2. CONSCIOUSNESS

2.1 Introduction

This chapter outlines a theory of consciousness that will be used throughout this thesis. A general failure to analyse what we mean by the physical world, perception and consciousness has been a central source of confusion in consciousness research and the first part of this chapter spends a substantial amount of time clarifying basic concepts about the phenomenal and the physical and linking them to the sources of our knowledge about consciousness. The philosophical approach that is used for this work is influenced by neurophenomenology (Varela 1996, Thompson et al. 2005), which combines cognitive science and neuroscience with a systematic analysis of human experience influenced by Continental philosophy – for example, the work of Husserl (1960). Although this approach might occasionally sound naïve, it is a necessary first step if we want to get clearer about what can and cannot be scientifically established about consciousness. Some of this material is also covered in Gamez (2007c, pp. 25-87) and it maps onto Metzinger's (2000) distinction between phenomenal and theoretical knowledge.

The first section in this chapter is a phenomenological examination of the relationship between the phenomenal and the physical, which is used to develop a definition of consciousness in Section 2.3. This is compared with some of the previous definitions that have been put forward and Section 2.4 examines and rejects popular metaphysical theories about consciousness, such as dualism, epiphenomenalism and physicalism, in favour of a correlates-based approach, which is explored in Section 2.5. A close reading of the brain-chip replacement experiment is used to show that we will never be able to separate out some of the potential correlates of consciousness empirically, which leads to a distinction between type I and type II

correlates of consciousness. Section 2.6 then covers the three type II theories of consciousness that have been selected to design and analyze a neural network in this thesis. The final part of this chapter develops a preliminary interpretation of the relationship between consciousness and action.

2.2 The Phenomenal and the Physical

> A person who grew up and lives in a certain limited environment has time and again encountered bodies of fairly constant size and shape, colour, taste, gravity and so on. Under the influence of his environment and the power of association he has become accustomed to find the same sensations combined in one place and moment. Through habit and instinct, he presupposes this constant conjunction which becomes an important condition of his biological welfare. The constant conjunctions crowded into one place and time that must have served for the idea of absolute constancy or substance are not the only ones. An impelled body begins to move, impels another and starts it moving; the contents of an inclined vessel flow out of it; a released stone falls; salt dissolves in water; a burning body sets another alight, heats metal until it glows and melts, and so on. Here too we meet constant conjunctions, except that there is more scope for spatio-temporal variation.
>
> Mach (1976, p. 203)

2.2.1 The Stream of Experience

Our theoretical studies and scientific experiments take place in a colourful moving noisy spatially and temporally extended stream of experience. This stream of experience is the most real thing that there is: everything that we do is carried out within it.[1]

Within waking life this stream of experience is highly structured. Some of the most characteristic structures are stable objects, which typically have a reasonably consistent set of properties that can be experienced on multiple occasions. For example, when I am examining a machine, I experience the front, turn it around to look at the back, and when I turn it around so that the front faces me again, I seem to experience the same set of sensations from the machine

[1] See Dennett (1992) and Blackmore (2002) for a criticism of this notion of the stream of experience.

as when I first looked at it. This stability of objects also extends over time: I speak about a *single* machine rusting because I can allow a subset of the machine's properties to change without thinking that a completely different machine has appeared in front of me. Whilst objects in waking life typically exhibit this kind of stability, objects in dreams or hallucinatory states are much less stable, and it is harder to return to the same view of an object or to perceive changes in a single object over time.

The stability of objects leads us to speak about their *persistence* when they are not under direct observation. Although I am not currently experiencing my motorbike, it is still out there in the garage and I can experience it again by going into the garage and taking off its cover. The difference between objects that we are currently perceiving and objects that are not currently being perceived by anyone is described by Lehar (2003) using his metaphor of a 'bubble' of perception that we 'carry around' with us, within which only a subset of the world's objects appear. Although objects appear *as* three-dimensional within this bubble of perception, I only experience part of them at any one time. From one position, I experience the outside of a cardboard box, but not the whole box, and I have to move relative to the box to experience more of its properties. Instead of simply saying that the box is there, I talk about *seeing* the box to indicate that I am *currently* experiencing the box, that the box is *within* my bubble of perception.

This interpretation of perception can be further analysed and broken down. For example, my visual perception is strongly linked to my eyes. In the stable world of waking life, the set of objects within my bubble of visual perception can be altered by covering my eyes or by damaging them in some way. The same is true of my ears and my bubble of auditory perception and my body and my bubble of somatic perception. In general, altering the sensory parts of my body alters the contents of my bubble of perception; it changes the subset that is 'extracted' from the totality of possible perceptions. This is a purely empirical observation and in a different world it could turn out that covering my big toe reduced the set of objects within my bubble of

visual perception. However, in this world, repeated experiments have shown that it is the eyes that are important for this. An alternative interpretation would be that it is the world that is changing when I cover my eyes, and not my bubble of perception. However, when I turn my head I continue to see the same objects with my other eye, and so I attribute the change to my perception and not to the world itself.

The states of my bubble of perception are also strongly correlated with the state of my brain. When I hit my head, the waking world is overlaid with bright points of light, damaging parts of my brain reduces my bubble of perception in different ways, and my bubble of perception can be altered by injecting or ingesting chemicals that are circulated by my blood to my brain.[2] These can change the colours, sounds and sensations in my bubble of perception, and they can even destroy the stability of my waking experiences entirely and make them similar to a dream. This correlation between perceptual changes and the brain is not logically necessary in any way – for example, it might have turned out that hitting a ring on my finger produced bright points of light. However, in this world, the strong correlations between my bubble of perception and the states of my senses and brain suggest that without my senses and brain I would not have a bubble of perception at all.[3]

As I move around I come across other objects that look the same as me and have a similar brain and body. These objects behave in a similar way to myself and speak about other objects in a similar way. The verbal reports of these human objects suggest that for most of the time they perceive different parts of the world that is experienced by me. When the senses or brains of these other people are damaged or altered by chemicals, their verbal reports change in the same way that mine changed under similar circumstances. These changes have no effect on the objects within my own bubble of perception, which gives me further evidence for my belief

[2] Chemicals that do not reach my brain do not have any effect.

[3] The possession of senses and a brain might be necessary for a bubble of perception, but they are not sufficient because some states of my senses and brain, such as deep sleep, are not associated with perception at all.

that changes to my brain do not induce changes in other objects. Some people's bubbles of perception contain objects or properties of objects that are not perceived by anyone else. Under these circumstances it becomes a matter of debate and consensus about which objects and properties are artefacts of people's bubbles of perception.[4]

2.2.2 The Physical World

The stream of experience is structured in subtle ways that can only be identified through systematic investigations. These regularities are often explained by hypothesizing invisible *physical* entities that have effects on the stream of experience. As systematic measurements confirm the regularities, the physical theories gain acceptance and their hypothesized entities are believed to be part of the world, even though they do not directly appear within the stream of experience. To make this point clearer I will give a couple of examples.

A classic example of a physical theory is the atomic interpretation of matter, which claims that large scale changes in the stream of experience are caused by interactions between tiny bodies. By hypothesising that gases consist of a large number of moving molecules, Bernoulli (1738) developed the kinetic theory of gases, which describes how pressure is caused by the impact of molecules on the sides of a container and links heat to the kinetic energy of the molecules. Although molecules had not been observed when the theory was put forward, their existence became accepted over time because of the theory's good predictions. More recently we have developed ways of visualising individual molecules, atoms and particles – for example, the scanning tunnelling microscope and bubble chamber. These techniques use a more or less elaborate apparatus to construct representations within the stream of experience that are interpreted as the effects of these particles.

[4] Children, mystics and madmen all experience non-consensual objects within their bubbles of perception. See Gamez (2007c, pp. 145-193) for a detailed discussion.

A second example of a physical theory is Newton's interpretation of gravity. To make more accurate predictions about the movement of objects relative to the Earth, Newton hypothesized an invisible force that attracts remote bodies. The magnitude of this gravitational force is given by Newton's equations, which can be used to calculate the acceleration of objects towards the Earth and to make reasonably accurate predictions about the movement of planetary bodies. Newton's theory of gravity was very controversial when it was put forward and Newton himself had no idea how one body could exert a force on another over a distance: "I have not been able to discover the cause of those properties from the phenomena, and I frame no hypotheses" (quoted from Gjertsen (1986, p. 240)). Over time Newton's theory gained acceptance because of the accuracy of its predictions and people gradually came to believe that the physical world was permeated by an invisible gravitational force. More recently, general relativity's claims about the effect of matter on the curvature of four-dimensional spacetime are no easier to imagine, and these counterintuitive claims are only taken seriously because of their accurate predictions.[5]

Almost every aspect of the stream of experience has been re-interpreted by modern science as forces, particles or waves that affect the stream of experience when they are within a certain frequency range (sound and light), of a certain chemical composition (smell and taste) or when they collide with the human body (touch). These appearances do not *resemble* the original forces, particles or waves in any way – light does not look like a photon; sound does not sound like a wave. Our scientific models of physical reality enable accurate predictions to be made about the transformations of objects in the stream of experience, but the forces, particles and waves that constitute these models are defined mathematically and have to be indirectly measured from within the stream of experience using scientific apparatus.

[5] Newton also introduced a notion of mass that is different from what we experience as weight in the stream of experience. If a pre-Newtonian person could have travelled to different planets, then they would have probably said that they were gaining and losing weight, rather than preserving a constant mass that was attracted by different gravitational forces.

2.2.3 The Phenomenal World

> The representation of space in the brain does not always use space-in-the-brain to represent space, and the representation of time in the brain does not always use time-in-the-brain.
>
> Dennett (1992, p. 131)

When we first encountered the stream of experience, it was neither objective nor subjective: it was just what was there as the world. However, the development of the notion of a non-experiential physical world forces us to re-interpret this stream of experience as a *phenomenal* world that is different from the physical world. This phenomenal world is the same stream of experience that we started with, but reinterpreted as a *representation* of the non-sensory physical world.

Many people try to limit the phenomenal world to simple sense experiences, such as red, the smell of burnt plastic, and so on, and make the assumption that we directly perceive the spatial and temporal aspects of the physical world.[6] The problem with this position is that there are no scientific or philosophical arguments for *resemblance* between our experiences of space, time and movement and these qualities in the physical world. In fact just the opposite is suggested by interpretations of perception put forward by Metzinger (2003), Lehar (2003), Gamez (2007c), Dawkins(1998), Revonsuo (1995) and many others, who claim that the brain generates a simulation of the physical world, in which space, time and colour are *all* representations within a completely virtual environment.[7] Although our virtual representations might have analogues in the physical world, there is no reason to believe that they resemble the

[6] This old assumption goes back to Locke (1997), who distinguished between the primary qualities of figure, solidity, extension, motion-or-rest and number, which are something like direct perceptions of qualities of the physical world, and secondary qualities, such as colour or smell, which are artefacts produced by the effect of the primary qualities on the senses.

[7] This is also supported by Russell's (1927) claim that physical matter is a source of events and not something that we are directly acquainted with. Kant's (1996) *Critique of Pure Reason* is another version of this position.

physical world, which has a completely non-sensory nature.[8] This suggests that phenomenal experiences cannot be reduced to simple sensory qualia that are superimposed on a direct experience of physical reality. If the phenomenal world is interpreted using a theory of qualia (a highly debatable point – see Section 2.3.1), then *everything* is qualia, including experiential space, time, movement and size. Since there is no such thing as a physical *experience*, the phenomenal world is everything in the stream of experience, and the physical theories of particles, gravity, and so on, lead us to reinterpret this stream of experience in relation to an invisible physical world.[9]

2.2.4 The Physical and Phenomenal Brain

Within the picture that I have presented so far, regularities in the stream of experience are explained using scientific theories based on the physical world, and we would expect that scientific theories about consciousness would conform to this model and be based on the physical brain, and not on the brain as it appears in the stream of experience. Before these scientific explanations can be sought it is essential to get as clear as possible about the distinction between the physical and phenomenal brain, which will help with the discussion of the hard problem of consciousness in Section 2.4.5.[10]

---

[8] This does not amount to scepticism about the physical world because space in the brain is represented by our phenomenal image of space. It is just that we cannot imagine or picture to ourselves what real space is actually like. This is also different from instrumentalism and anti-realism because one can be completely realistic about scientific descriptions of forces, quarks, electrons, and so on, and yet claim that they can only be described in an abstract language, and not imagined by human beings using the virtual phenomenal model associated with the brain.

[9] A more detailed version of this argument can be found in Gamez (2007c, pp. 71-83).

[10] This focus on the brain is not affected by Clark and Chalmers' (1998) suggestion that many cognitive processes might be carried out in the environment. Whilst some of our cognitive processes and even beliefs may be external to our brains, Clark and Chalmers (1998) are careful to point out that both experiences and consciousness are likely to be determined by the processes inside our brains. Velmans' (1990) interpretation of projection theory is also consistent with a strong link between the brain and consciousness because he claims that consciousness is generated inside the brain and projected out of it into the environment. The only people I am aware of who question a strong link between the brain and consciousness are Thompson and Varela (2001), who criticize an exclusive focus on the *neural* correlates of consciousness and claim that "the processes crucial for consciousness cut across brain–body–world divisions, rather than being brain-bound neural events." (Thompson and Varela 2001, p. 418).

