*.

By ridding subjectivity from phenomenal properties pan-phenomenalism must therefore provide an explanation of how non-subjective phenomenal ultimates can yield the subjective experience that we are all familiar with. Coleman puts forward a basic theory that sees conscious (subjective) awareness as being understood in terms of *phenomenal representation*: *i.e.*, "the representation of phenomenal quality by phenomenal quality."[57] What this theory holds is that the standard five sense organs serve the function of transmitting sensory

[57] Coleman, S. (2012) p. 159.

information (phenomenal input) up into the brain thereby causing a "phenomenal screen" to light up. Following this the phenomenal input is then centralised and *represented* within "the central perceptual/experiential domain of the subject" – this central perceptual/experiential *domain* is where subjective experience occurs.

I want to argue here that Coleman's *functional representationalism* theory fails to successfully resolve the *non-subject/subject gap*. The reason why I think it fails to bridge this gap is because it presupposes the existence of a subjective domain, this being the central phenomenal screen thus acting as the perceptual/experiential subjective domain. It is not clear to me how this domain operates or what is exactly responsible for generating it: Does this domain *emerge* through certain physical/brain activity? Or is it a more primitive domain that underlies high-order cognitive processing? Whatever is ultimately responsible for this subjective domain it cannot be phenomenal representation alone because phenomenal properties are in and of themselves ontologically devoid of subjectivity. Phenomenal properties therefore bear no metaphysical responsibility for generating this subjective domain. It seems to me that this domain must be functionally distinct from, though necessary for the *representation* of phenomenal qualities as processed within the brain. But phenomenal representation alone is insufficient for logically explaining the manifestation of subjective experience. Moreover, it certainly seems possible that phenomenal properties could undergo functional representation while remaining unexperienced because the subjective domain remains off-line. If this is possible, which I think it is, phenomenal representation is therefore metaphysically necessary but insufficient for yielding subjective experience.

Overall I find Coleman's version of pan-phenomenalism a highly plausible but currently incomplete theory. What it needs most, in my view, is a more robust account of the subjective domain, for this domain is ultimately responsible for bridging the non-subject/subject gap.

The last version of panpsychism that I want to examine here is pan-protopsychism. The idea of *proto* properties has recently gained traction as a promising metaphysical view for explaining the groundwork of phenomenal consciousness. David Chalmers has recently published a number of papers that critically assess the feasibility of *pan-protopsychism*.

2.3. – Pan-protopsychism and *the proto-phenomenal/phenomenal gap*

Pan-protopsychism is the view that the fundamental physical ultimates are *proto*-phenomenal. This is to say that the fundamental ultimates contain special properties that pertain specifically to the manifestation of certain mental states, specifically phenomenal conscious states. Chalmers writes that, "*protophenomenal* properties are special properties that are not phenomenal (there is nothing that it is like to have a single protophenomenal property), but can collectively constitute phenomenal properties, perhaps when arranged in the right structure."[58] I think it is important to assess here what Chalmers means by "special" properties. Chalmers writes that protophenomenal properties are special in that they are (i) distinct from structural properties, and (ii) there is an a priori entailment from protophenomenal properties to phenomenal properties. This second point refers to the *protophenomenal/phenomenal gap*, the first of many explanatory gaps facing pan-protopsychism. I agree with the first point, that protophenomenal properties are distinct from structural properties, however I think the *a priori* entailment from protophenomenal to phenomenal properties is more of a significant problem than Chalmers admits, mainly because it fails to account for subjectivity.

In my view, the most serious version of the combination problem for panprotopsychism is the *non-subject/subject gap*. How does subjectivity arise from a non-subjective (proto-conscious) foundation? Similar to how I argued that Coleman's functional representationalism fails to yield subjective

[58] Chalmers, D. (2013a), p. 13-14.

experience, I think the transition from protophenomenal properties to phenomenal properties also fails to yield subjective experience.

How then are we to understand subjective experience if the fundamental physical ultimates are devoid of subjectivity? Short of eliminating subjectivity all together, one approach is to view subjectivity as being an *emergent* phenomenon that occurs within complex physical systems, thus making subjective expereince a *systematic* phenomenon rather than a simply physical phenomenon. I will explore this idea of subjectivity being a systematic phenomenon in chapter four when examining Integrated Information Theory as a panpsychist theory of mind.

Another version of pan-protopsychism that I think is worth mentioning just briefly is what Chalmers has labelled *panqualityism*. Panqualityism can be formulated as a version of Russellian Monism. Russellian Monism is the view that all conscious experience is grounded in structural properties plus *quiddities*, with quiddities being "the fundamental roles specified by physics. Alternatively, we can say that quiddities are the categorical bases of the microphysical dispositions characterised by physics."[59] Quiddities are seen as providing the fundamental link between physical and mental properties, and the unification between mind and body is made possible only by this underlying 'common ground' of quiddities. Quiddities therefore serve as the neutral basis for all the laws of nature. A leading version of panqualityism is *constitutive Russellian panqualityism*, which holds the view that "qualities serve as quiddities and also serve to constitute human experience."[60] Chalmers explains that:

> *qualities* are the properties presented in experience: Intuitively, these are properties like redness, greenness, heat, and so on. It is important to note that qualities are not identical to phenomenal properties: when redness is presented to me in experience, I have a phenomenal property, but I need not be red. Instead, we would intuitively say that I am aware of redness, and that phenomenal properties involve awareness of qualitative

[59] Chalmers, D. (2013), p. 8.
[60] *Ibid.* p.25

> properties. Likewise, phenomenal properties are always instantiated by conscious subjects, but qualities need not be. We can certainly make sense of the idea of a red object that is not a subject of experience."[61]

Hence, qualities are protophenomenal properties that serve to constitute phenomenal properties, yet qualities are completely devoid of any subjectivity. One question to consider here is this: how do non-subjective qualitative properties yield subjective experience? This is a question that I cannot explore here, but as an alternative form of panpsychism, panqualityism also faces a *non-subject/subject gap*.

Clearly further philosophical investigation is required here for understanding how pan-protopsychism might bridge both the protophenomenal/phenomenal gap and the *non-subject/subject gap*. In my view, these two gaps are the most challenging of all explanatory gaps for panpsychism. Resolving these two gaps will become my aim in the fourth chapter when I develop Integrated Information Theory into a form of pan-protopsychism.

2.4 – Conclusion

My aim in this chapter has been to critically examine pan-experientialism, pan-phenomenalism, and pan-protopsychism as three contemporary versions of panpsychism. I have also looked closely at the combination problems for panpsychism, and how each respective version addresses these problems. I have argued that pan-experientialism and pan-phenomenalism both fail to successfully resolve their combination problems: pan-experientialism fails because there are strong arguments (*viz.,* Goff's epistemological argument and Coleman's metaphysical argument) against the summing of subjects; and pan-phenomenalism fails because phenomenal representation alone is metaphysically insufficient for yielding subjective experience – a subject domain is further required for resolving the non-subject/subject gap. And because pan-

[61] Chalmers, D. (2013), p. 24.

protopsychism has yet to be developed into a theory, there are currently no solutions offered for resolving the protophenomenal/phenomenal gap.

While each version of panpsychism has its respective merits, my conclusion here is that all three versions of panpsychism require further development to successfully resolve their combination problems. Given the strong arguments against subject-combination I cannot see how pan-experientialism might resolve the micro-subject/macro-subject gap. Without a solution to this problem I think pan-experientialism is the most implausible version of the three. I find Coleman's pan-phenomenalism the most promising version, but I think it needs to be further developed in regards to the subjective domain as a faculty in its own right. Greater clarity in this respect will better resolve the non-subject/subject gap for pan-phenomenalism. Pan-protopsychism will feature again in chapter four when I develop Integrated Information Theory (IIT) into a panpsychist theory of mind. I will also present solutions to both the protophenomenal/phenomenal gap and also the non-subject/subject gap. Prior to this I must firstly introduce IIT as a theory of consciousenss. IIT is a cutting-edge but controversial theory. In the next chapter I will turn my attention to IIT and I will begin by presenting a basic framework of the theory. Afterwards I will also examine whether or not it qualifies as a physicalist theory and how it stands up to the arguments against physicalism.

CHAPTER THREE

3.0 – Introduction

So far in this thesis I have examined two contemporary philosophical theories of mind: firstly, the prevailing view of physicalism in chapter one, and panpsychism in chapter two. In this chapter I will examine a recent theory of mind that has links to both physicalism and panpsychism. This theory is *integrated information theory*, or IIT for short.

Giulio Tononi first introduced IIT in 2004 and since then the theory has undergone two major revisions (Tononi, 2008; Oizumi, M. *et al*, 2014). IIT is a unique theory largely due to its methodology. Instead of taking physics and neurology as its starting point, IIT begins with phenomenology and works (back) towards the physical mechanisms of the brain, thereby developing a theoretical account of how the physical mechanisms must operate in order for them to generate phenomenal consciousness. IIT is an empirically rigorous theory and has been developed in line with the most recent neuroscientific evidence relating to brain function and psychological alterations, e.g. split-brain phenomena. Yet understanding how the brain "generates" or at least *becomes* conscious largely remains a mystery. IIT attempts to explain at least some of the mystery and in doing so the theory presents some very interesting philosophical aspects and implications within in it, especially panpsychist implications. My aim in this chapter is to firstly examine these philosophical aspects and implications and then assess how these relate to IIT in terms of understanding it as philosophical theory of mind.

I want to state clearly here that my aim in this chapter will be to focus strictly on the philosophical aspects and implications of IIT and not on the mathematical and quantitative methods used for measuring the quantity and quality of consciousness as formulated in IIT. To highlight these philosophical aspects and implications I will firstly present a basic theoretical framework of IIT wherein I will explicate in more detail the theoretical links between the phenomenological

axioms and the underlying physical mechanisms that are required for consciousness. Admittedly, the basic framework as I will present it is only a rough outline of IIT, but I believe it will be sufficient for my purposes in this chapter.

This chapter is split into two sections. In the first section I present the basic framework of IIT and I will provide a general exposition of the key terms and concepts involved. This exposition is important for the second section of this chapter, in which I will critically examine IIT as a physicalist theory of mind. Given its strong adherence to current neuroscientific and psychological evidence, IIT is by first approximation a physicalist theory of mind. By viewing IIT as a physicalist theory I will focus on how IIT stands up to the challenges against physicalism as set out in chapter one. This will involve an assessment of how IIT deals with the *knowledge argument,* the *conceivability argument*, as well as how IIT stands as a *realisation theory* of mind*.* In doing this I will highlight how IIT fails to conform to the more standard forms of physicalism, i.e. behaviourism, dualism, traditional identity theory, functionalism and computationalism. The reason why IIT fails to conform to traditional physicalism has largely to do with how integrated information is taken to be both a *fundamental* and *intrinsic* property of physical systems. I will conclude by arguing that it is the intrinsic nature of information as well as the claim that integrated information is a fundamental quantity that leads IIT away from physicalism and more towards panpsychism.

3.1 – Integrated Information Theory

Before examining IIT as a philosophical theory it is necessary to have a general understanding of its basic theoretical framework. Central to this framework are the phenomenological *axioms* relating to consciousness itself, and secondly the physical *postulates* relating to the underlying mechanisms responsible for generating consciousness. The phenomenological axioms define what IIT considers to be the self-evident features of consciousness. These axioms are: 1) *existence*: consciousness "is an undeniable aspect of reality"; 2) *composition*:

consciousness is compositional (structured); 3) *information*; consciousness is informative in that each experience differs from all other possible experiences; 4) *integration*: consciousness is integrated in that it is strongly irreducible to non-interdependent components; 5) *exclusion*: each conscious experience excludes all other possible experiences – "at any one time there is only one experience having its full content, rather than a superposition of multiple partial experiences."[62]

Holding these axioms to be self-evident and true to the nature of consciousness, IIT then proposes a set of physical postulates that correspond to the phenomenological axioms. These postulates specify the physical properties that systems must satisfy in order to generate consciousness. These postulates will become evident throughout this chapter so I will not provide details of them here. These axioms and postulates are then applied in IIT to present a bottom-up theoretical framework that illustrates, firstly, how the basic individual mechanisms are to operate in accordance with the postulates, then secondly, how systems of mechanisms must operate in order for the phenomenological axioms to be ultimately satisfied. In IIT the key feature that distinguishes an unconscious system from one that gains consciousness is *integrated information*. I will explicate this distinction in more detail by examining what Tononi claims to be a *minimally conscious* photodiode. By comparing and contrasting this minimally conscious photodiode with one that is unconscious I will illustrate how in IIT the physical mechanisms (and postulates) must operate in order to satisfy the phenomenological axioms, and how integrated information is the differentiating factor between an unconscious system and one that is conscious.

Although the theoretical process of IIT begins with phenomenology and works (back) towards the physical mechanisms, this process is also intended to be reversible such that IIT "offers a way to analyse systems of mechanisms to determine if they are properly structured to give rise to consciousness, how much of it, and of which kind."[63] This reverse methodology (from physics to

[62] See Giulio Tononi (2015) for specific definitions regarding these axioms.
[63] Oizumi, M. *et al* (2014), p. 2.

phenomenology) is important for practical scientific purposes, say in the case of needing to diagnose whether a brain-damaged patient is at all conscious. The aim of IIT is to be more than just a theoretical model of consciousness – IIT hopes to be practically useful for determining whether consciousness is or is not present in any sufficiently structured physical system.

3.2 – The basic framework of IIT

In what follows I will present the basic framework of IIT from the bottom-up. Through this framework I will firstly examine the individual mechanisms, then secondly a whole system of mechanisms. Afterwards I will critically examine this framework for the purpose of highlighting the philosophical aspects and implications of the theory. Individual elementary mechanisms are at the bottom of this framework, so these will be our starting point.

Individual mechanisms in IIT: distinguishing between unconscious and conscious mechanisms.

Tononi begins by defining an elementary mechanism in the form of a simple photodiode. A photodiode is a light-sensitive device containing two basic elements: a *sensor* that responds to variations in light (or electrical current) and a *detector* connected to the sensor that switches to "light" when the current is above a certain threshold and "dark" when below the threshold. Tononi then presents a thought experiment in which a photodiode and you, a conscious brain, are placed in front of a blank screen that is alternating on and off (between 'light' and 'dark'). As the blank screen is switched on and off the difference between light and dark is evident to us as we consciously observe the screen. At the same time the difference between light and dark is also evident to the photodiode due to the way in which the mechanism in the photodiode responds to the change in electrical current, thereby switching on or off. However, the photodiode is not seen as being in the slightest way conscious of the distinction between light and dark. That is, the photodiode has no conscious experience like we do – all is dark inside. So why does the photodiode lack consciousness?