The physical brain is part of physical reality: it is completely non-phenomenal and has never directly appeared in the stream of experience. It consists of the physical entities that are deemed by physicists to constitute physical reality, such as quarks, wave-particles, forces, ten-dimensional superstrings and so on. The physical brain is also defined by other properties, such as spatial extension, mass and velocity, which can be defined mathematically and must be carefully distinguished from their phenomenal representations.

The phenomenal brain is the totality of our possible and actual phenomenal experiences of the brain, including its texture, colour, smell, shape, taste, sound and so on. The phenomenal brain also includes phenomenal measurements of the physical brain, such as the experience of looking at an fMRI scan, or taking a reading from a thermometer with its bulb inside the brain. We can remember our phenomenal experiences of the brain and imagine them when the brain is not physically present.

2.2.5 Concluding Remarks about the Phenomenal and the Physical

This interpretation of the phenomenal and physical gives equal importance to the phenomenal and physical worlds and suggests that it is too early to *assume* that the phenomenal world can be reduced to the physical world - although it is not impossible that this could be established by later work. This understanding of the phenomenal and the physical also fits in with Varela's (1996, p. 347) claim that: "lived, first-hand experience is a proper *field of phenomena*, irreducible to anything else" and it has a lot in common with Flanagan's (1992) constructive naturalism and Searle's (1992) defence of the irreducibility of consciousness. How this starting point could be developed into a science of consciousness is discussed in detail in the rest of this chapter.

A second aspect of the phenomenal and the physical that is worth touching on at this stage is the ontological status of abstract properties, such as the volume of the brain or the

number of red objects in my visual field. Whilst the volume of the brain is not a physical entity like a force or particle, it is also not part of my stream of experience in the same way as a yellow flower or the smell of myrrh. This problem extends to the ontological status of language and mathematics, which are also not straightforwardly phenomenal or physical entities. Since this question is not particularly relevant to this thesis, it will be set aside here and I will use abstract properties, mathematics and language to describe the phenomenal and the physical worlds without taking a position about their ontological status.

2.3 What is Consciousness?

The distinction between the phenomenal and the physical will now be used to set out a definition of consciousness that will be employed throughout this thesis. After some clarifications of this definition, it will be compared with some of the other interpretations of consciousness that have been put forward.

2.3.1 Definition of Consciousness

The distinction between an invisible physical world and a phenomenal stream of experience suggests a simple definition of consciousness:

Consciousness is the presence of a phenomenal world. (2.1)

This definition is based on the distinction between phenomenal and physical reality and it suggests that phenomenal states and consciousness can be treated as interchangeable terms. Some clarifications of this definition now follow.

What is the best way speak about the consciousness of X?

There are many different ways of speaking about the consciousness that is associated with an object or person X and since some of these are potentially misleading, I will endeavour to adhere to the following general rules throughout this thesis:

- Unspecific terms, such as "the red flower", "the system", "the network", etc., could refer either to the phenomenal aspect of X, which I experience with my human senses, or to its underlying physical reality. Most of the time it does not matter whether the physical or the phenomenal aspect of X is being referred to, since it is assumed that phenomenal X corresponds to an underlying physical X, and that parts of physical X can affect our stream of experience.[11]

- Some conscious states might not include a subject or a perspective, and so it is potentially misleading to claim that X is *in* a phenomenal world. Difficult problems with spatial perception also make the use of 'in' problematic - see Gamez (2007c, pp. 25-87) for a discussion.

- The approach to consciousness in this thesis is based around the identification of correlations between the phenomenal and physical worlds (see Section 2.5), which may eventually lead to a causal theory of consciousness. However, until this point is reached it is inappropriate to use phrases like "The consciousness of X is *caused* by brain state Y" or "The brain state Y *gives rise to* the consciousness of X."

- I will be using the word "associated" to express the link between conscious states and X. The person or object X in front of me is an object in my phenomenal world and I can measure the physical aspects of this object. If X makes plausible claims about its

[11] It seems likely that all systems have both phenomenal and physical aspects, but I am leaving this open at this stage. Although it might be thought that some systems could have a completely non-phenomenal character – a dark matter machine for example, or perhaps a highly dispersed gas – it would still be possible to construct phenomenal representations of these systems, such as a picture.

conscious states or if I make predictions about the conscious states of X, then I will express this by saying “there are conscious states *associated* with X” or “there are phenomenal states *associated* with X.”

- Once we have an association between phenomenal states and a phenomenal/ physical X, then we can start to look for correlations between them. The specification of a correlation between a conscious state and a state of X is more technical than an association, and I will use “the consciousness *correlated* with X” to refer to a mathematical or statistical relationship between the consciousness associated with X and phenomenal/ physical X.
- Although “The conscious states *connected* with X” might seem to be a plausible alternative to “associated”, it implies a causal relation in one or both directions, which assumes too much at this stage.
- “The consciousness *of* X”, “conscious X” or “X’s consciousness” will be used as convenient synonyms for “the consciousness associated with X.”
- “What X is conscious of” will be used as a synonym for “The contents of the consciousness associated with X.”

The only deliberate exception to these rules will be when I am explaining or paraphrasing the work of other people.

Definition 2.1 has nothing to do with language

Most of my conscious states have little to do with language or narrative, although I use language to reflect on them and communicate them to other people. It might turn out that consciousness is constantly correlated with language or self-reflexivity, but this is not something that needs to be incorporated into the most basic definition of the phenomena that we are attempting to study and explain.

Phenomenal worlds might be completely different

When I experience a person within my phenomenal world they are surrounded by objects that are part of my phenomenal experience. However, the objects that I perceive might not be included in the other person's world – they could be immersed in a uniform field of blackness or pain, for example. When we look at a schizophrenic patient, such as Schreber, we say that he is associated with a phenomenal world, but this world might be very different from our own.[12]

There is nothing special about qualia

In Section 2.2.3 I argued that there is no fundamental distinction between classic qualia, such as red, and our experience of space, time, movement and number. This suggests that the concept of qualia is either redundant or should be used as a synonym for phenomenal experience in general. Theories of consciousness apply to the whole phenomenal world, and not just to the colourful smelly parts of it. Critical discussions of qualia and their standard interpretation can be found in Dennett (1988, 1992) and Churchland (1989).

The concept of consciousness is a new and modern phenomenon

This definition of consciousness helps us to understand why the concept of consciousness is a relatively new phenomenon. In the discussion of the phenomenal and physical I showed how the modern concept of the phenomenal is strongly linked to the physical world described by science, which is a recent product of a great deal of conceptual, technological and experimental effort. Earlier societies lacked this notion of physical reality, and so it is not surprising that the concept of consciousness is absent from Ancient Greek, Chinese and in the English language prior to the 17th Century (Wilkes, 1984, 1988, 1995). Consciousness is a new and modern problem because science is a new and modern phenomenon. The stream of experience was once understood in

[12] See Schreber (1988) for a description of this world and Nagel (1974) for a more detailed discussion of this point.

relation to an invisible world of gods and spirits; now it is interpreted as a *conscious* phenomenal representation of quarks, atoms, superstrings and forces.[13]

A single concept of consciousness

Many people, such as Armstrong (1981) and Block (1995), have tried to distinguish several different notions of consciousness, whereas Definition 2.1 is based on a single type of consciousness that is present when there is a phenomenal world and absent when there is not. States that are claimed to be conscious according to Armstrong's minimal consciousness or Block's access consciousness, for example, are not conscious according to Definition 2.1.

Awareness

It is worth distinguishing the presence of a phenomenal world from the related concept of awareness. Although many people link consciousness and awareness,[14] it is possible to interpret awareness as the presence of active representations in the brain that are not necessarily conscious. For example, when I am cycling along a canal and imagining a recent concert, then I might be said to have sensory awareness of the canal, although I am not conscious of it. Likewise, I might be attributed awareness of the sound of the refrigerator in my kitchen, but I only become conscious of it when the compressor cuts out. To avoid ambiguities of this kind, I will not use awareness in any technical sense in this thesis.

Consciousness and wakefulness

According to Laureys et. al. (2002, 2004) many patients in a vegetative state can be awake without being conscious and display a variety of responses to their environment:

---

[13] Many people around today have a different interpretation of the stream of experience that is often closely aligned with idealism (see Section 2.4.1) and rejects the scientific interpretation of physical reality – Tibetan Buddhism is one example. There is not space in this thesis to cover these other theories in detail and the primary focus will be on the scientific study of consciousness, which is closely linked to the Western atheistic viewpoint.

[14] For example, the *Oxford English Dictionary*'s (1989) third definition of conscious is: "The state or fact of being mentally conscious or aware *of* anything." (Volume III, p. 756).

> Patients in a vegetative state usually show reflex or spontaneous eye opening and breathing. At times they seem to be awake with their eyes open, sometimes showing spontaneous roving eye movements and occasionally moving trunk or limbs in meaningless ways. At other times they may keep their eyes shut and appear to be asleep. They may be aroused by painful or prominent stimuli opening their eyes if they are closed, increasing their respiratory rate, heart rate and blood pressure and occasionally grimacing or moving. Pupillary, corneal, oculocephalic and gag reflexes are often preserved. Vegetative patients can make a range of spontaneous movements including chewing, teeth-grinding and swallowing. More distressingly, they can even show rage, cry, grunt, moan, scream or smile reactions spontaneously or to non-verbal sounds. Their head and eyes sometimes, inconsistently, turn fleetingly towards new sounds or sights.
>
> Laureys et al. (2002, p. 178)

Vegetative patients are *awake* when they have their eyes open and vocalise or grimace. These patients are *conscious* when they are experiencing a phenomenal world, and Laureys et al. (2004) suggest some of the clinical signs that can be used to judge when this is the case.

2.3.2 Comparison with Other Theories of Consciousness

This section compares Definition 2.1 with some of the more influential theories of consciousness.

What it is like

According to Nagel (1974) an organism is conscious if there is something that it is like to *be* that organism. However, it is possible (although unlikely) that there are phenomenal worlds without any stable correlation with phenomenal or physical things, and so defining consciousness in terms of this association with phenomenal and physical objects is adding too much to the concept at this stage. Furthermore, Nagel's claims about the *subjective* character of experience suggests a necessary connection between consciousness and a perspectival self. Whilst some kind of self is undoubtedly important for higher organisms, it might not be an essential feature of

consciousness and there might be forms of minimal consciousness that are without subjectivity – see, for example, Metzinger's minimal notion of consciousness in Section 2.6.4.

Nagel (1974) discusses how we are unable to describe the experiences of creatures that are very different from ourselves – for example when we attempt to describe the phenomenology of a bat. This problem also occurs when we attempt to describe the consciousness of artificial systems, and it is covered in more detail in Section 4.4.2. Nagel's resistance to various reductionist theories of consciousness is also very much in line with the approach to consciousness that is taken in this thesis.

Minimal, perceptual and introspective consciousness.

Armstrong (1981) distinguishes between three types of consciousness. The first, called minimal consciousness, is present when there is mental activity occurring in the mind. When we are in deep sleep we might have knowledge and beliefs, but there are no events or occurrences going on, and so we are not minimally conscious. However, a person solving a problem in his or her sleep is minimally conscious because thinking is a form of mental activity. Armstrong's second type of consciousness is perceptual consciousness, in which we are aware of what is going on in our body and environment. Dreaming is minimally conscious, but we only become perceptually conscious when we wake up and perceive the world. Finally Armstrong identifies a third type of consciousness, called introspective consciousness, in which we have perception-like awareness of the states and activities of our mind. This notion of introspective consciousness was invoked to handle cases like 'unconscious' driving, in which we are perceptually conscious of the road, but not fully conscious of it because we are thinking about other things.

An initial difficulty with Armstrong's first two types of 'consciousness' is that it makes little sense to call something conscious that takes place whilst we are in deep sleep or 'unconsciously' driving, and so I will set Armstrong's notions of minimal and perceptual consciousness aside in this thesis. A central problem with Armstrong's third notion of

introspective consciousness is that it seems perfectly coherent that we could be aware of our own mental states without any form of consciousness being present, and such meta awareness is likely to be taking place all the time in the brain. For example, when we are driving 'unconsciously' and thinking about other things, low level sensory data is being passed to the parts of the brain that identify cars and plan motor actions, and these other parts could be said to be introspectively aware of the lower level data without any consciousness being present.

Higher order thought

Rosenthal (1986) starts by defining a *mental* state as a conscious or unconscious state that has sensory or intentional properties. These mental states are claimed to be conscious when they are accompanied by a higher-order thought and mental states without a higher order thought are said to be unconscious. Rosenthal claims that this presence or absence of higher order thoughts explains the consciousness or unconsciousness of mental states.

The problem with this account is that it is little more than a pseudo explanation that is introspectively and empirically unfounded. Rosenthal admits that we are unaware of our higher order thoughts, but claims that this is a necessary feature of his theory. If higher order thoughts were conscious, then an infinite chain of higher order thoughts would be needed to make each of the previous higher order thoughts conscious. To avoid this problem, Rosenthal claims that the higher order thoughts are unconscious and only become conscious when they are accompanied by third order thoughts. Whilst the unconsciousness of higher order thoughts is necessary to Rosenthal's theory it does mean that their existence cannot be established through introspection. Since higher order thought theory can hardly be said to be grounded in empirical data about the brain, it is left as something that 'explains' phenomenal consciousness on the basis of something that is itself completely ungrounded and unexplained.

Rosenthal (1986) argues that one of the benefits of his theory is that it offers some kind of explanation of consciousness and "If nothing were more basic to us than consciousness, there

would be nothing more basic in terms of which we could explain consciousness. All we could do then is try to make consciousness more comprehensible by eliciting a sense of the phenomena in a variety of different ways." (p. 352). The position of this thesis is that phenomenal experience is one of the most basic 'things' that there is and we need to elicit a sense of the phenomena in a variety of different ways before it we can start to hypothesize about its causes.[15]

Phenomenal and access consciousness.

Block (1995) claims that the word consciousness is used in two distinct ways, which he identifies as phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness). P-consciousness is experience and the experiential properties of a state are "what it is like" to have that state - for example, we have P-conscious states when we hear, see, smell, taste and have pains. On the other hand, access-conscious states are representational and their content is available as a premise in reasoning and for the rational control of action. Since many phenomenal contents are also representational, this distinction can be expressed by saying that it is in virtue of the phenomenal aspect of a state's content that it is P-conscious, whereas it is in virtue of a state's representational content that it is A-conscious. Block uses this distinction to argue against the claim that P-consciousness carries out a particular function, such as high level reasoning - a hypothesis that is often put forward in connection with cases of blindsight and epileptic automatism. Whilst A-consciousness is a functional notion, P-consciousness is not, although it might be systematically correlated with certain functions.

Block's separation of phenomenal consciousness from functions at the physical or information-processing level is entirely in keeping with the definition of consciousness in this thesis, which is based on a primary notion of phenomenal experience.[16] However, Block's notion

[15] Other criticisms of higher-order thought theory can be found in Gennaro (2004), Aquila (1990), Byrne (1997) and Rowlands (2001).

[16] However, Section 2.5 will argue that it does not make sense to speak about an *inaccessible* P-consciousness, which cannot be established through scientific investigation.

of access *consciousness* is much less convincing and hinges on his careful definition of what constitutes access to representational states, which enables him to claim that cases of blindsight and epileptic automatism are not A-conscious. It seems to make much more sense to separate the notion of a representational state from consciousness altogether and to speak about conscious and unconscious representational states – instead of introducing a second notion of consciousness to speak about non-phenomenal representational states. Block's claim that A-consciousness and P-consciousness have been historically confused is no doubt true, but this is not a reason to continue to speak about non-phenomenal *conscious* states when unconscious representational states are much more theoretically tractable.

2.4 Metaphysical Theories of Consciousness

One of the central questions in the philosophical study of consciousness has been whether the phenomenal and the physical are two separate realities or substances, or whether one can be reduced to the other. To answer this question a number of metaphysical theories of consciousness have been put forward.

2.4.1 Idealism and Phenomenology

Both idealism and phenomenology emphasise the phenomenal over physical reality. This type of theory ranges from Berkeley's (1988) claim that the concept of material substance is incoherent and ideas are the only reality, to Husserl's (1960) suggestion that we should suspend belief in the physical world and focus on the description of phenomenal experience, which might eventually enable us to ground science in phenomenological data. Although these theories are logically consistent and cannot be disproved, they have not developed a framework that can match science's success at prediction, and the hypothesis of a metaphysically real physical world leads to a much simpler interpretation of the phenomenal world. For example, it is much more useful

to interpret a stone as a real physical object that can be investigated in a variety of different ways, instead of as a collection of ideas that were put into our minds by God. For these reasons, I will set aside idealism and phenomenology in this thesis and focus on theories that accept the metaphysical reality of the physical world.

2.4.2 Interactionist Dualism

Interactionist dualism is the claim that the phenomenal world is a second thinking substance, which is completely distinct from the substance of the physical world (Descartes 1975, Eccles 1994). As our physical bodies move around in the physical world, our physical brains receive data through the senses and pass it to the thinking substance, where it becomes conscious. When our conscious phenomenal states decide upon an action, instructions are passed back to the physical brain, which controls the muscles. Interactionist dualism was first put forward by Descartes (1975), who suggested that data was passed between the two substances through the pineal gland. The main advantage of interactionist dualism is that it makes a very clear distinction between conscious and unconscious representations.

One of the major problems with this theory is that it has great difficulty explaining the interaction between the two substances. The pineal gland is now known to be closely linked to the maintenance of circadian rhythms, and no evidence has been found for the hypothesis that it is the central channel of communication between the phenomenal mind and the physical brain. In fact it is unlikely that there is a single ‘seat of awareness’ anywhere in the brain (Crick and Koch 2003, Edelman and Tononi 2000), and so the dualist has to explain how a shifting pattern of neural activation is passed on to a second substance and how the second substance causally influences the shifting pattern of activation in the brain. No plausible or testable theory about how this could take place has ever been put forward.

A second problem with interactionist dualism is that our greater understanding of the brain is making the thinking substance increasingly redundant. At one time we might have felt that a second substance was needed to explain something as mysterious as imagination, whereas we can now attempt to explain it as the offline activation of sensory processing areas (Kossyln 1994, Kreiman et al. 2000). Similarly, we might have thought that *thinking* needed a second substance to explain it, whereas we can now see how this could be explained as part of our language-processing and imaginative abilities (Damasio 1995). We are moving towards a situation in which we will be able to explain all of the *functions* of the physical brain in terms of neural processes, which will leave nothing for the thinking substance to *do*. This turns the thinking substance of interactionist dualism into a passive recipient of data from the parts of the brain that are the neural correlates of consciousness, with all the processing carried out by the brain's neural mechanisms. This is basically a version of epiphenomenalism, which will be considered next.

2.4.3 Epiphenomenalism

Epiphenomenalism is often put forward as a way of solving the problems connected with a two-way interaction between the thinking and extended substance. Since the physical world is thought to be causally closed, epiphenomenalism advocates a one way interaction in which the phenomenal world 'sits on top' of the physical world and receives information from the physical brain without having any causal influence on it.

This type of theory often emerges from some form of dualism and it can be argued that pantheism and Nagel's (1974) 'something it is like to be something' are also versions of epiphenomenalism. Many examples of physicalism are also implicit or explicit versions of epiphenomenalism, since they generally look to the physical world for the information-processing carried out by the mind and then seek some extra quality or function of the brain that

'throws up' passive phenomenal qualia, whose only function is the indication of underlying physical states.[17] A physicalism that was not epiphenomenal would need to give phenomenal states a causal role, but this is almost never the case, and so physicalism almost always ends up being epiphenomenal about consciousness.

The central and fatal problem with epiphenomenalism is that it completely undermines our ability to talk about phenomenal states. The descriptions of consciousness generated by the physical brain are not causally connected with phenomenal states, and so it is impossible for them to be *about* these states. To illustrate this point, consider a situation in which I am consciously perceiving a green apple. In this case, there are all kinds of causal links from the world to the activity in my visual cortex and epiphenomenalism claims that there are also causal links from the activity in my visual cortex to a second substance in which the green apple becomes conscious. However, since the causal links to the second substance only go in one direction, when I say that I am conscious of the green apple, the activity in my larynx muscles is driven entirely by the physical activity in my visual cortex, and it is completely independent of whether or not there is a conscious green apple in the second substance. This situation is illustrated in Figure 2.1.

[17] Jackendoff's (1987) theory is close to this position, although he does not explicitly embrace the metaphysics of epiphenomenalism: "The elements of conscious awareness are caused by/ supported by/ projected from information and processes of the computational mind that (1) are active and (2) have other (as yet unspecified) privileged properties." (p. 23). As Jackendoff points out, in this interpretation consciousness does not have any effect on the world: "Yet another way of looking at Theory II and its corollaries is as a claim that consciousness is *causally inert*. This may seem harmless enough until we realize its ugly consequence: *Consciousness is not good for anything*. The only way it can be good for anything is for it to have effects, and such possibility has just been denied. Again, the only construal of 'Consciousness is good for purpose *X*' within Theory II is as 'The computational states that cause/support/project consciousness are good for purpose *X*,' which does not exactly have the same ring of victory to it." (Jackendoff 1987, p. 26).

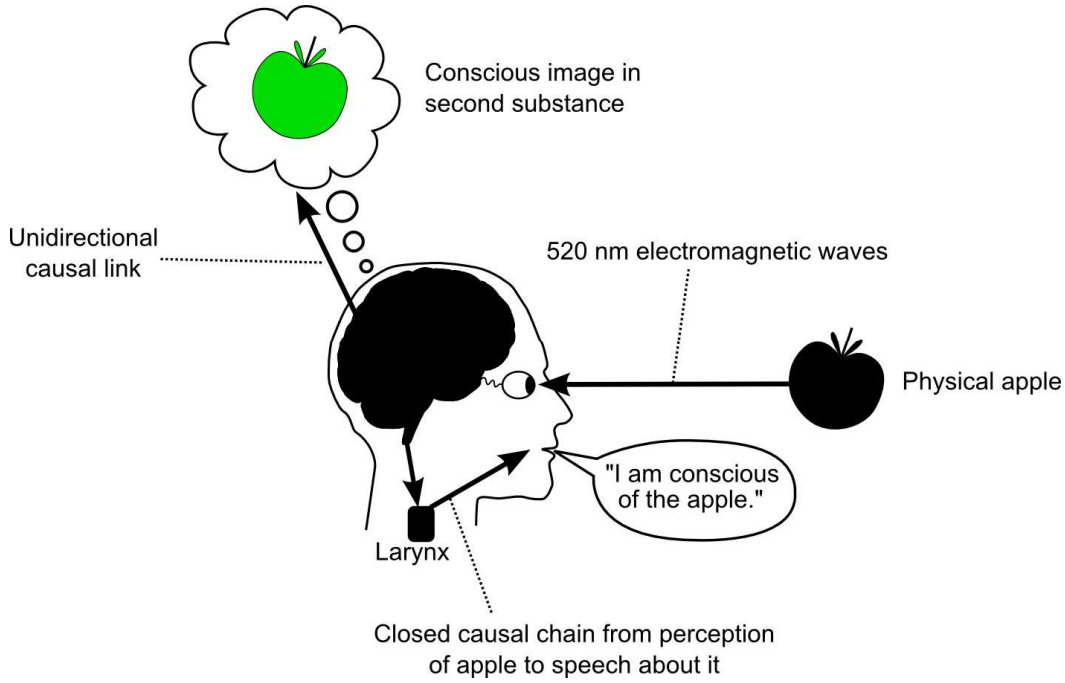

Figure 2.1. Within epiphenomenalism there is only a one-way causal chain from physical reality to the second substance, and so our statements about consciousness are completely independent of our actual consciousness

Since there is complete causal dissociation between the contents of our consciousness and our speech about it, I will continue to state that “I am conscious of the apple” regardless of whether I am actually conscious of an apple, a banana or not conscious at all (see Figure 2.2). If conscious experience cannot affect physical reality, then our physical bodies have no evidence for their claim to be conscious: there is simply no way in which our physical bodies could ever know that there is an epiphenomenal second substance.

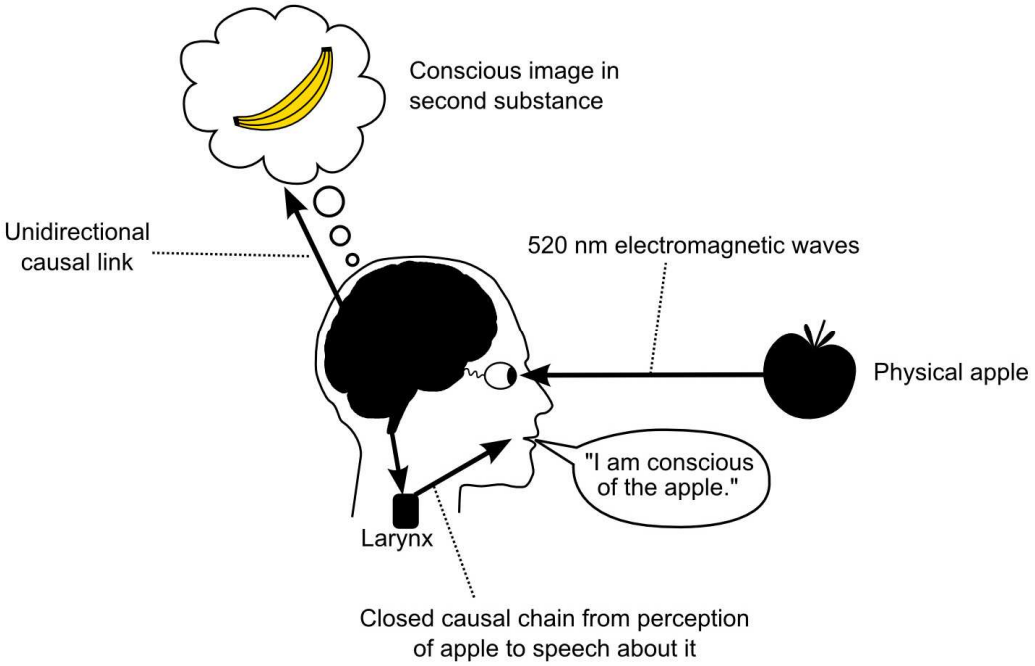

Figure 2.2. According to epiphenomenalism, the contents of our consciousness have no effect on our speech. Although the apple sense data is transformed into a conscious image of a banana, my physical brain and body continues to state that I am conscious of an apple. Even if I became conscious of this disparity, I would be unable to talk about it because there is no causal influence from my consciousness to the physical world.