According to IIT one reason why the photodiode lacks consciousness has to do with the quantity of information generated when discriminating between light and dark. To explain this Tononi writes:

> When the blank screen turns on, the mechanism in the photodiode tells the detector that the current from the sensor is above rather than below the threshold, so it reports "light"… For you, however, a light screen is different not only from a dark screen, but from a multitude of other images, so when you say "light", it really means this specific way *versus* countless other ways, such as a red screen, a green screen, a blue screen, this movie frame, and so on for every movie frame (not to mention for a sound, smell, thought, or any combination of the above)."[64]

In IIT the ability of a system to discriminate between alternative states is considered a determining factor for consciousness, and this ability is defined in terms of information. To see this, IIT states that when the photodiode switches from dark to "light" the mechanism is said to generate only 1 bit of information, while brain mechanisms, on the other hand, are said to generate vastly greater amounts of information. What ultimately distinguishes an unconscious system such as a photodiode from a conscious system is the system's ability to integrate information. Although this individual photodiode technically lacks integrated information (and therefore consciousness), further on I will look at how Tononi redesigns the photodiode in such a way for it to become *minimally conscious*. Prior to this, however, it is necessary to look at how the notion of integrated information is developed in the following thought experiment.

Mechanisms and integrated information: the camera thought experiment.

The camera thought experiment involves a digital camera with a sensor chip that consists of one million binary photodiodes, each of which contains a light sensor and a detector identical to the photodiode as described above. The camera as a

[64] Tononi, G. (2008), p. 217-218.

whole can therefore generate one million bits of information (or $2^{1,000,000}$ alternative states). According to IIT, the reason why the camera lacks consciousness is because "the [camera] chip is not an integrated entity: since its 1 million photodiodes have no way to interact, each photodiode performs its own local discrimination between a low and a high current completely independent of what every other photodiode might be doing… In other words, there is no intrinsic point of view associated with the camera chip as a whole."[65] This thought experiment highlights how IIT considers *integrated information* as the key feature for generating consciousness. Despite the vast quantity of information generated within the camera chip as a whole, the reason why the camera chip lacks consciousness is because the elementary mechanisms all operate in isolation. Put simply: no integration, no consciousness.

I think it is important to briefly discuss here the ontological status of elementary mechanisms as currently conceived in IIT. Within the current framework of IIT the notion of a mechanism "simply denotes anything having a causal role within a system, for example, a neuron in the brain, or a logic gate in a computer."[66] It is important to emphasise here that a basic elementary mechanism such as an individual photodiode is completely devoid of consciousness; there is nothing about the physics or the machinery for that matter that contains any trace of consciousness, and likewise for the camera chip containing one million photodiodes. To echo William James, the mere aggregation of elementary mechanisms yields nothing over and above the individual parts. In IIT the key for generating consciousness is not aggregation but integration, specifically integrated information. I will now look at how IIT describes integrated information as the key factor for generating consciousness.

Integrated Information in individual mechanisms and systems of mechanisms.

As stated above, each photodiode when operating generates 1 bit of information, meaning that it discriminates between two possible states: light and dark. The

[65] Tononi, G. (2008), p. 219
[66] Oizumi, M. *et al*. (2014), p. 3.

photodiode at this point does not possess any consciousness in and of itself, and similarly a computer chip with one million photodiodes all in operation remains completely unconscious. For consciousness to manifest within any complex system integration between the parts must be achieved.

For information to become integrated IIT states that "a system must be connected in such a way that information is generated by causal interactions *among* rather than *within* its parts. Thus a system can generate information only to the extent that it cannot be decomposed into informationally independent parts… In short, integrated information captures the information generated by *causal interactions* in the whole, over and above the information generated by its parts."[67] Tononi applies the symbol Φ to represent the integrated information generated by the system as a whole. A system gains consciousness when its Φ-value is greater than zero (Φ>0). More specifically, Φ represents the amount of information generated above and beyond the information present in each of the respective parts. To see this, the Φ-value of a system is measured by calculating the *relative entropy* (the difference between probability distributions) pertaining to the individual mechanisms within a system. Consider for example a system that is composed of two photodiodes, each with a mechanism generating 1 bit of information. This system as a whole thus contains 2 bits of information. If the two mechanisms remain disintegrated in such a way that no interaction is taking place between them, the relative entropy (or the probability distributions) between the two mechanisms does not exceed the total distribution of the system as a whole. In other words, the system as a whole is nothing more than the sum of its parts. As a result relative entropy equals zero: the two parts generate 2 bits of information respectively and because no integration is taking place between the parts, the system as a whole (which is simply the two individual mechanisms taken together) yields only 2 bits of information. Ultimately, the system as a whole does not generate information above and beyond the information generated by its parts, in which case Φ = 0, thus leaving the system unconscious.

[67] Tononi, G. (2008), p. 221.

However, when causal interactions do take place *among* rather than within the parts of a mechanism, IIT claims that such a mechanism will generate integrated information. The *information postulate* states that, "information is meant to capture the "differences that make a difference" from the perspective of the system itself – and is therefore both causal and intrinsic."[68] What I think is of philosophical interest here is how IIT holds information as being both *causal* and *intrinsic* to the physical activity of mechanisms and the consequential generating of consciousness. The intrinsic nature of information is one of the central philosophical aspects of IIT. I will examine this in more detail further on for it is one of the key reasons why IIT is considered to have panpsychist implications.

The integration postulate and the irreducible nature of integrated information.

The integration postulate holds that at the level of individual mechanisms, such as neurons in the brain or logic gates in a computer, "only mechanisms that specify integrated information can contribute to consciousness. Integrated information is information that is generated by whole mechanisms above and beyond the information generated by its parts. This means that, with respect to information, the mechanism is irreducible."[69] The irreducible nature of integrated information leads IIT to introduce the notion of "small phi" (φ) and "high phi" (Φ). Small phi (φ) relates to the integrated information generated by individual mechanisms, whereas "high phi" (Φ) refers to the integrated information at the system level, that being a system of mechanisms. Distinguishing between these two notions of phi is important for understanding the role they play in generating consciousness.

Small-phi (φ) and maximally irreducible cause-effect repertoire.

At the level of individual mechanisms, small phi is generated by what IIT calls the *maximally irreducible cause-effect repertoire* (*MICE*). The *exclusion* postulate

[68] Oizumi, M. *et al* (2014), p. 6.
[69] *Ibid*. p. 8.

holds that, in respect to this cause-effect repertoire, "a mechanism can have only one cause and one effect, those that are maximally irreducible; other causes and effects are excluded. The *core cause* of a mechanism from the intrinsic perspective is its maximally irreducible cause repertoire".[70] This core cause of the MICE in a mechanism is responsible for generating what IIT calls a *core concept*. In IIT each core concept (henceforth concept) is *qualitative*: the MICE of the mechanism "specifies what the concept is about (its *quale "sensu stricto"*), and its particular φ^{MAX} value… quantifies its amount of integration or irreducibility."[71] Most importantly, concepts are held as being phenomenal properties in the form of qualia. A basic or primitive concept might specify a particular shade of blue or perhaps a particular note in a chord of music. I will examine the phenomenal nature of concepts in more detail further on, but what I want to highlight here is that in IIT a concept (or *qualia*) is generated by the MICE of a mechanism and is specified by the phi-value, which states both the quantity and quality of integrated information intrinsic to a mechanism.

Systems of mechanisms (Φ) and their maximally irreducible conceptual structure.

At the system level, individual mechanisms can be interconnected to form systems of mechanisms. Whereas each individual φ mechanism generates a specific qualitative concept, a whole system of integrated mechanisms generates what IIT calls *integrated conceptual information* (Φ). When a power-set[72] of multiple mechanisms is interactive, wherein each mechanism is generating conceptual information in the form of a concept (qualia), this powerset set generates what IIT calls *conceptual structure*. The exclusion postulate states that, "only a conceptual structure that is *maximally* irreducible can give rise to consciousness – other constellations generated by overlapping elements are

[70] Oizumi, M. *et al* (2014), p. 9.
[71] See Figure 9 in Oizumi, M. *et al* (2014) for specific details regarding how a 'concept' is generated by the MICE of mechanisms.
[72] A power set consists of multiple interconnected mechanisms wherein each active mechanism generating conceptual information. This conceptual information, also referred to as a "constellation of concepts", is then "plotted" in concept space, thereby specifying both the quantity and quality of consciousness intrinsic to the system. See Figure 10 in Oizumi, M. *et al* (2014), p. 12.

excluded."[73] When conceptual structure is present in a system IIT states that the system is in possession of what it calls a *complex*. A complex is defined as follows:

> A *complex* is thus defined as a set of elements within a system that generates a local maximum of integrated conceptual information Φ^{Max} (meaning that it has maximal Φ as compared to all overlapping sets of elements). Only a complex exists as an entity from the intrinsic perspective… Once a complex has been identified, concept space can be called "*qualia space*" and the constellation of concepts can be called a "*quale 'senso lato'* [quale in the *broad sense*]. A quale in the broad sense of the word is therefore a *maximally irreducible conceptual structure (MICS)* or, alternatively, an *integrated information structure*.[74]

In IIT a complex is considered as being a single conscious entity with its own intrinsic 'point-of-view' replete with phenomenal experience. Tononi goes on to claim that at any given time "there may be a single *main complex* of comparatively much higher Φ that underlies the dominant experience (a main complex is such that its subsets have strictly lower Φ)."[75] Since a complex is held as being a conscious entity, and that a conscious system can possess not only a dominant complex (Φ^{MAX}) but also a subset of complexes with strictly lower phi, IIT appears to posit the existence of multiple mental 'points-of-view' within, for instance, the brain, such that there can be a *main complex* with multiple sub-complexes underlying it. Further on in this chapter I will present what I think are some philosophical issues concerning the relation between low φ and high Φ complexes, for I think that there are a number of combinatorial problems facing IIT. I will also address these issues in more detail in the next chapter when considering IIT as a form of panpsychism.

Summarising the basic theoretical framework of IIT from the bottom-up.

---

[73] Oizumi, M. *et al* (2014), p. 13.
[74] *Ibid.*
[75] Tononi, G. (2008), p. 221.

1) When individual elementary mechanisms causally interact thus generating integrated information (φ), this integrated state involves a *maximally irreducible cause-effect repertoire* (MICE), and this repertoire is the *core cause* underlying the generation of a *concept*. A concept is a conceptual structure that specifies phenomenologically "what" the MICE is about, e.g. $MICE^{BLUE}$ is a mechanism generating a phenomenal blue concept.

2) A set of integrated mechanisms forms a whole system. At the system level, the integration postulate states that only integrated conceptual structures can give rise to consciousness. When multiple φ mechanisms are systematically integrated they constitute an irreducible conceptual structure containing *integrated conceptual information* (Φ). Only *maximally irreducible conceptual structure* (MICS) can give rise to consciousness in the form of a complex.

3) When a system generates MICS the systems is said to possess a *complex*. A complex is a local maximum of integrated conceptual information (Φ^{MAX}) and only a complex exists as an entity possessing intrinsic perspective and phenomenal, subjective experience. A complex is therefore a conscious mind.

I have presented so far a broad outline of the basic theoretical framework of IIT. My aim here has been to focus on how IIT conceives the physical mechanisms based on the physical postulates, and to examine how these mechanisms must operate in order to satisfy the phenomenological axioms of consciousness. More generally, my aim has been to critically examine these physical mechanisms for the purpose of identifying how these mechanisms correlate to the proposed phenomenological states intrinsic to conscious systems. In what follows, I will consider recent criticisms of IIT and then, by viewing IIT as a physicalist theory of mind, I will analyse how IIT stands up to the arguments against physicalism as featured in chapter one.

3.3 – Understanding IIT as a physicalist theory of mind

Due to the strong empirical and neurophysiological basis underlying its theoretical framework, IIT is by first approximation a physicalist theory of mind. Prior to examining IIT as a physicalist theory of mind, I want to consider some recent criticisms that raise some philosophical issues with the theory.

3.4 – *Understanding the intrinsic nature of information*

Since first publication in 2004 IIT has undergone two revisions that take into account recent peer-review analysis and, in particular, criticisms that raise various concern about the mathematical calculus used for quantifying Φ (See Gamez, 2011; Beaton and Aleksander, 2012; Barrett, 2014; Tegmark, 2015). One of the most problematic issues for any information-based theory of mind is the very idea of information itself. All three formulations of IIT adopt the mathematical definition of information as developed by Claude Shannon (1948) in his mathematical theory of communication. Shannon's theory employs logarithmic functions for measuring the information "bits" in discrete systems such as the photodiode used in IIT. Tononi applies Shannon's theory for quantifying the information in a photodiode, the calculation being $\log_2(2) = 1$ bit of information as a result of the mechanism being in 1 of 2 possible alternative states, either light or dark. The more states a system can be in, the more information the system contains. For example, throwing a fair dice yields $\log_2(6) = 2.58$ bits of information due to the six possible states the dice can be in. One issue with using Shannon's data theory of information is that the quantitative calculus yields only what can be described as *extrinsic* information and not *intrinsic* information. One problem for IIT concerns the way in which extrinsic information is observer dependent, meaning that the information is relevant to the 'grain level' or the point at which the measurement is being taken. Adam Barrett highlights this problem and explains how Shannon's theory of information is a theory about "data" and not a theory about truly intrinsic information. Barrett explains that, "information can only be intrinsic to fundamental physical entities, and descriptions of information in systems modelled at a non-fundamental level necessarily rely on an extrinsic observer's

choice of level."[76] On claiming that Shannon's information theory fails to capture truly intrinsic information, Barrett proposes what he calls the *Field Integrated Information Hypothesis (FIIH)*.

Barrett hypothesises that "consciousness arises from information intrinsic to fundamental fields, and… to move IIT forward, what is needed is a measure of intrinsic information applicable to the configuration of a continuous field."[77] The FIIH views the universe as being composed of fundamental fields (e.g. gravitational fields, electromagnetic fields, etc.) which includes all the relevant particles that emerge due to the fluctuating nature of such fields. The FIIH holds that fundamental fields, including their particles, exist as the fundamental physical correlates of consciousness, and consequently, consciousness arises from information intrinsic to fundamental fields. Barrett's FIIH will become more important in the next chapter when discussing IIT as a version of panpsychism. This will also involve looking at how information is understood as being an intrinsic property of physical systems. And Barrett also claims that his FIIH has panpsychist implications in much the same way as IIT.

3.5 – Is IIT more a theory about qualia and less a theory about consciousness?

Anthony Peressini (2013) has criticised IIT as failing to solve the hard problem of consciousness. Since IIT starts with phenomenology and then develops in theory an account of how the underlying physical mechanisms must operate in accordance with the physical postulates, IIT appears to present phenomenology as the metaphysical basis of consciousness. By taking phenomenology as the basis of consciousness IIT presumes that all other aspects of consciousness such as perception and awareness, as well as the 'what-it-is-likeness' of subjective experience, are ultimately explicable in phenomenological terms. Peressini argues that, "the problem of subjective experience (SIL-consciousness) ought not to be thought of as identical to the problem of qualia."[78] According to Peressini,

[76] Barrett, A. (2014), p. 1.
[77] *Ibid.* p. 1
[78] Peressini, A. (2013), p. 190.

IIT is more a theory about qualia and less a theory about subjective (SIL) consciousness. By SIL-consciousness Peressini means the 'Something-It-is-Like-ness' in the same sense as Nagel's famous term 'what-it-is-likeness' relating to first-person subjective experience. SIL-consciousness refers to being a conscious *experiencer* whose experience consists of a rich multitude of phenomenal/qualitative mental states. Peressini goes on to claim that, "Tononi is making the same mistake that many philosophers make in this context, namely conflating SIL-consciousness of a creature [with] the qualitative properties of mental states."[79] The mistake, according to Peressini, is in thinking that phenomenal properties fully capture and in some way contain subjectivity such that subjective experience can be reduced to qualia and therefore ideally explained in phenomenological terms. I agree with Peressini's criticism here for the reason that reducing subjectivity to qualia appears back-to-front: qualia are generally taken to be properties belonging to and present within subjective experience, not the other way around.

Although I agree with Peressini here, I also think the philosophical problem that Peressini's is trying to highlight relates very closely to what Tim Bayne (2010) has called *the unity of consciousness*. What is most evident about human consciousness is not only the fact that it consists of a vast multitude of phenomenal properties, but that our conscious experience is phenomenally unified. It is the phenomenal unity of consciousness that forms what we commonly call *subjective* conscious experience. Bayne states the unity of consciousness as follows:

> Let us say that a subject has a unified consciousness if, and only if, every one of their conscious states is phenomenally unified with every other conscious state. We can think of such subjects as *fully unified*. Where a subject is fully unified, we can say that they enjoy a single *total conscious state*.[80]

[79] *Ibid*.
[80] Bayne, T. (2010), p. 15.

With the *unity of consciousness* thus stated by Bayne, I think another way to view Peressini's argument is that of understanding how the rich variety of phenomenal properties (including the five standard sense modalities) can be all fully unified into an overall holistic conscious experience. One problem for IIT as I see it is that the theory could be correct in claiming that integrated information yields phenomenal properties but incorrect in claiming that such integrated states yield the fully unified subjective consciousness that we are all familiar with. Peressini claims that IIT is better titled as IITQ: integrated information theory of *qualia*. If IITQ is true then the problem of understanding the *unity of consciousness* remains unresolved, the reason being that integrated information states (*i.e.* φ) do not *a priori* entail unified subjective consciousness at the system level (*i.e.* Φ). I will argue in the next chapter that there are philosophical issues facing IIT by holding all integrated information states (φ and Φ) as being subjective states; my argument on this matter will come under the section titled *IIT and the combination problems for panpsychism*.

3.6 – Why IIT is anti-functionalism.

Michael Cerullo (2011) has criticised IIT as failing to satisfy the principles of *structural coherence* and *organisational invariance*. Since IIT holds that consciousness is based in integrated information, Cerullo points out that IIT does not qualify as a form of behaviourism, dualism, or classical identity theory. Thus IIT is most closely aligned with functionalism. Viewing IIT as a functionalist theory, Cerullo reiterates David Chalmers' argument that any functionalist theory must satisfy the principles of structural coherence and organisational invariance. Chalmers defines the *principle of structural coherence* as being the "coherence between the *structure of consciousness* and the *structure of awareness.*"[81] By *awareness* Chalmers refers specifically to the contents of awareness which are to be understood as "those information contents that are accessible to central systems, and brought to bare in a widespread way in the control of behaviour. Briefly put, we can think of awareness as *direct availability*

[81] Chalmers, D. (1995), p. 17.

of global control." (*Ibid*) Chalmers goes on to state that awareness is a purely *functional notion* that is intimately linked to conscious experience, mainly because of the way in which consciousness involves awareness:

> Whenever there is a conscious experience, there is corresponding information in the cognitive system that is available in the control of behaviour, and available for report and global control. Conversely, it seems that whenever information is available for report and global control, there is a corresponding conscious experience. Thus, there is a *direct correspondence between consciousness and awareness*. (Chalmers, 1995. *my italics*)

Cerullo claims that IIT fails to satisfy the principle of structural awareness by arguing that "a person could be conscious of distinctions in the real world but not be able to act on them."[82] To illustrate his argument Cerullo uses the example of a perceptual illusion known as the 'simultaneous contrast illusion'. This illusion involves a solid bar of a set shade held in front of a dissimilar coloured background. When the colour gradient of the background changes, the shade of the bar appears to change as well when in fact it does not. To someone naïve about the illusion they are not aware that the bar remains the same shade throughout the entire experience and cannot report this information either verbally or in their behaviour. However, information about the colour of the bar is still available from retinal inputs, and the lower level visual processing stream does contain this information while the higher-level visual stream fails to realise this information, possibly because it uses heuristics to bias perception for other cognitive purposes such as predictability. Hence there is discordance between the lower and higher levels of the cognitive system, the consequence being that higher level processing fails to accurately generate the right experience, and only the lower level processing is correct in respect to the information about the shade of the bar. Cerullo therefore claims that "some parts of the brain are capable of distinguishing possibilities in the world (i.e. contain integrated

[82] Cerullo, M. (2011), p. 50.

information) yet these perceptions never reach awareness."[83] Thus, IIT fails to satisfy *structural coherence*.

Regarding the principle of *organisational invariance*, Chalmers defines the principle as, "any two systems with the same fine grained *functional organisation* will have qualitatively identical experiences. If the causal patterns of neural organisation were duplicated in silicon, for example, with a silicon chip for every neuron and the same patterns of interaction, then the same experiences would arise. According to this principle, what matters for the emergence of experience is not the physical makeup of the system, but the abstract pattern of causal interaction between its components."[84]

Cerullo highlights how organisational invariance implies a one-to-one mapping such that any two (or more) identical conscious experiences will necessarily involve the same functional organisation. However, if the mapping is not one-to-one then it is possible that a specific experience can be generated by multiple (perhaps even infinite) computations; "therefore there is nothing unique about computation (i.e. the fine-grained organisational structure) that relates it to experience and this destroys the vital link between experience and computation that is at the heart of functionalism."[85]

Cerullo therefore restates the principle of organisational invariance as: "Any two systems share the same fine-grained functional organisation *if and only if* they have qualitatively identical experiences."[86] By taking two systems that are isomorphic in terms of their computational functions, Cerullo demonstrates how there are other computational possibilities that generate the same *quantity* of integrated information yet differ in terms of their *values*, with each value representing one of the many other different probability states that the systems can be in. And because probability states reduce to information states and not functional or physical states, when two functionally identical systems possess

[83] Cerullo, M. (2011), p. 50.
[84] Chalmers, D. (1995), p. 20.
[85] Cerullo, M. (2011), p. 51.
[86] *Ibid*. p. 51.

different 'Φ' values then their *qualitative* experiences will differ as well. Cerullo's central point here is that computational functionalism does not categorically determine phenomenality: identical functional states can be in differing probability states and thus generate non-identical experiences. Thus IIT fails as a functionalist theory of mind.

These criticisms that I have raised so far can be summariesd as follows: Barrett claims that IIT fails to capture truly intrinsic information in its quantification of phi 'Φ'; Peressini argues that Tononi conflates the problem of subjectivity with the problem of qualia, which is a mistake; and Cerullo argues that IIT fails to satisfy the principles of *structural coherence* and *organisational invariance*, which means that IIT fails to be a functionalist theory of mind. I will now examine IIT as a physicalist theory of mind and how it stands up to the arguments against physicalism.

3.7 – Understanding IIT as a realisation theory

Earlier in section 3.2 I looked at what IIT considers to be an unconscious photodiode. In this section I want to look at how IIT presents what it calls a *minimally conscious* photodiode. By comparing and contrasting the unconscious photodiode with one that is minimally conscious, my aim here is to examine how a physical system such a photodiode is said to *realise* consciousness. A central tenet of IIT is that consciousness can be multiply realised and that neural activity is not necessary for generating consciousness.[87] This is demonstrated by IIT in relation to a *minimally conscious* photodiode. This minimally conscious photodiode is described as consisting of two elements: a detector *D* and a predictor *P*. When *D* is switched on by external inputs it activates element *P* which serves as a "memory" and also acts as a "predictor" of the next external input to *D*. Simple though it may appear, IIT claims that this photodiode satisfies the postulates of IIT: "both of its elements specify selective causes and effects

[87] See Figure 18 in Oizumi, M. *et al* (2014), p. 19 for details as to how an inactive system can generate a quale and why neural activity is not necessary for generating consciousness.

within the system (each element about the other one), their cause effect repertoires are maximally irreducible, and the conceptual structure specified by the two elements is also maximally irreducible."[88] Due to the cause effect repertoire in the system, the photodiode forms a complex, "albeit one having just two concepts and a Φ^{MAX} value of 1 (Figure 19C)."[89] Thus the photodiode is considered conscious, albeit minimally so. IIT therefore claims that throughout the duration of the repertoire the photodiode is having an actual conscious experience. IIT states that "the experience might be described roughly as "it is like this rather than not like this", without further qualifications."[90] But I think further qualification is definitely needed here and I want to call into question whether the minimally conscious photodiode can be said as having an actual conscious experience.

Firstly, regarding the intrinsic perspective of the photodiode IIT claims that, "the experience is only minimally specified, and in no way can [the experience] convey the meaning "light": *D* says something about *P*'s past and future state, and *P* about *D*'s, and that is all".[91] Why this photodiode should qualify as having an actual conscious experience is by no means clear to me. While the mechanism might satisfy the postulates of IIT, the conscious experience as had by the photodiode certainly leaves a lot to be desired. For instance, is the photodiode having a truly subjective experience? Apparently not, because the experience is meaningless and unintentional. Moreover, what IIT states here is simply a functional description: the fact that "*D* says something about *P*'s past and future state, and *P* about *D*'s" does not imply nor logically entail for that matter that an experience is being had by the system. I am therefore very sceptical that this photodiode is minimally conscious as IIT claims it to be.

All that I am willing to attribute to the photodiode in terms of being conscious is simply an experience of "this-ness", which might in reality be nothing more than a simple flash of "red-ness" or "C-major-ness" or another other simple token of

[88] Oizumi, M. *et al* (2014), p. 19.
[89] *Ibid.*
[90] *Ibid.*
[91] *Ibid.*

phenomenal experience. What I contend here is that the photodiode may not in fact be minimally *conscious* but rather minimally *phenomenal* – *i.e.*, there is "redness" being generated but no actual subjective experience as such. My contention here brings us back to the philosophical issue of whether qualia are intrinsically subjective or whether qualia can exist without being a subjects-of-experience. Recall in chapter two the differing views between Strawson and Coleman: Strawson claims that *all* experience is necessarily subjective experience, whereas Coleman argued that phenomenal properties could have objective reality without being intrinsically subjective. In respect to the photodiode, I am more inclined to believe the photodiode as generating and being in possession of a phenomenal concept/qualia. But given the "minimal" state of the photodiode, what I am not willing to attribute to it is an actual subjective, conscious experience. What might this mean for IIT? The photodiode, as I see it, is simply a qualia generator. If this is the case then integrated information does not appear to necessarily *realise* subjective experience, and this in turn calls into question whether integrated information does in fact generate subjective conscious experience as IIT so claims.

3.8 – IIT and The Conceivability Argument

In chapter one I examined the conceivability argument against physicalism, in which Chalmers presents the notion of a philosophical zombie. Recall that a philosophical zombie is someone physically, functionally and behaviourally identical to me but who completely lacks phenomenal consciousness – all is dark inside. Strangely enough, IIT claims that "zombie" systems are in fact possible. IIT confirms this by stating that, "it is conceivable that an unconscious system could show the same input/output behaviour as a "conscious" system… Thus, IIT admits the possibility of true "zombies", which may behave more and more like us while lacking subjective experience".[92] In this section I want to focus on and distinguish between philosophical zombies and what I will call *IIT zombies*, and whether the conceivability argument against physicalism also affects IIT. I argue

[92] Oizumi, M. *et al* (2014), p. 20-21.

that it does by presenting a conceivability argument against IIT. I will also consider how IIT might defend itself against my argument.

With philosophical zombies already described, I now want to look at how IIT describes a zombie system. An *IIT zombie* is a system that operates as a feed-forward system. A feed-forward system is in every way physically and functionally identical to a conscious replica system (*i.e.* one that contains a phi-complex) except that it has been "unfolded" in such a way that all the elementary mechanisms fail to interact sufficiently and in the necessary manner for integrating information. To see this in more detail, IIT demonstrates that over time (using 4 time steps) the state of each element in the system is fed forward through a chain of nodes, one node for each time step. "In this way, the states of upstream elements in previous time steps can be combined (converge) in a feed-forward manner to determine the state of elements down stream, but can never feed back on the elements upstream."[93] While the configuration of a feed-forward system has the same input/output functions as its conscious counterpart, one critical difference between these functionally identical systems has to do with the lack of "feed-back", "re-entry", or "recursion" connections in the zombie system. IIT therefore states that feed-back and re-entry connections could play a vital role in generating consciousness, so much so that "the presence or absence of feed-back could be directly equated with the presence or absence of consciousness."[94] Though feedback may play a crucial role in generating consciousness IIT does not view it as being a necessary condition. Rather, IIT states that "it is the potential for interactions among the parts of a complex that matters and not the actual occurrence of "feed-back" or "re-entrant" signalling, as is usually assumed… a complex can be conscious, at least in principle, even though none of its neurons may be firing, no feed-back or re-entrant loop may be activated, and no "ignition" may have occurred."[95] Though it is difficult to see how a complex system can be conscious without neural activity and with no feed-back and re-entry loops activated, what appears to be the necessary

[93] Oizumi, M. *et al* (2014), p. 21.
[94] *Ibid*. p. 19.
[95] *Ibid*. p. 20.

condition for consciousness in IIT is the *potential for interactions among the parts* of a complex, which, in other words, is *integrated information*.