2.4.4 Physicalism

One of the most popular theories about consciousness is that there is only one substance, the material world described by physics, and consciousness has something to do with the information, processes, functions or structures within this physical substance (Poland 1994, Kim 2005). This material substance is associated with phenomenal states when it is arranged into working brains, and not conscious when it is arranged into rocks or chairs. The advantage of dualism was that it could easily accommodate properties, such as redness or the smell of lavender, within a second substance. In rejecting this, physicalism leaves itself with the problem that phenomenal properties are absent from the world described by physics. However we arrange

the physical world we will never arrange it into redness or the smell of lavender.[18] These difficulties with integrating the physical and phenomenal worlds are discussed next.

2.4.5 The Easy, Hard and Real Problems of Consciousness

> In 1989 the philosopher Colin McGinn asked the following question: "How can technicolor phenomenology arise from soggy gray matter?" (1989: 349). Since then many authors in the field of consciousness research have quoted this question over and over, like a slogan that in a nutshell conveys a deep and important theoretical problem. It seems that almost none of them discovered the subtle trap inherent in this question. The brain is not grey. The brain is colorless.
>
> Metzinger (2000, p. 1)

Chalmers (1996) put forward a distinction between the 'easy' problem of explaining how we can discriminate, integrate information, report mental states, focus attention, etc., and the hard problem of explaining how phenomenal experience could arise from physical matter. Although solving the 'easy' problem is far from easy, we do at least have some idea how it can be done. On the other hand, although many theories have been put forward about the hard problem, it can be argued that we have no real idea about how to solve it.[19]

The hard problem of consciousness generally gains its intuitive force from an exercise in which we imagine (or perceive) a grey brain, imagine (or perceive) the colour red and then try to think how the colour red could be generated by the grey brain. This is a hard problem because we cannot *imagine* how the information-processing functions of the brain, for example, could lead to phenomenal red.

The problem with this attempt to imagine the hard problem of consciousness is that the physical brain is completely non phenomenal in character and so the hard problem of

[18] Although we have no problem *correlating* redness with electromagnetic signals of 428,570 GHz and lavenderness with molecules of Borneol, Geraniol, Linalool, Lavendulyl acetate, Linalyl acetate and Cineol.

[19] There has been extensive discussion in the literature on consciousness about whether Chalmers' hard problem is in fact a genuine problem and the different ways in which it can be tackled. Representative positions in this area can be found in Goguen and Forman (1995, 1996), Shear (1997) and Gray (2004).

consciousness can only be imagined by smuggling in our phenomenal representation of the physical brain and then trying to connect this phenomenal brain with a paradigmatic phenomenal red 'quale'. When we think that we are imagining the physical world we are actually imagining our phenomenal representation of the physical world. The *hard* problem of consciousness is a puzzle about how phenomena can cause phenomena, whereas the *real* problem of consciousness is about how the phenomenal world is connected with real physical neurons, which we can describe scientifically and mathematically, but cannot perceive or imagine in any way. This difference between the hard problem of consciousness and what I am calling the real problem of consciousness is illustrated in Figure 2.3.

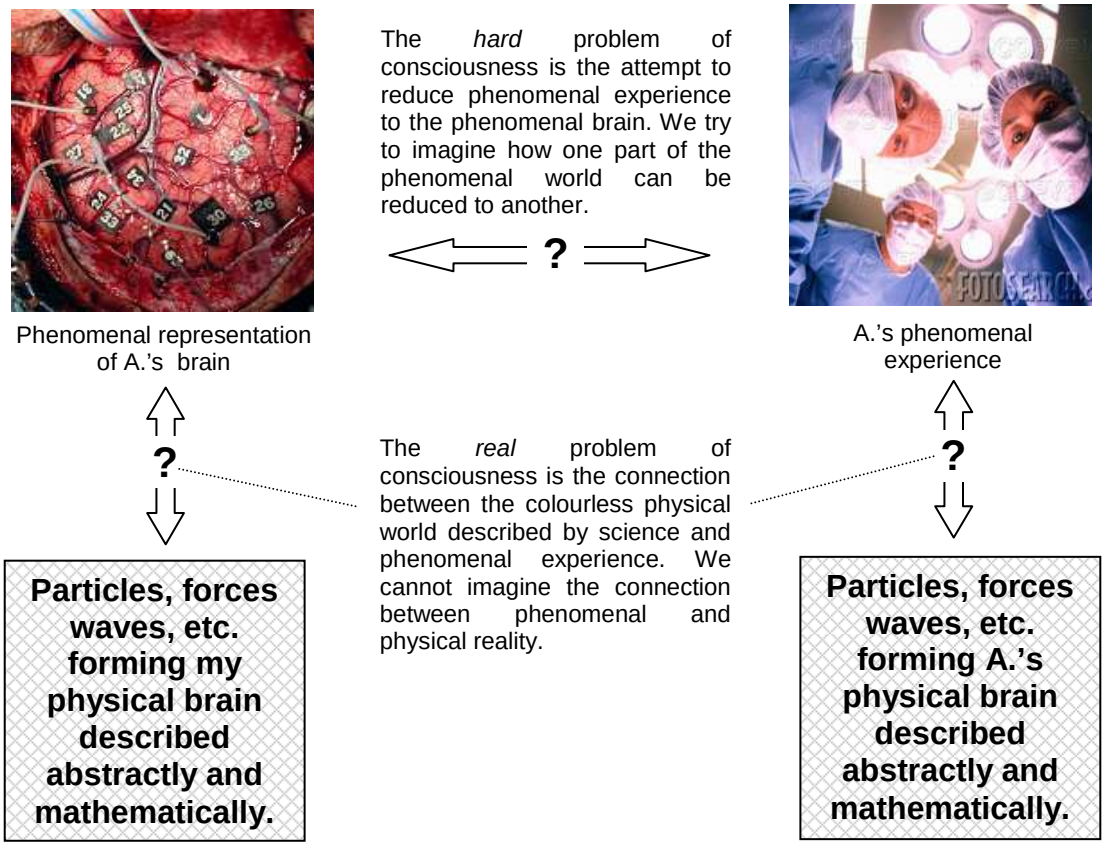

Figure 2.3. The relationship between the hard and the real problem of consciousness. The brain picture on the left is my phenomenal representation of person A's brain. The surgeon picture on the right is A's phenomenal reality (the operation is under local anaesthetic).

The hard problem of consciousness attempts to reduce one part of phenomenal reality (the colour red) to another part of phenomenal reality (the phenomenal brain). Discussions of consciousness often get intuitively or imaginatively stuck on this hard problem, which will be never be solved because intuition and imagination are simply not applicable.

Real scientific problems are solved by creating abstract descriptions of phenomenal observations and hypothesising forces or other features of the physical world that link these abstract descriptions with one other. In this respect, the real problem of consciousness is no different from any other scientific theory since we have phenomenal observations of brains and phenomenal observations of our experiences and science can look for regularities between them, which we may eventually be able to explain using a theory of consciousness. It is relatively easy to describe the brain because we can use mathematics, physics and biology to precisely specify its physical aspects. Precise descriptions of phenomenal states are much more of a challenge because up to this point we have relied on natural human language for our phenomenological descriptions. Whilst statements like “I am experiencing a red blob in the left hand corner of my visual field” might be adequate for our current research on consciousness, there are good reasons why a more precise language for phenomenology might be more appropriate for a science of human consciousness, and a number of arguments are put forward in Section 4.4 why a markup language, such as XML, is already needed for the description of the phenomenology of artificial systems.

Once we have obtained precise descriptions of the physical and phenomenal states we can look for correlations between them and use theories about consciousness to make predictions about the phenomenal states that are associated with the physical states and the physical states that are associated with the phenomenal states. The accuracy and falsifiability of these predictions (Popper 2002) will depend on the precision of the physical and phenomenal

descriptions. This scientific approach to the real problem of consciousness is illustrated in Figure 2.4.

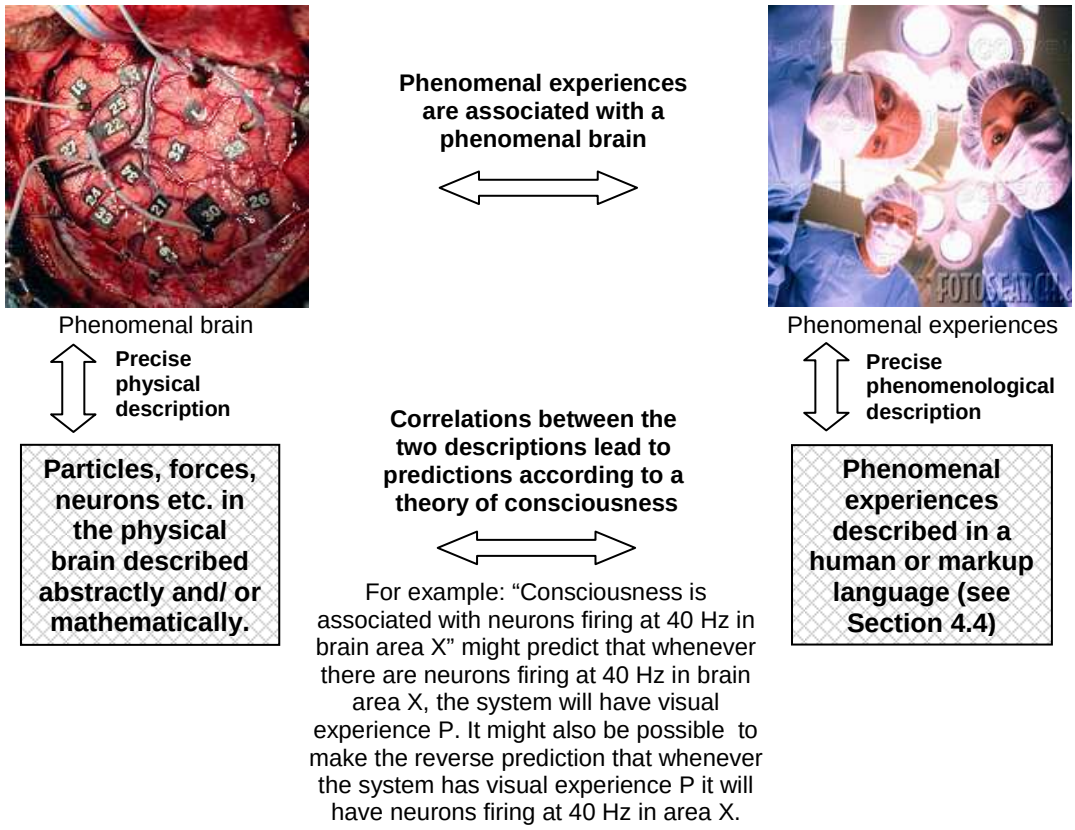

Figure 2.4. First stage in a scientific solution to the real problem of consciousness. Precise descriptions are formulated of the physical brain and the phenomenal experiences associated with the physical brain, and these are used to identify correlations between the physical and phenomenal worlds. The predictions that different theories of consciousness make about these correlations can then be experimentally tested.

If we can discover theories that make good and perhaps perfect predictions about the relationships between the physical and phenomenal worlds, then we might start to think about how we could *explain* these predictions. A good example of this move from prediction to explanation is given by the evolution of our theories about the expansion of gases. A key stage in this work was Boyle's law, published in 1662, which predicts that the pressure, P, and the volume, V, of a gas are related to a constant value, k, according to Equation 2.1:

$$PV = k \tag{2.1}$$

This equation is an empirical observation about the relationship between the pressure and volume of a gas, which can be used to predict how a fixed quantity of gas will respond to a change in pressure or volume according to Equation 2.2:

$$P_1V_1 = P_2V_2\,, \tag{2.2}$$

where P_1 and V_1 are the pressure and volume before the change and P_2 and V_2 are the pressure and volume after the change. These predictions made by Boyle's law were later *explained* by Bernoulli (1738), who showed how Equation 2.1 could be derived by applying Newton's laws to the motion of large numbers of molecules.

In the case of consciousness, if we can establish precise relationships between the phenomenal and physical descriptions, then we may eventually be able to move on to an explanation.[20] The form that such an explanation could take will probably only become clear once we have done a lot more work on the identification of correlations between the phenomenal and physical worlds, which will be covered next.[21]

[20] Since causal relationships are inherently temporal, it is coherent to claim that a phenomenal event causes a later physical event or a physical event causes a later phenomenal event, but it does not make sense to try to use a causal relationship to explain the co-occurrence of phenomenal and physical events at the same point in time - unless the common cause is something that is neither phenomenal nor physical and occurs before the simultaneous phenomenal and physical events.

[21] Coward and Sun (2007) put forward a general form for scientific theories of consciousness. Whilst their interpretation ignores the phenomenal/ physical distinction that has been argued to be essential for any science of consciousness, their suggestions about the hierarchical nature of scientific theories fit in well with the approach to synthetic phenomenology put forward in Chapter 4.