An IIT zombie is therefore a system whose parts have been unfolded in such a way that the requisite *interactions among the parts* are lost, which entails the loss of integrated information and therefore consciousness. Hence, an IIT zombie is physically identical to a conscious replica system, functionally identical in respect to its input/output functions, and is behaviourally identical as a consequence, yet completely lacks subjective experience. With IIT zombie thus defined, I now want to highlight the critical difference between philosophical zombies and IIT zombies.

The critical difference between philosophical zombies and IIT zombies is that of a *structural* difference. This structure difference is found in the fact that a philosophical zombie has not been unfolded in the way an IIT zombie has: philosophical zombies instantiate all the same physical and functional mechanisms, logical circuitry and information networks as found in a conscious human brain, all the while lacking phenomenal consciousness. Again, the critical difference is a structural difference: whereas a philosophical zombie has identical physical structure to its conscious counter-part, namely a human being, the IIT zombie has a very different physical structure due to being unfolded into a feed-forward system. So while they are physically identical in every other respect to philosophical zombies, IIT zombies are structurally very different.

Distinguishing between these two zombies is important for the following two reasons. Firstly, it is important to recognise that IIT zombies are not to be confused with nor thought of as being identical to philosophical zombies: IIT zombies are structurally very different from philosophical zombies. And secondly, by viewing IIT as a physicalist theory, it is worth considering whether the conceivability argument still affects IIT in the same way that it affects physicalism. In what remains of this section I want to consider how the conceivability argument might affect IIT.

The conceivability argument as developed by Chalmers turns on the notion of *logically possibility*: *i.e.* given two physically identical systems it is logically possible that while one is phenomenally conscious (that being me) the other (my zombie twin) remains phenomenally unconscious. The conceivability argument therefore illustrates the possibility that when given all the physical facts about the world, as well as all the logical truth relating to those facts, nothing about the physical fact logically entails the instantiation of phenomenal consciousness. Hence the possibility of philosophical zombies. Now, if IIT is correct in stating that consciousness *is* integrated information, what makes human beings conscious therefore is the way in which our brains integrate information thus giving rise to a phi-complex 'Φ^{MAX}'. Recall that a phi-complex has *qualia-space*, which is in effect the phenomenological domain of subjective experience. But if we grant IIT as being true in respect to the physical postulates, thereby holding the theory correct in terms of all the physical and functional facts – especially the functional facts – it is conceivable, *i.e.* logically possible, that a system capable of integrating information (hint: my zombie twin) still remains phenomenally unconscious. That is, given all the physical and functional facts about the brain, including any facts that we have about *information* (of which there are few), it is conceivable that a complex integrated information system such as the human brain remains phenomenally unconscious. The problem as I see it here is that nothing about the physical postulates logically entails any truths about phenomenal consciousness. Thus, I think the following conceivability argument can be lodged against IIT: that it is logically possible for a system to integrate information all the while remaining phenomenally unconscious.

I think there are two difficulties here for IIT regarding the conceivability argument as I have presented it above. The first difficulty is that the notion of 'integration' is a functional notion: if something is integrated it means that it is composed in such a way or that it functions in such a manner. And secondly, the concept of 'information' (which is not to be confused with the compound concept of 'integrated information') is *not* a phenomenological concept. It follows therefore that a physical system such as the human brain could be functioning perfectly and in accordance with the physical postulates and be processing

information within a phi-complex, all the while being phenomenally unconscious. What this means is that even if IIT is correct about integrated information being necessary for consciousness, the possibility for there to be philosophical zombies remains nonetheless. The consequence of this is that integrated information is a necessary but insufficient condition for phenomenal consciousness. Since it appears logically possible for an integrated information system to remain phenomenally unconscious, I think the conceivability argument against physicalism also stands as an argument against IIT.

To defend itself against the conceivability argument as I have presented it IIT needs a basic metaphysical principle that provides a logical link between physical/functional processes and phenomenal consciousness. I want to briefly consider one strategy that I think might assist IIT in explaining the logical connection between physical mechanisms of the brain and phenomenal consciousness.

To provide a logical connection between the physical and the phenomenal, I think one strategy available to IIT is to adopt David Chalmers *dual aspect theory of information*. On dual aspect theory Chalmers writes that:

> Information (or at least some information) has two basic aspects, a physical aspect and a phenomenal aspect. This has the status of a basic principle that might underlie and explain the emergence of experience from the physical. Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing."[96]

I want to elaborate here on Chalmers' *dual aspect theory* because I think it is relevant to my discussion both here in relation to the conceivability argument and in the next chapter when I focus on IIT as a form of panpsychism. My understanding of Chalmers' dual aspect theory is that the world – or be it the universe – is not purely physical but that it is also informational, which is to say

[96] Chalmers, D. (1995), p. 23.

that world contains not just physical properties but also informational properties, specifically information states. The physical aspect of Chalmers' theory views these information states as being "embodied" in physical processing, which in IIT would be the physical processing of any sufficiently complex system (most obviously the brain) that is capable of integrating information. The phenomenal aspect of information views conscious experience as "emerging" through the physical processing of information, which in the context of IIT is to say that conscious experience emerges by virtue of the integrated information intrinsic to the physical processes in the brain. Therefore, when information is integrated within physical processing (*e.g.* in the brain) phenomenal properties emerge not from anything physical *per se* but rather from the information states that are embodied and intrinsic to such physical processes.

By holding information as being intrinsic to physical processes in the manner just described, I think that IIT can potentially reject my conceivability argument on the grounds that information states, specifically *integrated information* states, are phenomenal states by virtue of information having a phenomenal aspect. Further philosophical scrutiny is needed on this matter, but I believe that Chalmers' *dual aspect theory* is compatible with IIT for the purpose of defending the theory against the conceivability argument and the logically possibility of philosophical zombies. I will focus more on the notion of information in chapter four when considering the panpsychist implications that arise by holding information as a fundamental and intrinsic property of the world.

3.9 – IIT and The Knowledge Argument

In chapter one I discussed the *knowledge argument* against physicalism, focusing specifically on Frank Jackson's argument involving Mary. The aim of the knowledge argument is to refute physicalism by demonstrating that knowledge of all the physical facts fail to *a priori* entail knowledge about phenomenal consciousness. From this Jackson infers that some facts about the world are non-physical, namely phenomenal facts, thereby concluding that physicalism is false.

Tononi (2008) refers to the knowledge argument in IIT and I find his response to it particularly interesting. I want to consider here whether Tononi's response can be thought of as a successful response to the knowledge argument.

In IIT phenomenal experience is fully described by the conceptual structure or the set of informational relationships that formulate a phi-complex. This description is captured in the central identity claim in IIT, which states that: *an experience is identical with the maximally irreducible conceptual structure (MICS, integrated information structure, or quale "senso lato") specified by the mechanisms of a complex in a state*. While IIT claims that in principle it is possible to *know* what sort of qualitative experience a complex is having, Tononi goes on to write that, "the IIT also implies that to be conscious – say to have a vivid experience of pure red – one needs to *be* a complex of high Φ; there is no other way. Obviously, although a full description can provide understanding of what experience is and how it can be generated, it cannot substitute for it: *being is not describing*."[97] I think Tononi makes a completely valid point here, which is that 'redness' cannot be properly known outside the context of *being* a complex having a red experience; the experience of redness exists only as a state or property of a complex. The point here is that if 'redness' is an intrinsic state of a system then it becomes practically impossible to know exactly what a particular experience is like from the extrinsic, empirical point of view; to know anything about about redness one must firstly *be* a complex before one can know anything about phenomenal consciousness.

Tononi also writes that, "The [knowledge] argument loses its strength the moment one realizes that consciousness is a way of being rather than a way of knowing. According to the IIT, being implies "knowing" from the inside, in the sense of generating information about one's previous state. Describing, on the other hand, implies "knowing" from the outside."[98] For Tononi the reason why the knowledge argument loses its strength is because there are two equally legitimate epistemic points of view: one is from the outside, which is empirical,

[97] Tononi, G. (2008), p. 234.
[98] *Ibid*. p. 234.

objective and descriptive, whereas the other is from the inside, which is mental, phenomenal and psychological. And only when a complex *is being* generated by mechanisms activated by the sensory inputs associated with colour will someone like Mary come to *know* what a colour visual experience is like.

What I find interesting to think about here is what Mary might say after learning about IIT while still in her black and white room. To explain consciousness, Mary might say something like this: i) to be conscious is to be a complex; ii) to be a complex is to be in a state of high Φ; iii) Φ is understood as *maximally irreducible conceptual structure* (MICS) or *integrated information structure* (*quale 'senso lato'*); and iv) the MICS and MICE are understood as a set of informational relationships generated by mechanisms and systems of mechanism respectively. Yet even with this knowledge of IIT in mind it still seems that Mary does not know what it is like to experience colour vision due to her on-going confinement. Knowledge of IIT therefore appears to be a little help to Mary, for she is still unable to logically deduce any facts about the phenomenal consciousness of other people.

However, because IIT holds that information is intrinsic to physical systems and that consciousness is integrated information, to know everything there is to know about conscious experience, including all phenomenological facts, is practically impossible from the outside, *i.e.* from the external, empirical point of view. This is because, to echo Bertrand Russell, physics tell us nothing about the intrinsic nature of the world. In IIT the only way to know about the intrinsic nature of the world is from the inside and the inside only, that is, by *being* conscious. While I think that Tononi presents a reasonable defence against the knowledge argument, I am not convinced that IIT can fully capture the *quality* of conscious experience as it claims it does, for it still seems that Mary, even with IIT in mind, cannot know all that much more about phenomenological facts of colour experience. Even if IIT can quantify consciousness, I am not convinced that it can logically deduce the quality of consciousness from analysing states of integrated information.

3.10 – Conclusion

My purpose in the chapter has been to explore the philosophical aspects and implications of IIT. After presenting the basic framework of IIT I turned my attention to some of the recent criticisms of IIT. The first of these came from Adam Barrett, who argues that Shannon's theory of information fails to properly capture truly intrinsic information. To advance IIT in respect to the notion of intrinsic information Barrett claims that truly intrinsic information exists only within fundamental physical entities, specifically fundamental fields, and proposes what he calls the *Field Integrated Information Hypothesis (FIIH)*. Barrett's FIIH also has panpsychist implications and I will consider his hypothesis in more detail in the next chapter.

Anthony Peressini criticises IIT as failing to resolve the hard problem of consciousness for the reason that Tononi conflates the problem of subjectivity with problem of qualia. I agreed with Peressini on this point, that the problem of subjectivity is not to be equated with the problem of qualia. Peressini therefore argues that IIT is more a theory about qualia – IITQ as he calls it – rather than a theory about subjective consciousness. Michael Cerullo also criticises IIT on the grounds that it fails to satisfy the principles of *structural coherence* and *organisational invariance*. Thus IIT fails to be a functionalist theory.

I then turned my attention to understanding IIT as a realisation theory of mind by focusing on what IIT presents as a minimally conscious photodiode. Given the "minimal" nature of the experience, I argued that this minimally conscious photodiode does not appear to have actual subjective experience. In my view, the photodiode appears more so to be a qualia generator rather than a minimally conscious subjective mind. My criticism here relates back to Cerullo's claim that IIT is more a theory about qualia than it is about subjective consciousness.

I then distinguished between philosophical zombies and what I called IIT zombies. Distinguishing this is important because IIT zombies are structurally very different from philosophical zombies. These two zombie systems are not to

be thought of as identical to one another. I considered whether the conceivability argument against physicalism might still affect IIT, and I argued that it is conceivable for a physical system generating integrated information to remain phenomenally unconscious. The implication here is that integrated information is a necessary but insufficient condition for generating phenomenal consciousness, which means that philosophical zombies are metaphysically possible creatures even if IIT is true. Thus, the conceivability argument against physicalism can also be laid against IIT. I considered one strategy that IIT could use as a defence against the conceivability argument, which is to adopt Chalmers' *dual aspect theory* of information. Since this theory holds that information has both a physical aspect and a phenomenal aspect, IIT could argue that the metaphysical basis of phenomenal consciousness is information, specially integrated information, and not anything physical *per se*. Because information has this phenomenal aspect, any integrated information state 'Φ' is necessarily phenomenally conscious. If this strategy is successful philosophical zombies are inconceivable creatures because integrated information metaphysically necessitates phenomenal consciousness by virtue of information having a phenomenal aspect. I believe that *dual aspect theory* is compatible with IIT and, if further developed, I think it could be used as a defence against the conceivability argument.

Finally I considered IIT in relation to the knowledge argument against physicalism. Tononi himself responds to the knowledge argument by stating that consciousness is not a matter of describing but rather a matter of *being*; that is, to know anything about consciousness one must firstly *be* conscious in order to describe consciousness. Tononi therefore dismisses the knowledge argument on the grounds that there are two equally legitimate epistemic points of view: that of knowing from the outside (*i.e.* external, empirical observation) and that of knowing from the inside (intrinsic, private subjective experience). I considered what Mary might say about phenomenal consciousness once having learnt about IIT in her black and white room. I argued that IIT appears to be of little help to Mary and that she cannot know exactly what the phenomenal facts of other people might be like even with her knowledge of IIT in mind. That Mary cannot

know what other phenomenal facts might be like does not, in my view, undermine IIT as a physicalist theory. The reason why I think this is because IIT could be construed as a very broad form of physicalism, perhaps in the form of say, Stoljar's o-physicalism. IIT clearly describes consciousness as an intrinsic property, and if qualia do qualify as categorical properties in the sense of being shapes in the qualia-space of a phi-complex, then IIT could qualify as a non-reductionist form of (broad) physicalism.

But as we have seen throughout this chapter, IIT fails to conform to the standard physicalist positions of behaviourism, dualism, classical identity theory, functionalism and computationalism. I therefore conclude that IIT cannot be considered a physicalist theory of mind, at least not in any traditional form. What I find most interesting about IIT is the fact that it has panpsychist implications. In my next and final chapter I will assess these panpsychist implications within IIT and then examine which version of panpsychism is most suitable for it. This will also involve identifying the combination problems most relevant to IIT. My aim in the next chapter is to demonstrate how I think IIT can resolve these combination problems and thereby succeed as a panpsychist theory of mind.

CHAPTER FOUR

4.0 – Introduction

In the previous chapter I examined IIT as a physicalist theory, with my conclusion being that IIT fails to conform to the standard physicalist positions of behaviourism, dualism, classical identity theory, functionalism and computationalism. Thus, IIT cannot be considered a physicalist theory of mind. Because IIT is claimed to have panpsychist implications, my aim in this chapter is to examine IIT as a panpsychist theory of mind. IIT is a unique and somewhat peculiar theory of mind; unique because it is the only theory that identifies consciousness as integrated information, and peculiar because it posits integrated information as a fundamental quantity – as fundamental as mass, charge, and energy. In this chapter I will look at what motivates IIT towards panpsychism. Tononi and Koch openly state that although IIT was not developed with panpsychism in mind the theory does have panpsychist implications. I will critically examine these implications and in doing so I will identify which of the three contemporary versions of panpsychism best suits IIT. This will also involve identifying the combination problems that apply to IIT as a panpsychist theory. I will demonstrate that IIT does not qualify as a form of either pan-experientialism or panphenomenalism but that it does qualify as a form of pan-protopsychism. IIT therefore faces two combination problems: i) the proto-phenomenal/phenomenal gap, and ii) the non-subject/subject gap. My ultimate task in this chapter is to present solutions to both these problems, thereby demonstrating that IIT succeeds as a panpsychist theory of mind.

4.1 The motivations for panpsychism in IIT

I think there are two claims in IIT that motivate the theory towards panpsychism. The first claim is that integrated information exists as a fundamental quantity – as fundamental as mass, charge, and energy. And because IIT identifies consciousness as integrated information, this entails that consciousness is a fundamental feature of the world. The second motivation can

be seen in how IIT views consciousness as being a graded property; "consciousness is not an all-or-none property, but is graded: specifically, it increases in proportion to a system's repertoire of discriminable states."[99] According to IIT, a photodiode can be minimally conscious as long as it is generating 1 bit of integrated information, and the greater the repertoire of the system the greater the consciousness overall. The minimally conscious photodiode is used as an example for demonstrating how consciousness can be multiply realised, with the added purpose of arguing that neural activity is not necessary for generating consciousness; any system capable of integrating information will, according to IIT, be conscious to some degree. The implication here is that many sufficiently complex systems can possess consciousness but only if they contain the requisite mechanisms for integrating information. In light of these implications, particularly the claim that integrated information (consciousness) is a fundamental quantity, I now want to consider the question: how close is IIT to panpsychism?

With the traditional definition of panpsychism in mind, that everything in the universe has some degree of consciousness, Tononi states that "many entities, as long as they include some functional mechanism that can make choices between alternatives, have some degree of consciousness."[100] This response immediately rules out *pure panpsychism*, the view that everything in the universe has a mind or is conscious. Thus, tables, chairs, cups and saucers are not conscious entities for obvious reasons. Clearly, biological entities such as ourselves possess the necessary physical mechanisms required for generating consciousness, which, given our natural evolutionary history, raises the question as to how far down the biological chain of organic life-forms consciousness might be found. In response to this question Tononi and Christof Koch have provided some clarification regarding their general position towards panpsychism. In a paper titled *Consciousness: here there but not everywhere* Tononi and Koch (2014) write that the IIT is not at odds with panpsychism insofar as integrated information is posited as a fundamental quantity: wherever there is integrated information,

[99] Tononi, G. (2008), p. 236.
[100] *Ibid*.

there is consciousness, and because integrated information is a fundamental quantity, so too is consciousness. The following paragraph captures how Tononi and Koch view IIT in relation to panpsychism:

> IIT was not developed with panpsychism in mind (*sic*). However, in line with the central intuitions of panpsychism, IIT treats consciousness as an intrinsic, fundamental property of reality. IIT also implies that consciousness is graded, that it is likely widespread among animals, and that it can be found in small amounts even in certain simple systems. Unlike panpsychism, however, IIT clearly implies that consciousness is not ubiquitous. Moreover, IIT offers a solution to several of the conceptual obstacles that panpsychists never properly resolved, like the problem of aggregates (or combination problem (James, 1980, Chalmers 2013)). It also explains why consciousness is adaptive, and can account for its quality. (Tononi and Koch, 2014, p. 6-7)

Panpsychism can also be seen in what Tononi describes as a "Φ-centric" view of the universe. By accepting integrated information as a fundamental quantity – as fundamental as mass, charge, and energy – Tononi writes that a valid view of the universe is this:

> A vast empty space that contains mostly nothing, and occasionally just specks of integrated information (Φ) – mere dust, indeed – even there where the mass-charge-energy perspective reveals huge conglomerations [with these conglomerations being planets, stars and galaxies]. On the other hand, one small corner of the known universe contains a remarkable concentration of extremely bright entities (where brightness reflects high Φ), orders of magnitude brighter than anything around them. Each bright "Φ-star" is the main complex of an individual human being (and most likely, of individual animals).[101]

[101] Tononi, G. (2008), p. 233.

This phi-centric view of the universe sees consciousness (in the form of Φ-stars) as being prevalent throughout the universe. Moreover, what this view entails is that consciousness is not to be seen as a phenomenon exclusive to life on Earth. Rather, if the universe is inherently geared for generating integrated information then we should expect to find various other conscious life-forms throughout the universe. On phi-centrism, Φ is just as much an inherent part of the universe as are mass, charge and energy. While I find Tononi's phi-centric view of the universe utterly fascinating, I also think that there are a number of philosophical problems with it. The most serious problem, in my view, is that of understanding IIT as a form of panpsychism, including the respective combination problems relevant to it. The combination problems for panpsychism are important problems and must be openly addressed by any theory claiming to be a version of panpsychism, and I think IIT needs greater clarity regarding thes problems.

Any theory claiming to be a version of panpsychism must either resolve the combination problem directly, or demonstrate why the combination problems do not apply to it. Although IIT was not developed with panpsychism in mind, Tononi and Koch claim that the IIT offers a solution to several of the conceptual/metaphysical problems confronting panpsychism, specifically the combination problem. It is not clear to me why Tononi and Koch claim this when they never openly address the combination problems for panpsychism, nor do they provide any explicit details as to how these problems are in fact resolved by IIT. Of course, Tononi and Koch could avoid the combination problems by asserting the point that because IIT was never developed as a panpsychist theory of mind the combination problem are simply irrelevant to it. However, simply dismissing the combination problems out of hand is unacceptable. This is because the combination problems are theoretically important to panpsychism and cannot simply be ignored. I think that IIT requires greater clarity in this respect: if IIT is a form of panpsychism then it must demonstrate how it resolves the combination problems. Failing such demonstration will mean that Tononi and Koch are not entitled to claim that IIT resolves the combination problems as they in fact claim. If IIT is to qualify as a panpsychist theory then it must be able to demonstrate how it resolves the combination problems for panpsychism, or

why the problems do not apply to it. To clarify this matter I will now consider which version of panpsychism is most suitable to IIT, and this will then determine which of the various combination problems are relevant to it as a form of panpsychism.

4.2 – Panpsychism? How IIT compares to contemporary versions of panpsychism.

In this section I will compare and contrast IIT against the three contemporary versions of panpsychism that I examined in chapter two. These three versions of panpsychism are: Strawson's *pan-experientialism*; Coleman's *pan-phenomenalism*; and Chalmers' *pan-protopsychism*. My purpose in doing this is to identify which version of panpsychism best applies to IIT, and this will determine which combination problems are relevant to it.

IIT states that conscious experience, along with phenomenal consciousness, exists only within the context of integrated information. And while integrated information is held as being a fundamental quantity, it is not ubiquitous. As I will now explain, IIT does not align with either pan-experientialism nor pan-phenomenalism for the following reasons.

Pan-experientialism holds that conscious *experience* (or experiential being) is "at the very bottom of things", which is to say that conscious experience is the *essential* nature of physical reality and the material world.[102] Pan-experientialism therefore defines reality in experiential and non-experiential terms: mind and consciousness are metaphysically grounded to experiential being, whereas anything non-conscious, *i.e.* anything simply physical/material, is based in non-experiential being. In Strawson's view, the only substance in reality is *energy*, and energy can take the form of experiential (mental) and non-experiential (physical) being.[103]

[102] Strawson, G. (2006), p. 186.
[103] For more detail regarding Strawson's pan-experientialism see: Strawson *Mind and Being* in G. Bruntrup and L. Jaskolla (OUP) 2015.

What differentiates IIT from pan-experientialism is the fact that in IIT *integrated information* is seen as constituting the intrinsic nature of reality. More precisely, in IIT conscious phenomenal experience is defined by the set of information relationships pertaining to the *qualia space* of a phi-complex. I think that two things can be said here that disqualify IIT as a version of pan-experientialism.

Firstly, since integrated information is not ubiquitous nor everywhere, it follows then that conscious experience *is not everywhere*, which means that conscious experience does not permeate all things, nor is it the essential nature of material reality. Tononi himself confirms this by writing that, "Information that is not integrated… is not associated with experience".[104] IIT is very clear on this point, which is that conscious experience exists only where there is integrated information, and since integrated information is not a ubiquitous feature of the world, neither is conscious experience. Secondly, IIT does not view conscious experience as something fundamentally intrinsic. Instead, *integrated information* is stated as an intrinsic property, and conscious experience is defined in terms of the information relationships that formulate the qualia-space in a Φ-complex. Because IIT does not view conscious *experience* as being fundamental and ubiquitous in the sense that pan-experientialism holds it to be, IIT does not qualify as a form of pan-experientialism.

IIT also stands contrary to *pan-phenomenalism* as developed by Sam Coleman. One of Coleman's central arguments was that it is logically possible for phenomenal/qualitative properties to exist without being necessarily bound to a subject-of-experience. Coleman advanced an 'Independence Argument' for pan-phenomenalism, claiming that phenomenal properties can exist independently of and outside the awareness of subjective experience. Coleman argues that, "There simply cannot be a property, a real feature of reality, which gains its existence precisely and only through someone's awareness of it… In fact, phenomenal qualities *must* (logically) be capable of existing unexperienced."[105] On pan-phenomenalism, qualitative properties (qualia) exist as properties of objects in

[104] Tononi, G. (2008), p. 233.
[105] Coleman, S. (2012), p. 152.

the external world, *i.e.,* properties that can exist outside and external to subjective conscious experience. Qualia are therefore not wholly mental properties but instead exist external to and independent of conscious awareness.

In comparison to pan-phenomenalism IIT does not view qualitative properties as existing outside the boundary of integrated information. IIT is very clear in its definition of qualia. The precise definition of a quale in IIT is this: *the conceptual structure generated by a complex in a state that corresponds to a local maximum of integrated conceptual information* Φ^{MAX} *(synonymous with "MICS" or "constellation" in "qualia space").*[106] From this definition it is clear how qualia (phenomenal properties) exist only within the *qualia space* generated by a phi-complex. It follows from this that qualia do not have objective external reality. Instead, qualia have only mental reality, for they exist only within the domain of integrated information. And because integrated information *is not everywhere*, neither are qualia. Thus, IIT does not qualify as a form of pan-phenomenalism.

The third version of panpsychism that I considered in chapter two is pan-protopsychism. Pan-protopsychism is defined by Chalmers as the view on which "fundamental physical entities are protoconscious. In more detail, let us say that *protophenomenal* properties are special properties that are not phenomenal (there is nothing it is like to have a single protophenomenal property), but can collectively constitute phenomenal properties, perhaps when arranged in the right structure".[107] Chalmers states furthermore that protophenomenal properties are to be considered as "special" properties that bear an especially close connection to phenomenal properties. Chalmers claims that this appeal to specialness can be understood by requiring that: i) protophenomenal properties are distinct from structural properties; and ii) there is an *a priori* entailment from protophenomenal properties to the phenomenal properties that they constitute. Of these two requirements I think the second one is the more difficult to explain and involves the following two tasks. The first task is to specify *what* exactly qualifies as a protophenomenal property, and then secondly, *how* these

[106] Oizumi, M. *et al*. (2014), p. 4.
[107] Chalmers, D. (2013), p. 13.

protophenomenal properties are able to collectively constitute phenomenal properties. Understanding the entailment between proto-phenomenal and phenomenal properties also involve bridging the *protophenomenal/phenomenal gap*.

Of the three contemporary versions of panpsychism considered so far, I believe that IIT does not qualify as a form of pan-experientialism nor pan-phenomenal. I do believe, however, that IIT does qualify as a form of pan-protopsychism. In the next section I will demonstrate why I believe this, and the reason has a lot to do with how information is held as being intrinsic to fundamental physical properties, specifically fundamental fields. So far I have argued that IIT does not conform to pan-experientialism or pan-phenomenalism, but that it does qualify as a form of pan-protopsychism. As a form of pan-protopsychism the first combination problem that IIT runs up against is the protophenomenal/phenomenal gap. There are other combination problems that confront IIT in this form, namely the *non-subject/subject gap*. In the following two sections I will focus on the combination problems for IIT as a form of pan-protopsychism and I will investigate how IIT can resolves these problems.

4.3 IIT and the Combination Problems for Panpsychism

As I highlighted in the previous section Tononi and Koch claim that the IIT "offers a solution to several of the conceptual obstacles that panpsychists never properly resolved, like the problem of aggregates (or combination problem…)"[108] While it may be clear to Tononi and Koch how IIT manages to resolve the combination problems for panpsychism, they never formally demonstrate how these solutions are achieved. The combination problems for panpsychism must be dealt with seriously by any theory aligning itself with panpsychism, for these problems stand as major theoretical gaps that must be closed if such a theory is to successfully qualify as a form of panpsychism.

[108] Tononi, G. and Koch, C. (2014), p. 6-7.

There are many combination problems for panpsychism (See Chalmers 2013b). Of all the combination problems for panpsychism, two of the most challenging problems are:

[1] Micro-phenomenal/Macro-phenomenal gap – how do micro-phenomenal properties combine to yield macro-phenomenal properties?
[2] Proto-phenomenal/phenomenal gap – how do proto-phenomenal properties combine to yield macro-phenomenal properties?