2.5 Correlates of Consciousness

2.5.1 Introduction

The discussion of metaphysical theories of consciousness has shown that systematic identification of the correlates of consciousness is an essential first step in the development of a scientific theory. Many people have started on this work and current investigations are mainly focused on the correlation between consciousness and the human brain, both because people are paradigmatic examples of conscious systems and because they are the only species that can make verbal reports about their phenomenal states. Although a great deal of work has been carried out on the neural correlates of consciousness in recent years (Chalmers 1998, Metzinger 2000), the firing of real biological neurons is not sufficient for consciousness, and might not even be necessary, and so this section covers a broad spectrum of potential correlates of consciousness (PCCs).[22]

The ultimate aim of the search for correlates of consciousness is to identify a list of necessary and sufficient conditions that would predict with certainty when a physical system is associated with phenomenal states and describe the contents of these states when they occur. Although our scientific theories would be much simpler if we found a single correlate of consciousness, it is possible that consciousness is correlated with a multiplicity of factors – for example, a particular combination of temperature and neural activity might be necessary. It is also possible that some factors will be partially correlated, which would only allow probabilistic predictions to be made about whether a system is conscious and what it is conscious of.

Adequate knowledge about the correlates of consciousness will enable us to predict phenomenal states from physical states and physical states from phenomenal states, but it will not prove that consciousness is causally dependent upon physical states any more than it will

[22] Without a commonly agreed definition of consciousness it is impossible to say whether we have identified *any* correlates of consciousness at this stage. For this reason, I will interpret all correlates of consciousness as *potential* in this thesis.

prove that physical states are causally dependent on consciousness. It is an open question how our theories about consciousness will evolve once we have mapped out the correlations between the phenomenal and physical worlds.

2.5.2 Potential Physical Correlates of Consciousness

The human brain is a paradigmatic example of a system associated with consciousness, and so any of its physical attributes are PCCs. None of these potential correlates is likely to be sufficient for consciousness because it is generally assumed that no consciousness is present when we are in deep sleep or a coma when the physical attributes remain unchanged.[23] Some examples of physical PCCs are as follows:

1. Volume of 1.4 litres.
2. Temperature of 310 K.
3. Weight of 1350 g.
4. Created after 1000 BCE.
5. Created through a process of natural selection.
6. Reflects light with a wavelength of 650 nm.[24]
7. Living neurons assembled from biological amino acids.
8. Haemoglobin.
9. Oxygen.
10. Rate of processing.

[23] This assumption may not hold if Zeki's (2003) notion of micro consciousnesses is correct. In this case one or more consciousnesses could be associated with a person in deep sleep or coma, which would not be verbally expressed because they are not integrated with the memory or vocal systems.

[24] I am using this as a convenient shorthand for the fact that the brain looks pinkish. In fact almost every non-black object reflects light of 650 nm to some degree and more care would be needed to formulate an accurate physical description of this property of the brain.

2.5.3 Potential Neural Correlates of Consciousness

Activity in biological neurons has been shown to be strongly correlated with consciousness and a large number of experiments have been carried out that have attempted to distinguish between neural activity that takes place when we are not conscious – in deep sleep or a coma, for example – and neural activity that is correlated with conscious experience. The emerging consensus is that the neural correlates of consciousness are likely to be distributed over many different brain areas - see, for example, Edelman and Tononi (2000), Crick and Koch (2003) Dehaene and Naccache (2001) or Zeki et al. (1998, 2003) - and the coordination between these areas might be achieved by synchronization of neural firing (Singer, 2000), NMDA synapses (Flohr, 2000), connections to thalamic nuclei (Newman et. al., 1997) or some combination of these mechanisms. The distributed neural correlates of the conscious model of our bodies are described in Melzack (1992) and Damasio (1995, 1999). Further discussion of the neural correlates of consciousness can be found in Chalmers (1998), Metzinger (2000) and Noë and Thompson (2004).

2.5.4 Potential Functional and Cognitive Correlates of Consciousness

The human brain can be analysed from the perspective of the large number of functions that it carries out, many of which might be correlated with consciousness. These range from the low level input and output functions of ion channels and neurons, up to higher level functions, such as perception, memory and cross-modal integration. The brain also carries out a number of cognitive functions that have been linked to consciousness, such as emotional evaluation of a situation, internal representations of the self, imagination and attention.

2.5.5 Experimental Determination of the Correlates of Consciousness

> We will focus on the notion of consciousness as such by contrasting pairs of similar events, where one is conscious but the other is not. The reader's conscious image of this morning's breakfast can be contrasted with the same information when it was still in memory, and unconscious. What is the difference between conscious and unconscious representations of the same thing? Similarly, what is the difference between the reader's experience of his or her chair immediately after sitting down, and the current habituated representation of the feeling of the chair? … All these cases involve contrasts between closely comparable conscious and unconscious events.
>
> These contrasts are like experiments, in the sense that we vary one thing while holding everything else constant, and assess the effect on conscious access and experience.
>
> Baars (1988, pp. 18-19)

To decide which PCCs are *actually* correlated with consciousness we need to measure the level of consciousness when the potential correlates are present individually and in different combinations, until we find the set that is systematically correlated with consciousness.[25] For example, if the human brain has attributes W, X, Y and Z, and removing Z and W has no effect on the consciousness of the system, but removing either X or Y individually or X and Y together leaves the system unconscious, then we can conclude that X and Y are necessary for consciousness. However, we can only conclude that X and Y are *sufficient* for consciousness if the human brain has no other attributes in addition to W, X, Y and Z that might be correlated with consciousness. For example, if the attribute C was left unchanged during the experiments, then it is possible that X + Y is not sufficient for consciousness and C has to be included as well. Some of the problems connected with this experimental process will now be covered in more detail.

[25] It is possible that there is more than one set of correlates of consciousness. For example, neurons constructed with silicon chemistry and neurons constructed using carbon chemistry may both be correlated with consciousness.

Selection of potential correlates

The first step in establishing the correlates of consciousness is to choose an initial set of potential correlates for experimentation. Since we know almost nothing about the link between the phenomenal and physical worlds, we cannot exclude anything with certainty, but we are likely to make more rapid progress if we start with a list of candidates that are broadly compatible with the Western scientific outlook.[26] To begin with, we can exclude potential correlates that are hard or impossible to test, such as the property of being created after 1000 BCE. However, this still leaves a potentially infinite number of testable PCCs, which we can only narrow down using our intuition about their potential link with consciousness.

A first problem with the use of intuition for this task is that our intuitions about consciousness are all taken from our phenomenal experiences and we have never experienced a direct link between phenomenal and physical reality. However, we do have a lot of experience of correlations between our phenomenal experiences and our phenomenal measurements of the physical world, which can be imagined and intuited. The intuitive exclusion of factors will have to be limited to human cases because we have never directly experienced animal or machine consciousness and any 'observations' of animal or machine consciousness have been extremely indirect, inferential and based on what we believe about human consciousness. Although we cannot reliably intuit whether a stone, for example, is capable of conscious states, we can discard many of the unique attributes of stones from our initial list of potential correlates because it is likely to be more profitable to start with attributes of humans, which we know to be conscious already.

A second problem with the use of intuition is that it can vary widely between people. For example, some people have an intuition that size is relevant to consciousness because all of the conscious systems that they have encountered have been within a certain range of sizes. This

[26] See Footnote 13.

leads to clashes of intuition in which some people are unwilling to believe that a system the size of the population of China could be conscious, whereas others do attribute consciousness to larger or smaller systems. When clashes of intuition do occur, it is generally better to leave the attribute as a PCC so that its validity can be established scientifically. In the longer term it is hoped that our intuitions about consciousness can be grounded by identifying the regularities in experience that gave rise to them.

Measurement of the physical system

To identify correlations between the physical and phenomenal worlds we need to measure changes in the physical system. Most of the potential physical correlates can be gauged using standard weight, volume and chemical measures and we have a wide range of ways of monitoring neural activity in the brain, such as EEG, fMRI, PET or implanted electrodes.[27] The functional and cognitive correlates of consciousness can be measured using psychological tests, and the functions of particular brain areas can be probed using patients with brain damage, animal models or by applying transcranial magnetic stimulation. All of these measurement techniques produce phenomenal representations of different aspects of the physical brain.[28]

Measurement of consciousness

Experiments on the PCCs also have to measure whether consciousness is associated with the system and, if consciousness is a graded phenomenon, the *amount* of consciousness that is present. Since consciousness cannot be detected with scientific instruments, its presence is

[27] These technologies are in the early stages of development and their low temporal and/ or spatial resolution limits the precision with which the neural correlates of consciousness can be identified.

[28] One potential measurement issue is that we might have to measure the system's *capacity* for some functions as well as the actual exercise of them within the system. For example, if it is possible to have conscious experiences that do not involve imagination, then it could be argued that imagination is not a necessary correlate of consciousness. However, this does not rule out the possibility that a *capacity* for imagination is a necessary correlate. The latter can only be ruled in or out by seeing if there are any conscious (probably brain damaged) people who lack all capacity for imagination. An example might be the amnesiac patients studied by Hassabis et al. (2007), who are not only bad at remembering the past, but at imagining new experiences as well.

established through first person reports in language, first person observations that are remembered and reported later, or through behaviour that is interpreted as the result of conscious experience – this is the only technique that can be used with animals, such as monkeys, which are trained to respond to a stimulus that is assumed to be conscious.[29] In all of these cases, the presence of consciousness is established through *behaviour* - our own behaviour when we write down our introspective observations, the verbal behaviour of a reporting subject or non-verbal animal or human behaviour.

A first problem with behavioural measures is that they are often inaccurate, especially when some form of brain damage is involved. This can occur when people are reporting everything in good faith with no intention of deceiving the experimenter. For example patients with Anton's syndrome claim to be able to see perfectly when in fact they are clinically blind and anosognosia patients will make claims about being able to use a paralyzed limb, for example, and confabulate wildly to explain its lack of movement (Ramachandran and Blakeslee 1998).

A second issue with the measurement of consciousness through immediate or deferred behaviour is that certain types of behaviour could themselves be correlates of consciousness. Since some behaviours, such as the statement "I am conscious right now", are more correlated with consciousness than anything else that can be varied in an experiment, this possibility cannot be completely ruled out. However, it does seem reasonable to suppose that a verbal report of my dream was not necessary for the occurrence of the dream, which I would have experienced independently of any external behaviour.

A third problem is that the probing of the conscious states might affect the conscious states themselves, either by distorting our memories of the conscious states or by priming us to

[29] See, for example, Logothetis' (1998) work on the neural correlates of consciousness. In these experiments macaque monkeys were trained to pull different levers in response to different images and Logothetis recorded from a variety of visual cortical areas in the awake monkey whilst it performed a binocular rivalry task.

interpret the situation in a particular way. Furthermore, as Dennett points out in his discussion of Orwellian and Stalinesque revisions (Dennett, 1992: pp. 101-38), the ordering of events can be ambiguous at small time scales, and so when we report our conscious experience of a visual illusion, for example, there is an ambiguity between a false memory of something that did not consciously take place and a correct memory of a false conscious event. Dennett (1992) uses this ambiguity to argue that there is no single Cartesian Theatre in which a determinate stream of consciousness takes place and there are just multiple drafts of narrative fragments under constant revision by multiple brain processes. These multiple drafts can be probed at different times and places to precipitate different narratives from the subject, but there is no single canonical stream of consciousness.

The most serious problem with a behavioural measure of consciousness is that it limits us to experiments that change the behaviour of the system. If an experiment does not alter the system's behaviour between the time of the experiment and the system's death, then it is impossible to tell if it has changed the system's phenomenal states. The behaviour-neutral experiment might have changed the consciousness of the system (in this case, the attributes under investigation are necessary and perhaps even sufficient for consciousness), or it might have had no effect at all on the system's consciousness (the attributes are extraneous factors that should be eliminated from the list of potential correlates), and we have no way of telling which is the case. The physical aspects of a system that were covered in Section 2.5.2 are the most behaviour-neutral, since size, temperature and material can all be changed whilst the behaviour is held constant, which makes it impossible to measure the correlation between any of these factors and consciousness. To make this point clearer I will look at an experiment that is often discussed in the literature in which part of the brain is replaced by a functionally equivalent chip.

2.5.6 Brain Chip Replacement

To identify the necessary and sufficient correlates of consciousness each PCC needs to be tested independently. Consciousness might be correlated with some of the functions carried out by the physical brain and/or with the biological material of the brain, and so we need experiments that change the material of the brain whilst holding the functions constant, and experiments that change the functions of the brain whilst holding the material constant. One way of holding the functions constant and changing the material is to replace part of the brain by a functionally equivalent silicon chip. For example, if the replacement of part of the lateral temporo-occipital cortex with a functionally equivalent chip caused a person to lose consciousness of movement information,[30] then we could conclude that the brain's biological substrate and functions are *both* necessary for consciousness. Although this is currently only a thought experiment, people are working on the development of a silicon hippocampus,[31] and so it might be possible to carry out this experiment in the future.