What I want to demonstrate here is that [1] does not apply to IIT but that [2] does apply to IIT. The reason why [1] does not apply to IIT is rather straightforward. It has to do with the fact that the IIT does not explicitly divide phenomenal properties – *qualia* – into distinct micro and macro categories. In IIT phenomenal properties (synonymous with *qualia*) are understood within the context of integrated information, specifically within "the conceptual structure generated by a complex in a state that corresponds to a local maximum of integrated conceptual information Φ^{MAX}."[109] This conceptual structure is also referred to as 'qualia space' or Q-space, which is defined as a set of *informational relationships*. Tononi clarifies the qualitative nature of Q-space by stating that:

> the set of informational relationships in Q generated by the mechanisms of a complex in a given state (q-arrows between repertories) specify a shape in Q (a quale). Perhaps the most important notion emerging form this approach is that *an experience is a shape in Q.* According to the IIT, *this shape completely and univocally specifies the quality of experience*.[110]

It is clear from what Tononi says here that phenomenal properties/qualia are to be understood as *shapes* delineated by the informational relationships formulating the conceptual structure/qualitative space of a phi-complex. Again, IIT never categorically divides qualia into micro and macro properties; all qualia exist as specific shapes within the overall *maximally irreducible conceptual*

[109] Oizumi, M. *et al*. (2014), p. 4.
[110] Tononi, G. (2008), p. 228.

structure of Q-space. Given this definition of qualia, IIT avoids the *micro-phenomenal/macro-phenomenal gap*.

Since phenomenal properties are understood in terms of informational relationships (shapes) in Q-space belonging to a phi-complex, IIT therefore holds that phenomenal properties exist only within the context of integrated information. Because IIT does not posit the existence of conscious experience outside the context of integrated information, phenomenal properties do not pervade the world in the sense of having external, mind-independent reality. Contrary to pan-phenomenalism, IIT claims that phenomenal properties exists strictly within the context of integrated information, and while stated as being fundamental, integrated information is not ubiquitous, which means that phenomenal properties are not ubiquitous either. Hence why IIT does not conform to pan-phenomenalism.

By denying the foundations of reality as possessing phenomenality I think IIT faces a metaphysical problem. The problem here is one of understanding how phenomenal properties can be generated with a phi-complex at the fundamental level when all other fundamental properties are non-phenomenal in nature. To reiterate the problem, IIT needs to explain how integrated information as a fundamental quantity *a priori* entails the existence of phenomenal properties from an otherwise non-phenomenal foundation. Resolving this problem I think requires looking closer at the information postulate and the role it plays in yielding phenomenal consciousness.

Because the information postulate states that information is both causal and intrinsic, one way for IIT to explain how the integration of information yields phenomenal properties is to hold information as being proto-phenomenal. As a form of pan-protopsychism IIT must do two things: Firstly, it must specify what the fundamental physical entities are that qualify as being proto-phenomenal, and then secondly, explain how these proto-phenomenal properties *a priori* entail phenomenal properties. By viewing IIT as a form of pan-protopsychism I

now want to focus on how I think IIT can bridge the protophenomenal/phenomenal gap.

4.4 – Bridging the proto-phenomenal/phenomenal gap for IIT.

There are two parts to the project of explaining how IIT can bridge this gap. The first part of the project is to specify what exactly qualifies as a proto-phenomenal property. This requires making some claim about the ontological status of fundamental physical entities, and in particular the *intrinsic* nature of such entities. The second part of the project is to then explain *how* protophenomenal properties *a priori* entail phenomenal properties. This involves explaining how the collective constitution of protophenomenal properties *a priori* entails phenomenal properties and conscious experience. Given that in IIT phenomenal properties exist strictly within the context of *integrated information*, one way to think about protophenomenal properties is in relation to how *information* is considered as both causal and intrinsic.

In the previous chapter I explained how IIT applies Claude Shannon's theory of information for the purpose of quantifying integrated information as an overall Φ-value. Chalmers explains how Shannon himself was not concerned with a *semantic* notion of information, whereby information is seen as information *about* something. Instead, Shannon's account of information is more a *syntactic* notion of information. Shannon's definition of information is defined by Chalmers as follows:

> The most basic sort of information is the *bit*, which represents a choice between two possibilities: a single bit (0 or 1) selected from a two-state space is said to carry information. In a more complex case, a "message" such as "0110010101" chosen from a space of possible binary messages carries information in a similar way. What is important, on Shannon's account, is not any *interpretation* of these

> states; what matters is the *specificity* of a state within a space of different possibilities.[111]

Chalmers formalises Shannon's notion of information by appealing to the concept of *information space*. "An information space is an abstract space consisting of a number of states [called] *information states*, and a basic structure of *difference relations* between those states."[112] In IIT the information postulate defines information as "differences that make a difference", which can be described in more detail as: the *differences between information states* is what makes a difference to a system from its own intrinsic perspective. Using a light switch as an example, Chalmers explains that despite the infinite number of positions a light switch can be in, there are only two states of the switch that are relevant to the light being on or off, these being "up" and "down". It is the difference between these two states – "up" and "down" – that are the *differences that make a difference* to the light being on or off. But as I pointed out in chapter three, Adam Barrett is critical of Shannon's theory of information because of the way in which information is quantified by an external observer in relation to non-fundamental discrete states such as those of the photodiode used in IIT. I want to return now to Adam Barrett's criticism of IIT and how he conceptualises truly intrinsic information.

Barrett's main contention with IIT and its application of Shannon's theory of information is that "information can only be intrinsic to fundamental physical entities, and descriptions of information in systems modelled at a non-fundamental level necessarily rely on an extrinsic observer's choice of level."[113] Barrett argues that the proposed Φ-values do not properly quantify integrated information as a truly fundamental quantity. The reason for this is that complex systems such as the photodiode are currently modelled on discrete elements, e.g. the discrete states of a photodiode, and these states are non-fundamental states. The problem is that when it comes to the measuring of integrated information

[111] Chalmers, D. (1996), p. 278.
[112] Chalmers, D. (1996), p. 278.
[113] Barrett, A. (2014), p. 1.

'Φ' within any discrete system, the measurements are being made at the level chosen by an extrinsic observer instead of being taken at the fundamental level. To properly capture truly intrinsic information Barrett claims that, "[a] true measure of intrinsic integrated information must be frame invariant, just like any fundamental quantity in physics. That is, it must be independent of the point of view of the observer".[114] To better conceptualise the fundamental nature of intrinsic information Barrett proposes what he calls the *Field Integrated Information Hypothesis (FIIH).*

The FIIH proposes that intrinsic information exists only as a property of fundamental physical entities with such physical entities being fundamental fields. Barrett explains that in contemporary physics "fields" are postulated as being the fundamental physical ingredients of the universe, and the more familiar quantum particles arise due to the fluctuating nature of fields.[115] I think it is worth stating here how Barrett defines the nature of fundamental fields as understood in theoretical physics:

> In theoretical terms, a field is an abstract mathematical entity, which assigns a mathematical object (e.g. scalar, vector) to every point in space and time… So, in the simplest sense, a field has a number associated with it at all points in space. At the very microscopic scale, ripples, i.e., small perturbations, move through this field of numbers, and obey the laws of quantum mechanics. These ripples correspond to the particles that we are composed of, and there is precisely one fundamental field for each species of fundamental particle."[116]

Two examples of fundamental fields are the electromagnetic and gravitational fields, with the corresponding quantum particles being the photon and graviton respectively. Barrett explains that all particles are divided into matter particles and force-carrying particles, and all particles have an associated field such that

[114] Barrett, A. (2014), p.1.
[115] See Barrett (2014) for a more comprehensive definition of how fundamental fields are understood in contemporary physics.
[116] *Ibid.* p.1.

"all the forces of nature can be described by field theories which model interactions, i.e., exchanges of energy, between fields."[117] On the FIIH, fundamental fields are seen as the physical correlates of consciousness. This brings Barrett to hypothesis that:

> Consciousness arises from information intrinsic to the configuration of a fundamental field. The amount of consciousness generated by a patch of field is the amount of integrated information intrinsic to it. When a patch of field generates a large quantity of intrinsic integrated information, mathematically there is a high-dimensional information structure associated with it. The geometrical and topological details of this structure determine the contents of consciousness. The task now is to correctly mathematically characterise intrinsic integrated information, and construct equations to measure it.[118]

Since the current Φ measurements in IIT apply only to discrete systems, Barrett's main argument in favour of the FIIH is that due to fundamental fields being continuous in space, "a new mathematical formalism" is needed to properly quantify the intrinsic integrated information present within fundamental fields. [119] Consequently, if fundamental fields are responsible for generating consciousness then the FIIH (like IIT) does imply a form of panpsychism, one that sees the universe populated by "germs" of consciousness in the form of integrated information states, *i.e.*, Φ-complexes. And while *not everywhere*, Φ-complexes are held as being fundamental elements of the universe – as fundamental as mass, charge and energy.

Aware of the discomfort that many feel towards panpsychism Barrett writes that another way to think about consciousness being fundamental to physical reality is to attribute "potential consciousness" to matter. It is this notion of *potential* consciousness that I think can play an important role in understanding IIT as a

[117] Barrett, A. (2014), p. 2.
[118] *Ibid.*
[119] *Ibid.*

form of panpsychism, specifically pan-protopsychism. A lot hinges, however, on what is meant by the word 'potential' and how potential consciousness is metaphysically relevant to actual consciousness. On the notion of potential consciousness Barrett writes that:

> The quantity of potential consciousness is simply the quantity of integrated intrinsic information. But only when there is a large amount of intrinsic integrated information with a sufficiently rich structure to be worthy of being compared to a typical healthy adult human waking conscious moment, should we say that the integrated information has "actual consciousness" associated with it."[120]

I want to take issue here with how Barrett distinguishes between "potential consciousness" and "actual consciousness", and I think the distinction is problematic for two reasons. The first reason is that the distinction Barrett makes between potential and actual consciousness is *chauvinistic*. This is due to the comparison he makes between the integrated information as had by a *typical healthy human being* and all other non-human integrated information states. But IIT stands firm on the point that consciousness is identical to integrated information, so no matter how distant an integrated information state might be from that of human consciousness, it is nevertheless actually conscious even if minimally so. Obviously a minimally conscious photodiode has nowhere near the same level of consciousness as a human being, however I do not think human consciousness should be the measuring board for determining whether or not something is considered actually conscious, especially when IIT clearly states that the photodiode is minimally conscious.

Secondly, Barrett's distinction between potential and actual consciousness distorts the identity claim that consciousness is integrated information. Other than distinguishing it from human consciousness, Barrett never actually draws a definitive line between potential and actual consciousness, which makes it hard to see where potential consciousness ends and actual consciousness begins.

[120] Barrett, A. (2014), p. 4.

Moreover, to draw such a line is by no means an objective measure of consciousness, for such a line would again be arbitrary due to it being drawn from the perspective of an external observer.

This problem is made all the more difficult by the fact that, according to Barrett, "potential consciousness would still be assigned phenomenal content, so it is perhaps more elegant to just use a single term "consciousness" for the whole spectrum of integrated information."[121] Barrett is clearly aware of the difficulty in distinguishing between potential and actual consciousness, so much so that he almost gives up on the term "potential" altogether in preference of the single term "consciousness". I agree with Barrett's last point that the term consciousness should apply to the whole spectrum of integrated information no matter how minimal it might be. The benefit of doing this is that it avoids drawing an arbitrary line between human and non-human integrated information states. However, doing away with the notion of "potential" consciousness is no longer useful for comforting those who feel uneasy with panpsychism, which is why Barrett introduced the notion in the first place. I now want to present another way to think about the notion of potential consciousness, which I think will help with understanding IIT as a version of pan-protopsychism.

In my view, the word 'potential' is synonymous with the prefix 'proto': for something to be *potentially conscious* is for it to be *proto-conscious.* Rather than applying the notion of "potential consciousness" to integrated information as Barrett does, I think the notion can be applied more specifically to the most basic fundamental physical entities themselves, namely fundamental fields. As Barrett explained, fundamental fields are seen as possessing intrinsic information, which is equivalent to saying that the *intrinsic nature* of fundamental fields is *pure information*. By applying Chalmers definition of information to fields, where each field can be understood as an information state within a larger information *space* (one that contains perhaps an infinite number of information states) the fundamental nature of the universe can be seen as being one large information

[121] Barrett, A. (2014), p. 4.

domain. Chalmers also refers to such a view as the "it from bit" conception of the world, in which "it" (the physical world) emerges from "bits" of information, thus making information fundamental to the nature of reality.[122] Because I take the notion of 'potential' as being synonymous with the prefix 'proto', I want to propose here that the information intrinsic to fundamental fields plays a *proto-conscious* role in the form of being *proto-phenomenal*.

I think it is worth noting here that a number of electromagnetic field theories of consciousness have been recently developed, with each theory hypothesising in their own respective ways that phenomenal consciousness (including perception) is based in and generated by the electromagnetic patterns within the brain (See Libet, 1994; McFadden, 2002; Pockett, 2013). One such theory is Susan Pocket's electromagnetic (EM) field theory of consciousness, on which "conscious perceptions (and sensations, inasmuch as they can be said to have independent existence) are identical to certain spatiotemporal electromagnetic patterns generated by the normal functioning of waking mammalian brains."[123] Barrett's FIIH builds on Pockett's EM theory by claiming that integrated information is the key factor when it comes to correlating phenomenal consciousness with fundamental fields: conscious perception and phenomenal sensation exist only when the information intrinsic to the electromagnetic fields is integrated in the form of a Φ-complex. It follows from the notion of integrated information that when electromagnetic fields are *not* integrated phenomenal consciousness cannot and does not exist. Tononi himself is very clear on this point, stating that, "information that is not integrated… is not associated with experience."[124] My proposal here is that fundamental fields are protophenomenal entities by virtue of the intrinsic information existent within them. By holding fundamental fields as being protophenomenal I think IIT now has a foundation from which it can explain how the fundamental ultimates of the universe *a priori* entail phenomenal properties within integrated information

[122] Chalmers, D. (1996), p. 303.
[123] Pockett, S. (2013).
[124] Tononi, G. (2008), p. 233.

states (*i.e*. phi-complexes). I now want to propose the following solution to the protophenomenal/phenomenal gap for IIT:

Solution to the protophenomenal/phenomenal gap for IIT.

Fundamental fields are protophenomenal entities by virtue of possessing intrinsic information. When the information in fundamental fields become integrated, thereby forming an integrated information state (Φ-complex) the integrated information generates a *maximally irreducible conceptual structure*, which is synonymous with a 'quale' and 'qualia space'. Phenomenal properties are *shapes* in qualia space and are determined by the information relationships (q-arrows/q-edges) that formulate Q-space within a Φ-complex. In short: fundamental fields are proto-phenomenal states that *a priori* entail phenomenal properties when integrated in the form of a Φ-complex.

My first aim in presenting the above solution has been to firstly reconsider how the notion of *proto* (or potential) consciousness can be applied to fundamental physical entities. My second aim here has been to demonstrate that, as a form of pan-protopsychism, IIT is able to successfully bridge the proto-phenomenal/phenomenal gap. The solution I have presented above makes one amendment to Barrett's FIIH: whereas Barrett applies the notion of potential consciousness to instances of integrated information which lack the sufficiently rich conscious experience as had by human beings, I propose instead that proto/potential consciousness should be applied to the information intrinsic to the fundamental fields themselves. To summarise: fundamental fields are proto-phenomenal properties by virtue of possessing intrinsic information. Only when fundamental fields become *integrated* do they generate phenomenal properties within the *maximally irreducible conceptual structure* (qualia-space) as had by a Φ-complex.

So far I have focused on IIT as a form of pan-protopsychism and I have presented a solution to the protophenomenal/phenomenal gap. Resolving this problem however leads leads to another major combination problem for IIT. The

problem is that, while proto-phenomenal properties *a priori* entail phenomenal properties, they do not *a priori* entail subjective conscious experience. This is because proto-phenomenal properties are ontologically devoid of subjectivity; there is nothing it is like to be a proto-phenomenal property. All that proto-phenomenal properties can generate are phenomenal properties or qualia. Likewise, qualia are not in and of themselves subjective entities either, for they are simply the qualitative aspects of subjective conscious experience. In the previous chapter I agreed with Peressini's argument that IIT conflates the problem of subjective consciousness with the problem of qualia. In Peressini's view, IIT is more a theory about qualia than it is a theory about subjective consciousness. The consequence here is that as a form of pan-protopsychism IIT runs up against the *non-subject/subject gap*. I now want to turn my attention to this combination problem and how I think IIT might be able to resolve it.

4.5 – IIT and the non-subject/subject gap

In the final section of this thesis my aim is to focus on how IIT might explain the manifestation of subjective consciousness as a property of integrated information. I think it is worth reiterating Peressini's criticism of IIT just to clarify the issue here. The issue is that "the problem of subjective experience (SIL-consciousness) [*i.e.* subjective experience] ought not to be thought of as identical to the problem of qualia… That is, the question of the nature of a creatures SIL-consciousness is not assumed to be the same as the question of the nature of the creatures [phenomenal]-conscious states."[125] Peressini's criticism here is that the problem of phenomenal consciousness is not the same as the problem of understanding the 'what-it-is-like-ness' of subjective experience, which is to say that phenomenal properties do not fully capture nor explain the nature of subjective experience. This is because phenomenal properties appear to be properties that make up the experience belonging to the subject – the subject-of-experience. The metaphysical argument underlying Peressini's criticism of IIT is that phenomenal properties are not metaphysically responsible for generating subjectivity. In the previous chapter I claimed that another way to

[125] Peressini, A. (2013), p. 190.

think about Peressini's argument is in relation to what Bayne calls the *unity of consciousness*. Although I agree with Peressini's criticism, I think that the problem for IIT is that while the theory might be correct about integrated information yielding phenomenal consciousness, what it fails to account for is how integrated information yields the *total unified field* of subjective conscious experience. What this means then is that subjective experience, when understood as a holistic and unified experience, is a problem in its own right; one that is arguably more problematic than the problem of understanding the nature of qualia.

In IIT the property of subjectivity is found in the context of integrated conceptual information ('Φ^{MAX}') relating to a Φ-complex. A Φ-complex is described as having its own intrinsic perspective. What this seems to suggest is that subjective experience *emerges* as a consequence of integrated conceptual information at the system level. To overcome Peressini's criticism and the metaphysical argument that qualia are not metaphysically responsible for generating the *unified field* of subjective experience, IIT needs to consider how subjectivity exists within the context of integrated information without holding qualia at all responsible for the existence of subjective experience. I will now turn my attention to the notion of emergence and how IIT might explain the existence of subjective experience within the context of integrated information.

Because IIT is not a form of *pure panpsychism* another way to think about the existence of subjectivity within the context of integrated information is through the metaphysical framework of *emergent panpsychism*. Chalmers defines emergent panpsychism as follows:

> Emergent panpsychism holds that macroexperiences are not grounded in microexperiences, but instead are strongly emergent from microexperiences, from microphysics, or from both. Strong emergence involves the emergence of ontologically novel entities that are not grounded in the base entities. On a common conception of strong emergence, the base entities do not metaphysically necessitate the

> emergent entities, but instead they are connected by contingent laws of nature. On this conception of emergent panpsychism, there will be contingent laws of nature connecting microexperiences (or micro physics) to macroexperience.[126]

I believe that emergent panpsychism is compatible with IIT. I will now attempt to illustrate how I think IIT can be understood as a form of emergent panpsychism. The view I am about to propose differs slightly from how Chalmers defines emergent panpsychism, for I am going to substitute micro and macro "experiences" with the more generic term micro and macro "properties".[127] The reason for this will become clear further on. But as I will explain, micro and macro properties do remain mental properties, they are just not explicit 'experiences' as stated in Chalmers' definition of emergent panpsychism.

One of the clearest conceptual distinctions made in IIT is the distinction between *integrated information* (small-phi 'φ') and *integrated conceptual information* (high-phi 'Φ'). IIT defines these two notions as follows:[128]

Integrated information (φ): Information that is generated by a mechanism above and beyond the information generated by its (minimal) parts. φ measures the integration or irreducibility of mechanisms (integration at the mechanism level).

Integrated conceptual information (Φ): Conceptual information that is generated by a system above and beyond the conceptual information generated by its (minimal) parts. Φ measures the integration or irreducibility of a constellation of concepts (integration at the system level).

---

[126] Chalmers, D. (2013b), p. 15.
[127] I am substituting the word 'experience' with the word 'property' because small-phi do not have any actual experience in and of themselves, for they are not subjective in nature like a high-phi complex. In IIT, only a complex with Φ^{MAX} has subjective *experience* in any meaningful sense of the term. I do not think this substitution for 'properties' in place of 'experience' is philosophically problematic for considering IIT in relation to emergent panpsyhism.
[128] See Glossary in Oizumi, M. *et al* (2014), p. 4 for these and other definitions.

Associated with small-phi (φ) is the notion of a *concept*, defined as: A set of elements within a system and the maximally irreducible cause-effect repertoire it specifies, with its associated value of integrated information φ^{max}. The concept expresses the casual role of a mechanism within a complex.

Associated with high-phi (Φ) is the notion of a *complex*, defined as: A set of elements within a system that generates a local maximum of integrated conceptual information Φ^{MAX}. Only a complex exists as an entity from its own intrinsic perspective.

An important difference between small-phi 'φ' and high-phi 'Φ' is that small-phi are said to exist at the level of individual mechanisms whereas high-phi exists at the system level. In IIT only a complex with Φ^{MAX} is said to have its own intrinsic perspective, which implies that it has subjective experience given that perception is central to the subjective nature of consciousness. In light of how IIT defines small-phi and high-phi, it is clear that small-phi complexes lack the property of intrinsic perspective, which means that small-phi lack the subjective experience that comes with conscious perception. It follows therefore that small-phi 'φ' is non-subjective in nature and thus not a subject-of-experience. A Φ-complex, on the other hand, is subjective in nature due to having its own intrinsic perspective. Thus, the property of subjectivity – that of being a *subject-of-experience* – applies only to a phi-complex with Φ^{MAX}.

Due to the small-phi/high-phi distinction, I hereby hold small-phi 'φ' concepts as micro-properties in the same sense that emergent panpsychism makes the distinction between micro and macro experiences. Because small-phi are non-subjective entities they therefore do not have conscious experiences themselves, so for this reason I think it is wrong to describe small-phi as micro-experiences. Importantly however, the conceptual information generated by a small-phi is phenomenologically involved in formulating the overall constellation of qualia-space as had by a Φ-complex. I propose here that the individual mechanisms (φ), which are responsible for yielding conceptual information, can be viewed as

generating the phenomenality that goes into formulating the overall qualia-space for the whole system. In other words, when conceptual information at the level of mechanisms becomes integrated conceptual information 'Φ^{MAX}' at the system level, the resulting qualia-space is represented in the form of subjective experience *for* the *dominant 'Φ' complex* (*i.e.,* the subject-of-experience). The phenomenal concepts (conceptual information) inherent to small-phi thus qualify as micro-properties given how they are responsible for generating phenomenal mental states, *i.e.* qualia. When these phenomenal concepts are projected into qualia-space at the system level, the overall constellation of qualia-space *represents* the totality of integrated information within the system as a whole, thereby generating subjective experience *for* the dominant complex (Φ-complex). A Φ-complex therefore stands as a macro-property (a property of the system as a *whole*) and only macro-properties (Φ-complexes) are subjective in nature. In short: small-phi 'φ' concepts are micro-properties, for they exist at the level of individual mechanisms and are responsible for generating the conceptual information that formulates qualia-space (Q-space). The high-phi complex (Φ^{MAX}) is a macro-property due to it existing at the system level, where the overall constellation of concepts in qualia-space is represented as subjective experience *for* the system as a whole (Φ-complex).

Using the framework of emergent panpsychism I think the emergence of subjectivity within integrated information can be understood as follows. Macro-properties (high-phi complexes) are not grounded in micro-properties (small-phi concepts), but instead are strongly emergent from small-phi concepts. Strong emergence sees the emergence of ontologically novel entities (high-phi complexes) as not being grounded in the base (micro) entities (small-phi concepts). As strong emergence states, the base entities (in this case, small-phi/qualia) do not metaphysically necessitate emergent entities (Φ-complexes/subjectivity). On this view of strong emergence qualia are not metaphysically responsible for the existence of subjectivity within Φ-complexes: qualia (as concepts) are simply *shapes* within qualia-space that are phenomenally experienced by the dominant complex, which emerges at the system level wherein the integrated conceptual information of qualia-space is

represented in the form of subjective experience – *i.e.* experience *for* the dominant complex.

Admittedly, my conception of IIT as a form of emergent panpsychism is only a rough sketch. However, I believe that my description of IIT as a form of emergent panpsychism is exactly what IIT needs in order to escape Peressini's criticism that IIT conflates subjectivity with qualia, as well as my own concern that IIT fails to explain the *unity of consciousness.* The view I have described here sees subjective experience, *i.e.* the *unification of conscious experience*, emerging at the system level, the level of the dominant Φ-complex. Importantly, the dominant complex is not metaphysically based in the micro-properties (small 'φ'/qualia) but instead emerges when the whole system integrates the totality of conceptual information thereby generating Φ^{MAX} – *i.e.* a complex with intrinsic perspective and subjective experience. While the current formulation of IIT may remain more a theory about qualia, I think emergent panpsychism provides a reasonable theoretical framework for explaining, at least in principle, how subjectivity emerges through integrated information at the system level.

One concern that arise here as a consequence of viewing IIT as a form of emergent panpsychism is this: If small-phi (φ) only generate qualitative properties and are not in and of themselves conscious entities, then not all integrated information states are conscious states. That is, integrated information can remain unconscious but only at the level of small-phi (φ). This concern may dissipate if consciousness is to be understood primarily at the level of *whole* systems (*i.e.* at the level of high-phi 'Φ' only) rather than at the level of individual mechanisms. This is how I believe consciousness should be understood in respect to IIT: that consciousness exists solely for high-phi complexes only and not for small-phi complexes. In my view small-phi are phenomenological mechanisms that are not in their own right conscious states – only high-phi are truly conscious, subjective entities.

4.6 – Conclusion

I began this chapter by highlighting what motivates IIT towards panpsychism, with the two central motivations being: i) integrated information is a fundamental quantity, and ii) consciousness is not an all or none property but is graded – many entities, so long as they possess mechanisms for integrating information are conscious. I also outlined what Tononi himself calls a "phi-centric" view of the universe. This view sees the universe as being populated by what Tononi calls "Φ-stars" which represents the dominant main (Φ) complex of individual human beings and animals here on Earth. But this view also implies that consciousness is not a phenomenon restricted to life on Earth. Rather, consciousness should be found right throughout the universe and perhaps in a myriad of forms.

In section two I compared and contrasted IIT against the three contemporary panpsychist theories of pan-experientialism, pan-phenomenalism and pan-protopsychism. I demonstrated that IIT is not a form of either pan-experientialism or pan-phenomenalism. This is because IIT defines both conscious experience and phenomenality strictly within the context of integrated information, and because integrated information is not everywhere, neither are conscious experiences or phenomenality. Because the information postulate in IIT holds information as being both causal and intrinsic, I proposed that IIT is best understood as a form of pan-protopsychism. This is due to the way in which information can be seen as playing a proto-phenomenal role in the formation of a phi-complex. I therefore claimed that of all the combination problems facing IIT as a panpsychist theory, the first problem confronting IIT is the protophenomenal/phenomenal gap.

To explain how IIT can bridge the protophenomenal/phenomenal gap, I appealed to Adam Barrett's *Field Integrated Information Hypothesis (FIIH)*, in which Barrett explains that truly intrinsic information exists only within fundamental physical entities, namely fundamental fields. To appease the discomfort that many feel towards panpsychism, Barrett distinguishes between *potential* consciousness and *actual* consciousness and claims that actual consciousness should only be assigned to any integrated information state with a

sufficiently rich structure similar to that of a typical healthy adult human conscious moment. Any integrated information state that lacks such a rich conscious moment can be viewed as being potentially conscious. The problem with this distinction, as Barrett notes, is that potential consciousness would still contain phenomenal content, and so he proposes that perhaps the singular term of consciousness should be assigned to even such minimal integrated information states.

I argued that Barrett's distinction between potential consciousness and actual consciousness is not all that helpful in terms of understanding "potential" consciousness. I proposed instead that potential consciousness should be assigned to the fundamental physical entities themselves, specifically fundamental fields. I referred to a number of other fundamental field theory such as Pockett's (2013) electromagnetic field theory to support my view that fundamental fields carry the potential for generating consciousness.

By holding the information intrinsic to fundamental fields as being potentially conscious, I claimed that fundamental fields qualify as being protophenomenal properties. For IIT to resolve the protophenomenal/phenomenal gap, I proposed the following solution: that when the information intrinsic to fundamental fields is integrated in the form of a phi-complex, this protophenomenal information *a priori* entails phenomenal properties, which are shapes in qualia-space (Q-space) as described in a phi-complex. Fundamental fields therefore carry the potential for generating consciousness, but only when fields are integrated in the form of a phi-complex does the intrinsic information *a priori* entail phenomenal properties. By holding intrinsic information as being protophenomenal I believe that my solution resolves the protophenomenal/phenomenal gap for IIT.