The central problem with this experiment is that the chip carries out exactly the same functions as the brain area that it is replacing, and so the overall functioning of the brain – and the behaviour of the person - is not altered by the operation. As Moor (1988) and Prinz (2003) point out, neither an external observer nor the person who received the chip would observe any effect of the implant on consciousness. An outside observer would not detect the replaced part because the function of the lateral temporal-occipital cortex would still be carried out by the chip. The person would continue to report and describe the movement information processed by affected area, even though there might not be any consciousness of movement present. From an

[30] This example is based on a patient studied by Zihl et. al. (1983, p. 315), who completely lost her ability to perceive motion after bilateral cerebral lesions in the lateral temporo-occipital cortex: "She had difficulty, for example, in pouring tea or coffee into a cup because the fluid appeared to be frozen, like a glacier. In addition, she could not stop pouring at the right time since she was unable to perceive the movement in the cup (or a pot) when the fluid rose. Furthermore the patient complained of difficulties in following a dialogue because she could not see the movements of the face and, especially, the mouth of the speaker."

[31] See http://www.newscientist.com/article.ns?id=dn3488.

outside point of view, this would not even seem like a confabulation because the visual system would be working perfectly.

A first-person perspective does not help matters either. Since the chip is functionally connected to the rest of the brain in the same way that the lateral temporal-occipital cortex was before the operation, the person's language centres should report phenomenal movement in the same way that they did before, and so they will continue to think that they are experiencing movement, even if they have no consciousness of movement. Searle (1992, pp. 66-7) thinks that the person might feel *forced* to say that they are experiencing movement whilst they remain conscious of the fact that there is no phenomenal movement present. However, if the person was conscious of this compulsive language behaviour, then they would be able to remember and report it at a later time, which would be a functional change in the system that has been excluded by this experiment. It seems that even a first-person perspective cannot be used to decide whether consciousness is affected by the replacement of biological neurons with a functionally equivalent chip.

Against this Chalmers (1996) argues that verbal behaviour and consciousness would be very tenuously connected if we could lose our conscious experience of movement and yet continue to describe movement using language. The problem with this objection is that the implantation of a chip involves invasive surgery and it is not uncommon for people with brain damage to be systematically mistaken about their experiences and confabulate to an extraordinary extent to cover up their deficiency. As was pointed out in the previous section, people with Anton's syndrome are blind and yet insist that they can see perfectly and hemineglect patients will bluntly assert that a paralysed arm is functionally normal. Faced with these cases, it cannot be assumed that it is impossible for us to be systematically mistaken about our phenomenal states. Further criticisms of Chalmers' argument can be found in Van Heuveln et. al (1998) and Prinz (2003).

The brain-chip experiment can be applied to part of the brain or to the entire brain and in all cases the system's behaviour will remain constant. The same argument applies to other experiments on the brain's material, such as a change in temperature or the use of synthetic blood to probe the link between haemoglobin and consciousness.[32] Both a change in temperature and the exchange of real for artificial blood would leave the behaviour of the patient untouched, and we would be left wondering whether it removed the consciousness and left the behaviour intact or had no effect on consciousness. As Harnad (2003) points out, all our attributions of consciousness to a system are based on its behaviour, and so something that does not change the behaviour cannot be separated out as a correlate of consciousness:

> The only way to sort out the relevant and irrelevant properties of the biological brain, insofar as consciousness is concerned, is by looking at the brain's behaviour. That is the only non-telepathic methodology available to us, because of the other-minds problem. The temptation is to think that 'correlations' will somehow guide us: Use brain scanning to find the areas and activities that covary with conscious states, and those will be the necessary and sufficient conditions of consciousness. But how did we identify those correlates? Because they were correlates of behaviour. To put it another way: When we ask a human being (or a reverse-bioengineered robot) 'do you feel this?' we believe him when he says (or acts as if) he feels something – not the other way round: It is not that we conclude that his behaviour is conscious because of the pattern of brain activity; we conclude that the brain activity is conscious because of the behaviour.
>
> Harnad (2003, p. 74)

If some PCCs cannot be ruled in or out, then *we will never be able to identify a list of necessary and sufficient correlates of consciousness and we will never be able to tell for* ***certain*** *whether a system is associated with phenomenal states*. This distinction between correlates of consciousness that can and cannot be separated out will now be formalized as a distinction between type I and type II correlates of consciousness. Type I PCCs are behaviour neutral and so

[32] The temperature change would have to be carried out so that it did not affect the functionality of the brain or allowed the same functionality to take place over longer time scales. The synthetic blood would have to be one of the varieties that was not based on haemoglobin.

their link with consciousness cannot be experimentally tested; type II PCCs do affect the behaviour of the system and their impact on consciousness can be measured. This distinction will now be discussed in more detail and it will used to address questions about the potential consciousness of non-biological systems in Chapter 4.

2.5.7 Type I Potential Correlates of Consciousness

Type I PCCs are either behaviour-neutral or they cannot be separated from the behaviour that is used to measure consciousness in a system. Their key characteristic is that no experimental measure of their connection with consciousness can be devised or suggested. Many PCCs are type I because they can be changed independently of the functional properties of the system. The brain-chip replacement experiment illustrates how this is true for the material substance of the brain and the rest of the physical PCCs in Section 2.5.2 are all type I as well. The second class of type I PCCs is linked to our ability to remember and/or report phenomenal experiences. A change to the system that eliminates its ability to express its phenomenal experiences or prevents it from remembering them for later expression cannot be used to test for correlations with consciousness because it destroys the measuring instrument that is needed for the experiments. Memory and vocalisation/ behaviour can be removed individually – for example, in short term memory loss patients or REM sleep - but if both are lost together, then we can no longer measure consciousness in the system. For example, if Zeki's (2003) notion of micro-consciousness is correct, there could be consciousness in deep sleep and coma, which cannot be remembered or reported because key brain areas are inactive or damaged. This suggests that some forms of global integration and binding might also be type I PCCs: if there is no integration between the visual cortex and other parts of the brain, then there will be no reports or memories of visual

experience. The loss of integration could have eradicated visual consciousness from the system or it could have eliminated the system's ability to remember and report visual experience.[33]

2.5.8 Type II Potential Correlates of Consciousness

Type II PCCs can be separated out using behaviour and there is no overlap with the parts of the system that are used for measuring or reporting consciousness. When a type II PCC is removed or altered, the system's reports of conscious states can change. Activity in particular brain areas is a type II correlate because we can vary this activity through transcranial magnetic stimulation or observe brain damaged patients and measure the change in consciousness through verbal or other behaviour. Functional correlates also fall into this category because it is conceivable that we could disable a person's capacity for imagination or emotion, for example, and then probe their conscious states.

2.6 Three Theories of Consciousness

2.6.1 Introduction

The distinction between type I and type II PCCs can be applied to theories about consciousness:

- Type I theories of consciousness cannot be experimentally validated, either because they are based on type I correlates or because they are metaphysical statements of belief about the world that can never be tested. This type of theory is essentially an *a priori* statement of belief about the world that sets out a framework for interpretation and is completely un- or pre-scientific in character.
- Type II theories of consciousness *can* be empirically verified through experiments because they are based on type II PCCs.[34]

[33] It is even conceivable that we are conscious when dead, but unable to produce any form of behavioural output.

It would be an impossible task to examine all type II theories in this thesis, and I have decided to focus on Tononi's (2004) information integration theory of consciousness, Aleksander's (2005) axioms and Metzinger's (2003) constraints, which will be used to make predictions about the consciousness of the neural network that is described in Chapter 5.[35] Tononi's information integration theory of consciousness was chosen because it is a numerical method that can be used to automatically make predictions about the conscious parts of an artificial neural network. Aleksander's (2005) axiomatic theory was selected because it has been influential within the machine consciousness community and it provides a nice link between cognitive mechanisms and phenomenal consciousness. Metzinger's (2003) constraints were chosen because they are comprehensively worked out at the phenomenal, functional and neural levels and three of his constraints can be used to define a minimal notion of consciousness. Taken together, these three theories cover the cognitive characteristics of consciousness and some of its potential neural correlates, and it is fairly clear how they could be used to analyse a system for consciousness. Although I am focusing on these three theories in this thesis, the approach to machine consciousness that I am developing is quite general and can easily be extended to other type II theories.

The rest of this section gives an overview of Tononi's, Aleksander's and Metzinger's theories of consciousness, which will be used to demonstrate how detailed predictions can be made about the consciousness of a system using different theories. As Crick and Koch (2000) point out, a comparison between predictions and empirical measurements will eventually determine which theories are accepted and rejected by science:

---

[34] All theories about consciousness operate within a framework of assumptions that is *a priori* at some level. However, type I theories of consciousness will never be empirically verifiable within the current scientific paradigm, whereas it may be possible to test type II theories.

[35] The most serious omission is global workspace theory (Baars 1988), which has been influential in research on consciousness and machine consciousness. An overview of machine consciousness work in this area can be found in Section 3.5.6.

> … while gedanken-experiments are useful devices for generating new ideas or suggesting difficulties with existing ideas, they do not lead, in general, to trustworthy conclusions. The problem is one that should be approached scientifically, not logically. That is, any theoretical scheme should be pitted against at least one alternative theory, and *real* experiments should be designed to choose between them.
>
> Crick and Koch (2000, p. 103)

In this thesis, Tononi's, Aleksander's and Metzinger's theories are being used to demonstrate how detailed predictions can be made about the consciousness of a system, and in the future it is hoped that it will be possible to compare these detailed predictions with a system's reports about consciousness. For this purpose only minor improvements or criticisms are necessary and no attempt will be made to integrate the three theories together or to put forward a new theory of consciousness.

2.6.2 Information Integration

The theory of information integration was developed by Tononi and Sporns (2003) and elements of it are also covered in Edelman and Tononi (2000). Information integration is measured using the value Φ, which is the amount of causally effective information that can be integrated across the informational weakest link of a group of elements. The information integration theory of consciousness is the claim that the capacity of a system to integrate information is correlated with its amount of consciousness and the quality of consciousness in different parts of the system is determined by the informational relationships (Tononi 2004). To test the link between information integration and consciousness Tononi and Sporns (2003) and Tononi (2004) evolved neural networks with different values of Φ and showed how they are structured in a similar way to the parts of the brain that are correlated with consciousness.

To measure the information integrated by a subset of elements, S, the subset is divided into two parts, A and B. A is then put into a state of maximum entropy (A^{HMAX}) and the entropy of B is measured. In neural terms, this involves trying out all possible combinations of firing

patterns as outputs from A, and measuring the differentiation of the firing patterns produced in B. The *effective information* (EI) between A and B is a measure of the entropy or information shared between them, which is given in Equation 2.3:

$$EI(A \rightarrow B) = MI(A^{HMAX}; B), \tag{2.3}$$

where MI(A; B) is given by:

$$MI(A; B) = H(A) + H(B) - H(AB). \tag{2.4}$$

Since A has effectively been substituted by independent noise sources, there are no causal effects of B on A, and so the entropy shared by A and B is due to the causal effects of A on B. EI(A→B) also measures all possible effects of A on B and EI(A→B) and EI(B→A) are in general not symmetrical. The value of EI(A→B) will be high if the connections between A and B are strong and specialized, so that different outputs from A produce different firing patterns in B. On the other hand, EI(A→B) will be low if different outputs from A produce scarce effects or if the effect is always the same.

The next stage in the measurement of effective information, is the repetition of the procedure in the opposite direction by putting B into a state of maximum entropy and measuring its effect on A, giving EI(B→A). For a given bipartition of the subset S into A and B, the effective information between the two halves is indicated by Equation 2.5:

$$EI(A \leftrightharpoons B) = EI(A \rightarrow B) + EI(B \rightarrow A). \tag{2.5}$$

The amount of information that can be integrated by a subset is limited by the bipartition in which EI(A⇋B) reaches a minimum, and to calculate this *minimum information bipartition* the analysis is run on every possible bipartition. Since EI(A⇋B) is bounded by the maximum information available to A or B, EI(A⇋B) has to be normalised by H^{MAX}(A⇋B) when the effective information of each bipartition is compared (Equation 2.6).

$$H^{MAX}(A \leftrightharpoons B) = \min\{H^{MAX}(A); H^{MAX}(B)\}, \quad (2.6)$$

The *information integration* for subset S, or Φ(S), is the non-normalised value of EI(A⇋B) for the minimum information bipartition.

Tononi and Sporns (2003) define a complex as a part of the system that is not included in a larger part with a higher Φ. To identify the complexes it is necessary to consider every possible subset S of *m* elements out of the *n* elements of the system starting with *m* = 2 and finishing with *m* = *n*. For each subset Φ is calculated and the subsets that are included in a larger subset with higher Φ are discarded, leaving a list of complexes with Φ > 0 that are not included within a larger subset with greater Φ. The *main complex* is then defined as the one that has the maximum value of Φ, and Tononi (2004) claims that this main complex is the conscious part of the system. To substantiate his link between Φ and consciousness, Tononi (2004) compares different network architectures with structures in the brain and shows how the architectures associated with high Φ map onto circuits in the brain that are associated with consciousness. The details of the algorithm that was used to calculate Φ are given in Section 7.4.2 along with some optimisations that were developed for large networks.

Information integration is a type II theory because it makes testable predictions about the link between consciousness and high Φ. For example, subjects should only report that they are conscious of information that is held in the main complex and it might be possible to change the amount of information integration in animals and measure the effect on consciousness. The main weakness of Tononi's approach is that it is based on extremely simplified networks consisting of 10-20 elements, which makes it a rather speculative interpretation of circuits in the brain consisting of hundreds of millions of neurons. The positive side of this approach is that it links up with other work on effective connectivity and binding and it is less dependent on a subjective interpretation of the system's constituent parts than other methods – for example, to apply

Tononi and Sporns algorithm we do not have to decide whether a particular layer represents emotions.[36]

2.6.3 Aleksander's Axioms

Aleksander and Dunmall (2003), Aleksander (2005) and Aleksander and Morton (2007c) have developed an approach to machine consciousness based around five axioms that are claimed to be minimally necessary for consciousness. According to Aleksander, this is a preliminary list of mechanisms that could make a system conscious, which should be revised as our knowledge of consciousness develops – a useful starting point that can be used to test ideas and develop the field. These axioms were deduced by Aleksander using introspection and he also identifies neural mechanisms that could implement them in the brain. Each of the axioms will now be covered in more detail.

1. Depiction

Depiction occurs when a system integrates sensory and muscle position information into a representation of an 'out there' world. The key characteristic of depiction is that visual or other perceptual information is integrated with proprioceptive information to give the sensation of something that is *out there*, which is very different from a photographic representation. Aleksander claims that this axiom is implemented in the brain by cells that respond to a particular combination of sensory and muscle information, such as the gaze-locked neurons discovered by Galletti and Battaglini (1989). These cells respond to small visual stimuli only when the monkey's eyes are pointing in a particular direction: if the monkey changes its direction of gaze, different cells respond to the same visual stimulus. Other senses exhibit depiction as well, with touch being the next most depictive, followed by hearing and then smell

[36] See Section 7.4.7 for some other criticisms of information integration.

and taste, which are hardly depictive at all. Depiction is the most important axiom and it is a key mechanism for conscious representation.

2. Imagination.

Imagination occurs when the system recalls parts of the world that are not physically present and this ability can be used to plan actions by constructing sequences of possible sensations. Imagination is linked to the sustained activation of depictive firing patterns, which is likely to depend on feedback or re-entrant connections in the brain. Research on mental imagery suggests that the parts of the brain that are used in sensation are reactivated in imagination (Kossyln 1994, Kreiman et al. 2000), with the difference that they can be active in different combinations, so that we can imagine things we have never encountered before. Many different theories have been put forward about how information in the brain areas involved in perception or imagination is bound together. Aleksander and Dunmall (2000) claim that this is done by associating the different sensory areas with a single location in muscular space, which unifies them into a single object that feels out there in the world. The vividness of imagination decreases in proportion to the degree to which the senses are capable of depiction, and so our most vivid imagined sense is vision, followed by touch and then audition. Smell and taste are almost impossible to imagine or remember accurately.

3. Attention

Attention refers to the process of selecting what we experience in the world and what we think about in our imagination. Our attention can be attracted automatically, for example when we hear a loud noise, or we can purposefully select the parts of the world that we depict or imagine. In the human brain, the superior colliculus is one of the areas that is involved in the selection of the eye position as part of the process of visual attention.

4. Volition

The terminology that is used to describe this axiom has shifted over time, with Aleksander and Dunmall (2003) referring to it as “planning”, whereas Aleksander and Morton (2007c) refer to it as “volition” to distinguish it from rule-based planning processes. This axiom refers to the fact that we are constantly thinking ahead, considering alternatives and deciding what to do next. The neural machinery for this process is the same as that in axiom 2, since the re-entrant neural connections that facilitate imagination also enable the network to move through sequences of states to plan actions. Volition is conscious when it involves depictive areas and the emotions are used to select the plan that is to be executed.

5. Emotion

We have feelings, emotions and moods and use them to evaluate planned actions. Some emotions, such as pleasure and fear, are hardwired at birth, whereas others develop over the course of our lives – for example, the feeling of hurt that we experience when we have been rebuked. Aleksander expects that the neural firing patterns associated with emotions will have distinctive characteristics, which enable them to be associated with perceived and imagined depictive events. As planning proceeds, predicted states of the world trigger neural activity in the emotion areas that determine which plan is selected for execution.

Aleksander’s axioms are a clear set of mechanisms that are a useful starting point for work on machine consciousness. Although I am reluctant to follow Aleksander (2005, pp. 33-4) in claiming an identity between neural activity and conscious sensations, I am happy to interpret the axioms as potential cognitive correlates of consciousness, and to interpret the neural mechanisms behind the axioms as potential neural correlates of consciousness. Aleksander’s axioms are a type II theory because they have been established through introspection and it should be possible to test their correlation with consciousness - for example, by finding people

who lack one or more of the axioms and asking them about their conscious experience. The axiomatic theory also predicts that people without a link between motor information and sensory input should be incapable of depiction, and it might be possible to test this using lesions in a monkey. Aleksander's neural implementation of the axiomatic mechanisms, which he calls the kernel architecture, is summarized in Section 3.5.1.[37]

2.6.4 Metzinger's Constraints

Metzinger (2003) sets out a detailed theory of consciousness that is based around eleven constraints on conscious experience:

1. Global availability
2. Window of presence
3. Integration into a coherent global state
4. Convolved holism
5. Dynamicity
6. Perspectivalness
7. Transparency
8. Offline activation
9. Representation of intensities
10. "Ultrasmoothness": the homogeneity of simple content
11. Adaptivity

[37] A critical discussion of Aleksander's axioms can be found in Bringsjord (2007). One of the problems raised by Bringsjord is the lack of formalization of the axioms, which is addressed to some extent by the definition given in Section 7.6.2.

These constraints should be met by any fully conscious mental representation and Metzinger (2003) gives detailed descriptions of their neural, functional and computational correlates. Metzinger's constraints are all based on type II correlates of consciousness because their phenomenal, functional and neural aspects can be introspectively and experimentally measured in a system. A brief summary of the constraints now follows.[38]

1. Global availability

Phenomenal information is globally available for deliberately guided attention, cognitive reference and control of action. Our attention can be attracted by or directed to any part or aspect of our conscious mental content and we can react to this content using a multitude of our mental and bodily capacities. Globally available cognitive processing is characterized by flexibility, selectivity of content, and a certain degree of autonomy. One of the functions of global availability is to increase the behavioural flexibility of the system, enabling many different modules to react to the same conscious information, and it also supports goal directed behaviour and the selective control of action. The neural correlates of global availability are not clear at present and form part of the general question about how different areas of the brain are integrated together. One theory is that large scale integration is mediated by the transient formation of dynamic links through neural synchrony over multiple frequency bands (Varela, Lachaux, Rodriguez, and Martinerie 2001) and Tononi and Sporns' (2003) information integration offers a way of measuring the degree of global integration (see Section 2.6.2). In contrast to constraints 2-10, global availability is a functional constraint and it is described by Metzinger as a third-person version of constraint 3.

[38] Metzinger (2003) also gives an account of the phenomenal self model and intentional relation. Whilst these are important aspects of human consciousness, they are less relevant to this thesis and I will only cover Metzinger's constraints here.

2. Window of presence

We experience conscious events in a single now within which a number of things happen simultaneously. In this now events can be represented as having duration or integrated into temporal figures, such as a musical tune. Events within the now have an organisation and vividness that is lacking from events outside it, and the window of presence is embedded in a unidirectional flow of events, which join and leave it. This constraint is supported by short term memory, which keeps phenomenal contents active for some time after the stimuli have disappeared from the receptive field. Functionally this constraint involves the definition of windows of simultaneity, so that all physical events registered within each window are temporally identical. By avoiding the definition of temporal relations within each window the fine structure of physical time becomes transparent to the system[39] and temporal elements can be ordered in a sequence. The neural correlates of this constraint are not well known, although some form of recursion will be necessary to sustain past events. Metzinger cites Pöppel's (1972, 1978, 1985, 1994) theories about how oscillatory phenomena in the brain could provide a rigid internal rhythm, which could generate the elementary integration units.

3. Integration into a coherent global state

Phenomenal events are bound into a global situational context within which we are *one* person living in *one* world. Other situations are not phenomenally possible - the phenomenal world and the phenomenal self are *indivisible*. This constraint also refers to the fact that phenomenal events are densely coupled: as we interact with the world, the states change whilst the apparently seamless integrated character of the overall picture is preserved. One function of global availability is to reduce the ambiguity of the world down to a single compressed representation and a single consciousness is also most appropriate for a single body. Metzinger discusses how this constraint functions as a stable background for imaginative planning that cannot be

[39] See constraint 7.

transcended by the system, so that alternative simulated worlds can be compared with a representation that is tagged as the actual world and the system does not get lost in its own simulations. A global conception of the whole is also necessary in order to understand other objects and events as parts of the whole. The neural correlates of global availability are similar to those for constraint 1 and Metzinger mentions Flohr's (2000) hypothesis about the role of the NMDA receptor complex in achieving large scale integration of ongoing activity. Tononi and Sporns' (2003) information integration measure (see section 2.6.2) is also applicable here.

4. Convolved holism

Phenomenal wholes do not exist as isolated entities, but appear as flexible nested patterns. We experience phenomenal wholes – horse, house, person – that are parts of larger wholes – stables, city, crowd - and can be broken down into smaller wholes that form their parts – legs, body, head, walls, windows, roof, etc. This constraint functions to integrate information together into a unified superstructure and the binding of information at different levels could be achieved using temporal coherence on different frequency bands, as discussed for constraint 1.

5. Dynamicity

Our conscious life emerges from a series of psychological moments that are integrated over time and represented as being in states of presence, duration and change - they are not a succession of isolated events. Whilst constraint 2 refers to the single now that exists at any point in time, this constraint refers to the integration of events over longer periods and to the change in objects over time - something like a temporal version of convolved holism. The functional mechanisms behind dynamicity constitute and represent the transtemporal identity of objects for the system, making information about temporal properties of the world and the system globally available for the control of action, cognition and guided attention. Metzinger does not have any suggestions about the neural correlates of this constraint.

6. Perspectivalness

Phenomenal space is always tied to an individual perspective. We experience things from somewhere and it is impossible to imagine a way of seeing objects that would encompass all of their aspects at once. We are also phenomenologically aware of *being someone*, of being a self in the act of experiencing the world. From a functional point of view, perspectivalness represents the limits of the space that we can causally influence and enables a system to become the object of its own attention and self-directed actions. A phenomenal self is also a necessary precondition for the possession of a strong *epistemic* first-person perspective and for social cognition. The neural correlates of this constraint include the networks involved in the representation of our bodies, the vestibular organ, visceral forms of self-representation and the nuclei involved in the homeostatic regulation of the internal milieu. Damasio (1995, 1999) and the second half of Metzinger (2003) go into the neural correlates of this constraint in detail. A substantial part of Metzinger's work is dedicated to understanding more complex forms of the phenomenal self model, which are not covered in this thesis.

7. Transparency

When we look at the world we do not see a series of neural spikes or streams of data from our optic nerves. We simply see the objects around us and this transparency of our representations is due to the attentional unavailability of earlier processing stages and our inability to introspect the vehicle properties of a representation (we see a red flower, and not the neurons generating the representation of a red flower). This transparency of our mental content forces us into naïve realism about the world: we see the world and not the means by which a representation of the world is constructed by our brains. A world cannot be present without transparency at some

point in the system, and so this constraint forms part of the minimal notion of phenomenal experience.[40]

One of the functions of transparency is to remove complex processing from the system and present the final result in the form of naïve realism, which forces the system to take it seriously because it is no longer 'just a representation'. One of the reasons why the brain is transparent is because it has no senses in it that could present it to itself as an object – it is notably without pain receptors, for example. However, this is not in itself enough for the emergence of transparency, since there is no reason why we should not perceive the incoming data from the retina, for example, as spiking neuron activity instead of light. Transparency is fundamental to phenomenal experience, but unfortunately, as Metzinger notes, "almost nothing is known today about the neural basis of phenomenal transparency." (Metzinger 2003, p. 178).

8. Offline activation

Phenomenal mental content can be active without sensory input, which enables absent objects to be recollected or dreamt and it can also be used in planning. Offline activation also makes the difference between possibility and reality available to the system, supports social cognition by enabling us to simulate other people's first person perspectives, and minimises the risks associated with exploratory activity in the world. Offline phenomenal states are characterised by the fact that they are constructed from sequences of non-stimulus correlated states and this lack of covariance with the environment is an essential feature of their causal role. In the human brain the same neural areas are frequently used for perception and for simulating possible perceptual and motor situations, and brain areas that reactivate perceptual areas, such as the hippocampus, are important for this constraint as well.

[40] It is also discussed in Haikonen (2003).

9. Representation of intensities

Phenomenal experience has a quantitative dimension: colours can vary in brightness, some sounds are louder than others and pain has a variety of different levels. This representation of intensities has the function of guiding the attention of the organism to stimuli of maximum interest and it also reflects the intensity of stimuli in the environment. For example, pain directs attention to a damaged area, and the higher the pain the more our attention is focused on that area. The neural correlates of this constraint are likely to be the firing rates of the neurons and the timing of their spikes.