Resolving the protophenomenal/phenomenal gap leads to another combination problem for IIT, this being the non-subject/subject gap. In section four I developed IIT into a form of *emergent panpsychism* as a means to resolve this gap. To demonstrate this I modified David Chalmers' definition of emergent panpsychism in order to make the distinction between micro properties (qualia

in this case) and macro properties (subjective-of-experience in this case). On my view of emergent panpsychism, macro properties (Φ-complexes with subjective experience) strongly emerge from micro properties (φ-complexes). Importantly, macro-properties are not grounded in micro-properties, which means that qualia do not metaphysically necessitate subjective consciousness. Instead, subjectivity is an ontologically novel property at emerges at the system level only and not at the base level of individual mechanisms.

While my conception of IIT as a form of emergent panpsychism is only a rough sketch, I believe that emergent panpsychism allows IIT to escape the criticism made by Peressini that IIT is more a theory about qualia and less a theory about subjectivity. I also think that it provides an explanation for the unity of consciousness in the form of subjective experience. Although I think that IIT is best suited to pan-protopsychism more than any other contemporary version of panpsychism, I also believe that IIT can resolve the non-subject/subject gap by appealing to emergent panpsychism as I have developed it in this chapter.

Thesis Conclusion

In this thesis I have presented a philosophical investigation into IIT as a theory of consciousness, and my aim in this thesis has been to demonstrate why I believe panpsychism is a more promising theory of mind than physicalism. To demonstrate this I began in chapter one with an analysis of physicalism as a philosophical doctrine. This involved detailing how the doctrine has been understood traditionally in the forms of *reductive* and *non-reductive physicalism* and in the more contemporary form of *supervenience physicalism.* I also presented what I think are three critical arguments against physicalism, these being Frank Jackson's *knowledge argument,* David Chalmers' *conceivability argument*, and Joseph Levine's *explanatory gap argument*. I argued that when taken together these three arguments raise serious philosophical concerns for physicalism and leave the doctrine in a good deal of doubt. Furthermore, given that some philosophers have attempted to broaden physicalism, *viz.,* Stoljar's o-physicalism and Chalmers' naturalistic dualism, I argued that these broader forms of physicalism runs the risk of collapsing into panpsychism. In light of this, I therefore argued that we have very good reasons for taking panpsychism seriously.

In chapter two I turned my attention to three contemporary versions of panpsychism, namely pan-experientialism, pan-phenomenalism and pan-protopsychism. All three versions run up against different forms of the combination problem. And while these three versions have their respective merits, I argued that each version faces certain internal issues that prevent it from successfully resolving its respective combination problems. Further philosophical work on each version respectively could lead to stronger solutions in relation to closing their respective metaphysical gaps. I am not confident that pan-experientialism can close the micro-subject/macro-subject gap, mainly because I think that both Goff and Coleman present strong arguments for why subjects cannot combine. I am more confident about Coleman's pan-phenomenalism, but I think a stronger conception of the central/perceptual subjective domain is needed to better explain the association between

phenomenal representation and subjective experience. And since pan-protopsychism has yet to be developed into a formal theory there are no solutions to its combination problems. However, my analysis of the panpsychist implications in IIT led me to develop the theory into a form of pan-protopsychism, so I have developed two solutions for it in this respect.

In chapter three I turned my attention directly to IIT as a recent theory of consciousness. I firstly outlined a basic framework detailing the phenomenological axioms and physical postulates, along with how IIT distinguishes between individual mechanisms with small-phi 'φ' and a whole system of mechanisms with high-phi 'Φ'. With this basic framework in hand I then examined IIT as a physicalist theory and how it stands up to the arguments against physicalism as featured in chapter one. Regarding the knowledge argument, I looked at how Tononi claims that consciousness is a matter of *being* more than a matter of *describing*. That is, to *know* anything about consciousness one must firstly *be* conscious before one can *describe* it. According to Tononi, the reason why Mary knows nothing about what it is like to see 'red' (and all other colour experiences) is simply due to her never having *been* phenomenally conscious of such experiences. I argued, however, that even with IIT in mind, it still seems that Mary cannot know all that much about phenomenal consciousness. That is, Mary cannot confidently deduce any phenomenological facts simply through knowing IIT. I concluded by arguing that it is unclear and certainly questionable as to whether IIT can capture the *quality* of consciousness as it so claims.

Regarding the conceivability argument and philosophical zombies, I highlighted the fact that IIT zombies are structurally very different from philosophical zombies. This means that IIT zombies are irrelevant to the conceivability argument against physicalism. However, because IIT claims that consciousness *is* integrated information, I argued that it is still conceivable for a system to generate integrated information and still be phenomenally unconscious. This is because the act of integration is a functional notion. Hence it is conceivable that all the mechanisms in a system are functioning appropriately and generating

integrated information even though no phenomenal experience is being had by the system. I argued therefore that IIT is susceptible to the possibility of integrated information being phenomenally unconscious. To defend IIT against the conceivability argument I considered how IIT could adopt Chalmers' *dual aspect theory* of information, on which information has both a physical aspect and a phenomenal aspect. I think that dual aspect theory could provide IIT with a metaphsycial basis for explaining the logical connection between the physical properties of complex systems and the manifestation of phenomenal consciousness. In this sense, when information is integrated within physical systems such as the brain, integrated information can be said to yield phenomenal consciousness by virtue of information having this phenomenal aspect. I think that this is one strategy available to IIT for dealing with the possibility of philosophical zombies.

I also examined IIT as a realisation theory of mind by focusing on what it presents as being a "minimally conscious" photodiode. In this section I questioned whether this photodiode could be described as having an actual conscious experience, especially when this minimal experience is described as being "meaningless" to the photodiode. I argued that conscious experience is typically *meaningful*, but since the photodiode is seen as having a meaningless experience I am not convinced that the photodiode is having an actual conscious experience as such. I therefore argued that the photodiode is not minimally conscious but rather "minimally phenomenal", in which case the photodiode could be described as generating qualia while not being actually conscious in and of itself. Through analysing IIT as a physicalist theory I concluded chapter three by claiming that the theory cannot be classified as a form of physicalism.

Because IIT is stated as having panpsychist implications my aim in chapter four was to philosophically investigate how IIT is best understood as a form of panpsychism. I demonstrated that IIT does not conform to either pan-experientialism or pan-phenomenalism, but that it does qualify as a form of pan-protopsychism. As a version of pan-protopsychism IIT confronts two combination problems: i) the protophenomenal/phenomenal gap, and ii) the

non-subject/subject gap. To resolve the first gap I appealed to Adam Barrett's FIIH in which fundamental fields are described as possessing intrinsic information. I claimed that fundamental fields qualify as protophenomenal properties. By holding intrinsic information as protophenomenal, when fields are integrated into a phi-complex, the act of integrating intrinsic information *a priori* entails phenomenal properties with the *qualia-space* of a phi-complex.

To resolve the non-subject/subject gap I developed IIT into a form of *emergent panpsychism*. Because the property of subjectivity applies only to a high-phi 'Φ' complex and not to small-phi 'φ' complexes, I demonstrated through the framework of emergent panpsychism that subjectivity emerges only at the level of the *whole* system 'Φ' but not at the level of individual mechanisms 'φ'. Thus, subjectivity is a macro property of high-phi complexes that emerges over the micro properties (small-phi complexes in this case). Subjectivity is therefore an ontologically novel property that is distinct from the micro properties from which it emerges. I believe that my view of IIT as a form of emergent panpsychism presents a solution for IIT to bridge the non-subject/subject gap.

In my view, the most important philosophical aspect of IIT is the notion of *information*. Everyone is familiar with information in some form or another, and modern science happily talks about information in the world: computers process information, the brain integrates information, and apparently quantum particles transmit information via entanglement. But what *is* information exactly? Is information categorically physical? And might information states be the categorical bases of all physical and indeed mental states? The problem is that the ontological status of information is by no means clear, nor is there any consensus within philosophy or modern science as to what information *is* exactly. Further philosophical consideration is clearly required on this matter. But despite seeming non-physical in nature, my current intuition is that information – including integrated information – is something fundamentally physical. The reason why I think this is based mainly on the way in which the information postulate in IIT views information as being both *causal* and *intrinsic*.

If information is causally efficacious in the physical world, then information is arguably something physical too.

In its current form IIT (3.0) is an incomplete theory of consciousness. Only time will tell whether it can be developed more confidently and in accordance with future empirical evidence regarding brain activity and indeed the world more broadly. And it remains to be seen whether IIT can confirm some of the predictions it makes about the actual nature of conscious experience. At the same time, however, IIT is a cutting-edge theory of consciousness. If IIT is heading in the right direction then we need to pay close attention to the philosophical implications of the theory, especially those that imply panpsychism. Although Tononi himself endorses these implications, he does not clearly explain how IIT is to be understood as a panpsychist theory. Nor does he explain how the theory resolves the combination problems for panpsychism. IIT therefore requires much needed philosophical attention in all of these respects. My aim and purpose in this thesis has been to analyse IIT as a theory of consciousness: my conclusion is that IIT is best understood as a form of pan-protopsychism and not a form of physicalism. And if IIT is correct about consciousness being a fundamental and intrinsic feature of the world, which I believe it is, then this makes panpsychism a promising theory of mind.

Bibliography

Barrett, A. B. (2014) *An integration of integrated information theory with fundamental physics*. Front. Psychol. **5**:63. doi:10.3389/fpsyg.2014.00063

Bayne, T. (2010) *The Unity of Consciousness: A cartography* in M. de Caro & F. Ferretti (eds.) 2007 *Cartographies of the Mind: Philosophy and Psychology in Intersection*. Dordrecht: Kluwer, 201-10.

Beaton, M. and Aleksander, I. (2012). World related integrated information: enactivist and phenomenal perspectives. *Internation Journal of Machine Consciousness* 4, 439-455. doi: 10.1142/S1793843012400252

Campbell, K. (1970). *Body and mind*. Garden City, N.Y., Anchor Books.

Coleman, S. (2012) *Mental Chemistry: Combination for Panpsychists*, dialectica Vol.66, No.1.

Coleman, S. (2013) *The Real Combination Problem: Panpsychism, Micro-subjects, and Emergence*, Erkenntis (1):1-26.

Chalmers, D. J. (1995) *Facing up to the hard problem of consciousness*, Journal of Consciousness Studies, 2(3):200-19

Chalmers, D. J. (1996). *The conscious mind*, *In search of a fundamental theory*. New York, Oxford University Press.

Chalmers, D. (2013a) *Panpsychism and Panprotophyscism*, forthcoming in Amherst, http://fragments.consc.net/.

Chalmers, D. (2013b) *The combination problem for panpsychism*, in (G. Bruntrup and L. Jaskolla, eds.) *Panpsychism*, Oxford University Press.

Cerullo, M.A. (2011) *Integrated Information Theory*, *a promising but ultimately incomplete theory of consciousness*, Journal of Consciousness Studies, 18, No. 11-12, p. 45-58

Dennett, D. (1995). The unimagined preposterousness of zombies. *Journal of Consciousness Studies* 2 (4):322-26.

Gamez, D. (2011). Information and Consciousness. *Etica Pol.* 13, 215-234. Available online at: http://hdl.handle.net/10077/5803

Goff, P. (2009) *Why Panpsychism Doesn't Help Us Explain Consciousness*, dialectica Vol. 63, No.3.

James, W. (1890/1950) *The Principles of Psychology*, Vols. 1 & 2, New York, Dover Publications.

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

Kim, J. (1993). *Supervenience and mind: selected philosophical essays*. New York, NY, USA, Cambridge University Press.

Krikpe, S. (1980). *Naming and necessity*. Cambridge, Mass, Harvard University Press.

Libet, B. (1994). A testable field theory of mind-brain interaction. *Journal of Consciousness Studies* 1, p. 119-126.

Levine, Joseph (1983). "Materialism and qualia: The explanatory gap." *Pacific Philosophical Quarterly* 64:354-61.

Lewis, D. (1966). 'An Argument for the Identity Theory', *Journal of Philosophy* 63.

Ludlow, P., Nagasawa, Y., & Stoljar, D. (2004). *There's something about Mary: essays on phenomenal consciousness and Frank Jackson's knowledge argument*. Cambridge, Mass, MIT Press.

McFadden, J. (2002). The conscious electromagnetic information (cemi) field theory: the hard problem made easy? *Consciousness Studies* 9, p. 45-60.

Oizumi M, Albantakis L, Tononi G, (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Comput Biol 10(5): e1003588. Doi:10. 1371/journal.pcbi.1003588.

Pockett, S. (2013). Field theories of consciousness. *Scholarpedia* 8:4591. Doi: 10:4249/scholarpedia. 4951

Peressini, A. (2013) *Consciousness as Integrated Information, a provisional philosophical critique*, Journal of Consciousness Studies, 20, No. 1-2, pp. 180-206

Ryle, Gilbert (1949). *The Concept of Mind*. Hutchinson and Co.

Searle, J. (1992). *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* 27:379-423 [Reprinted in C. E. Shannon and W. Weaver, *The Mathematical Thoery of Communication.* Urbana: University of Illinois Press, 1949]

Skrbina, D. (2005) *Panpsychism in the West*, Cambridge, MIT Press.

Skrbina, D. (ed.) (2009). *Mind That Abides: Panpsychism in the New Millennium*. John Benjamins Publishing.

Smart, J.J.C. (1956). 'Sensations and Brain Processes'. *The philosophical Review*, Volume 68, Issue 2, 141-156.

Stoljar, D. (2001). Two Conceptions of the Physical. *Philosophy and Phenomenological Research.* 62, 253-281.

Strawson, G. (1994) *Mental Reality*, Cambridge, MA: MIT Press.

Strawson, G. (2006) *Panpsychism? Reply to commentators with a celebration to Descartes*. Journal of Consciousness Studies 13 (10-11).

Stawson, G. (1999) *The Self and the Sesmet*, Journal of Consciousness Studies, 6, No.4, pp.99-135.

Strawson, G. (2015) "Mind and Being: the primacy of panpsychism", in *Panpsychism: Philosophical Essays*, ed. G. Bruntrup and L. Jaskolla, Oxford University Press.

Tegmark, M. (2007) The Mathematical Universe. *Found. Phys.* Available online at: **arXiv:0704.0646**

Tononi, G. (2008) "Consciousness as Integrated Information: a provisional manifesto." *Biol. Bull.* 215: 216-242.

Tononi, G. (2015), "Integrated Information Thoery". Scholarpedia, 10(1):4164.

Tononi, G. and C. Koch (2014) "Consciousness: Here, There, but not Everywhere." arXiv:1405.7089.

Yablo, S. (1993). Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53 (1):1-42.