10. "Ultrasmoothness": the homogeneity of simple content

Unlike the real world, simple phenomenal experiences have a structureless density and are homogenous at all levels of magnification. There is no internal structure, no temporal texture and the graininess of neuron firing is invisible at the phenomenal level. This constraint is linked to transparency because the homogenous atomic nature of simple sensory content could be generating the transparency of sensory awareness. One of the functional properties of homogeneity is that it prevents us from introspectively penetrating into the processing stages underlying the activation of sensory content, which is essential for the production of an untranscendable reality (constraint 3) and for reducing the computational load. At the neural level homogeneity might be related to our brains' limited spatial and temporal resolution: we could only perceive the space between the grains of our neural representations if we had a second, more fine grained, neural mechanism Without this, the data that we get is just the data that we get, and we have no access to the spaces or graininess within it.

11. Adaptivity

The adaptivity constraint states that phenomenal mental content must have come about through natural selection. If we want to understand how consciousness could be *acquired* in the course of

millions of years of biological evolution, we must assume that it possesses a true teleofunctionalist description. Metzinger claims that this third person objective constraint could affect the ability of artificial systems to experience emotions: “artificial systems as known today do not possess genuinely *embodied goal representations*, because they are not ‘evolutionarily grounded’ – neither their hardware nor their software has developed from an evolutionary optimization process.”(Metzinger 2003, p. 199).

One of the ways in which Metzinger argues for this constraint is using Davidson’s Swampman thought experiment:

> Lightning strikes a dead tree in a swamp while Davidson is standing nearby. His body is reduced to its elements, while entirely by coincidence (and out of different molecules) the tree is turned into his physical replica. This replica, the Swampman, is a physical and functional isomorph of Davidson; it moves thinks, talks, and argues just as the original Donald Davidson did. Obviously, it has precisely the same kind of phenomenal experience as Donald Davidson, because phenomenal content locally supervenes on the brain properties of the replica. On the other hand, the intentional contents of Swampman’s mental state are not the same – for instance, it has many false memories about its own history be they as conscious as they may. The active phenomenal representations in Swampman’s brain would be strongly conscious in terms of the whole set of constraints listed so far, but they would not satisfy the adaptivity constraint, because these states would have the wrong kind of history … It would enjoy a rich, differentiated cognitive version of conscious experience tied to a first person perspective, but it would still be consciousness in a weaker sense, because it does not satisfy the adaptivity constraint holding for ordinary biological consciousness. (Metzinger 2003, p. 206).

The relation of consciousness to its present and past environment is useful for understanding the relationship between consciousness and action (see Section 2.7). However, this constraint has a number of serious problems. To begin with, very little of our bodies is the same as when many of our memories were generated, and so everyone has false or partially false memories about their early history. Secondly, evolutionary arguments linking present states of the organism with a past environment tend to rely on simplistic notions of evolution that ignore the complex

feedback loops between the organism and its environment and the constraints of physics and chemistry. Third, many parts of the human body and mind evolved for very different purposes than they presently serve, and so it is senseless to attempt to tie their present meaning to their present or past environment. Finally, I cannot see the benefit in saying that without this constraint the consciousness would be weaker, when the phenomenal experience is said to be the same in both cases.

Within the framework of his constraints, Metzinger defines a minimal notion of conscious experience as follows:

> The phenomenal presence of a world is the activation of a coherent global model of reality (*constraint 3*) within a virtual window of presence (*constraint 2*), both of which are transparent in the sense just introduced (*constraint 7*). The conjunction of satisfied *constraints 2*, *3*, and *7* yields the most elementary form of conscious experience conceivable: the presence of a world, of the content of a world-model that cannot be recognized *as* a model by the system generating it within itself. Neither a rich internal structure nor the complex texture of subjective time or perspectivalness exists at this point. All that such a system would experience would be the presence of one unified world, homogenous and frozen into an internal Now, as it were. (Metzinger 2003, p. 169).

This suggests that a robot implementing constraints 2, 3 and 7 should experience a minimal phenomenal state that is without the differentiation, subjectivity and cognitive capacity of biological consciousness. In general Metzinger stresses that consciousness is a matter of degrees and higher degrees of constraint satisfaction will lead to higher degrees of phenomenality in a system.[41]

[41] A critical discussion of Metzinger's work can be found in Legrand (2005). There is also a certain amount of overlap between Metzinger's constraints and Taylor's (2007) discussion of the components of consciousness.

2.7 Consciousness in Action

> Suppose someone were thus to see through the boorish simplicity of this celebrated concept of "free will" and put it out of his head altogether, I beg of him to carry his "enlightenment" a step further, and also put out his head the contrary of this monstrous conception of "free will": I mean "unfree will," which amounts to a misuse of cause and effect. … The "unfree will" is mythology; in real life it is only a matter of *strong* and *weak* wills.
>
> Nietzsche (1966, p. 29)

> There is no question that consciousness is important for language, for artistic, mathematical, and scientific reasoning, and for communicating information about ourselves to others.
>
> Koch (2004, p. 234)

2.7.1 Introduction

In this chapter I have kept the physical and phenomenal apart and emphasized the search for correlations between them. One consequence of this approach is that it does not make sense to speak about phenomenal objects carrying out physical functions or physical objects carrying out phenomenal functions - although phenomenal states might be correlated with physical functions. At the current stage of consciousness research, it is only possible to talk about the relationship between phenomenal events and phenomenal actions and between physical events and physical actions - with the hope that we will eventually be able to identify systematic correlations between the two. This strict separation means that a phenomenal event, such as the perception of a red object, will never have to be invoked to *explain* a physical event, such as the nerve signals sent to a muscle.[42]

Although the exact mechanisms of physical action are poorly understood, we can conceive how complete descriptions could be carried out at the physical level that explain how networks of neurons could control a human body driving a car or carry out sophisticated

[42] It must be emphasised that this separation of causal chains does not imply any separation of substances between the phenomenal and the physical.

processes of reasoning. Such descriptions would be framed solely within the language of physics and biology and they would be complete without any mention of consciousness or the phenomenal aspects of imagination or emotion. These physical descriptions would completely explain the transformations of the physical world, but they would leave out the phenomenal aspect of reality, which has been argued to be at least as important as the physical. In order to understand the relationship between consciousness and action at the phenomenal level, we need to use concepts such as red, imagination and emotion to explain how we can make decisions that change the stream of experience. This level of explanation is much less well understood and the final part of this chapter will take a brief look at some empirical observations about consciousness and use them to comprehend how we consciously and unconsciously carry out actions.

This section starts with some phenomenological and experimental observations about consciousness, which demonstrate that our naïve preconceptions about the relationship between consciousness and action are often wrong. Section 2.7.3 then offers a tentative classification of the different aspects of conscious and unconscious action, which is used to develop an interpretation of conscious control and conscious will in sections 2.7.4 and 2.7.5. Finally, Section 2.7.6 takes a look at our experience of conscious will.

2.7.2 Observations about Consciousness and Action

This section offers some general observations about consciousness that will be used to develop and support an interpretation of the relationship between consciousness and action. Since this is a subsidiary theme in this thesis, I will not be examining the large amount of research that has been carried out in detail.[43] Instead, the aim of this section is to offer some broad support for the

[43] Some of the other work in this area is covered by Velmans (1991).

interpretations of conscious control and conscious will that are put forward in sections 2.7.4 and 2.7.5.

Almost all conscious mental states[44] can become unconscious, but not vice versa

When we are driving a car we can be conscious of the controls and the road, but we can also process this information unconsciously if we are thinking about other things.[45] However, we cannot make the processes that regulate our heart beat conscious, even if we can exert voluntary control over them with appropriate feedback (Yellin 1986). When we carry out a task unconsciously it is not clear whether its associated mental states are structured in the same way as when the task is carried out consciously.

Unconscious representational mental states can be used to guide action and for limited problem solving

People who suffer from epileptic automatism can perform tasks as complex as diagnosing lung patients without conscious awareness (Cooney 1979). In our everyday lives we execute many complex tasks unconsciously that were learnt when we were carrying them out consciously at an earlier stage in our lives.

Most of the time we are zombies

This point follows from the last. Most of the time we are acting in and responding to the world unconsciously whilst our consciousness is focussed on something completely different. Detailed discussions of the unconscious control of behaviour can be found in Crick and Koch (2003), Koch (2004) and Milner and Goodale (1995).

---

[44] See sections 4.3.2 and 4.3.3 for definitions of a mental state and a representational mental state that apply to both natural and artificial systems.

[45] This point has been disputed by Searle (1992) and by Dennett (1992, p.137), who claims that it is an example of rolling consciousness with swift memory loss. The unconscious processing of complex information is demonstrated by the work on visual masking, which has shown that unconscious words or digits can be processed at the lexical and semantic levels (Kouider and Dehaene 2007).

Unconscious processing is not good at dealing with new situations

When we encounter a problem with a task that we are executing unconsciously, we often turn our attention to the problem and solve it consciously (Baars 1988, Koch 2004, Underwood 1982). For example, suppose that an amateur carpenter is hammering in a nail whilst thinking about his wife. If the nail bends, he will probably stop thinking about his wife and consciously decide either to extract the nail or to straighten it out in situ. This observation should be qualified with the fact that many complex problems can be solved unconsciously. For example, part of my mind is often working on a problem unconsciously and the solution pops into my head spontaneously without any conscious processing. In my case this only happens for fairly abstract problems, but dancers, for example, might be capable of solving complex motor problems unconsciously.

Consciousness and learning

There seems to be a strong link between conscious information processing and the learning of new skills, which generally have to be carried out consciously before they can be initiated and executed automatically. As Koch explains:

> … a zombie agent can be trained to take over the activities that used to require consciousness. That is, a sequence of sensory-motor actions can be stitched together into elaborate motor programs by means of constant repetition. This occurs when you learn how to ride a bicycle, sail a boat, dance to rock-and-roll, climb a steep wall, or play a musical instrument. During the learning phase, you are exquisitely attentive to the way you position and move your hands, fingers, and feet, you closely follow the teacher's instructions, take account of the environment, and so on. With enough practice, however, these skills become effortless, the motion of your body fluid and fast, with no wasted effort. You carry out the action beyond ego, beyond awareness, without giving any thought as to what has to be done next. It just comes naturally.
>
> Koch (2004, p. 235)

Although there is some evidence that we can learn unconsciously as well as consciously - for example Reber's (1967) work on the learning of artificial grammars - the information that is acquired in these experiments is fairly basic (see Shanks (2005) for an overview and criticisms).

Consciousness is not an all or nothing phenomenon

Each individual has periods of full consciousness and periods of barely conscious experience. When I am late for work and waiting for a train I am extremely conscious of the tension inside me, the situation on the platform, the clock and the possibility that I might get fired. As I travel back from work and drift in and out of sleep on the train, I am barely conscious at all. When we are fully conscious we are maximally conscious of the objects at the centre of our attention and barely aware of objects at the periphery. For example, I am currently most conscious of my laptop in front of me and barely conscious of the street scene outside my window. It is likely that minimally conscious brain-damaged patients experience considerably less and more intermittent consciousness than normal people or patients with locked-in syndrome (Laureys et al. 2004). It also seems likely that some animals are phenomenally conscious to a lesser degree than a fully conscious human – see Crook (1983), Baars (2000) and Seth et al. (2005) for discussions of animal consciousness.

The time scale of consciousness

Libet's (1982) experiments measured the duration of neural activation that is necessary for conscious experience. Using electrodes he stimulated the somatosensory cortex of conscious subjects with trains of pulses of different frequency, duration and intensity, and asked the subjects to report whether they felt any sensations. Libet found that there was a minimum intensity below which no sensation was elicited, no matter how long the pulses were sustained. Furthermore, when a stimulus was above this intensity threshold it could only elicit a conscious sensation if it was continued for around 0.5 seconds - pulse trains shorter than this did not enter

conscious awareness. Libet concluded from these experiments that 'neuronal adequacy' for conscious sensation is only achieved after half a second of continuous stimulation of the somatosensory cortex. This suggests that it takes approximately this much time to integrate all of our sensory information into a single coherent conscious experience that can be reported. These timing experiments confirm the observation that we are mostly zombies. On time scales of less than half a second we react and respond to stimuli unconsciously and automatically. Over longer time scales we build up conscious models, which set the framework for our unconscious actions.

Consciousnesses and voluntary action

Libet (1985) carried out an experiment to measure the timing relationship between our consciously experienced decisions and the start of the neural events that lead to voluntary action. In this experiment subjects were asked to hold out their hand in front of them and flex their wrist whenever they wanted to. At the same time the subjects watched a rotating spot of light and were asked to report the location of the spot when they became conscious of their decision to act. Libet also recorded the readiness potential, which is a slow negative shift in electric potential that precedes voluntary motor actions and can be detected using electrodes on the scalp. In these experiments, Libet found that the readiness potential preceded the subjects' experience of a voluntary decision to act, which suggests that the action of flexing the wrist was initiated unconsciously, rather than as the result of a conscious decision.[46]

2.7.3 Conscious and Unconscious Action

These empirical observations about consciousness show that in many circumstances we react automatically to the world or unconsciously initiate actions that we have not consciously decided to do. To clarify the relationship between consciousness and action, the sequence of events that

[46] Libet's timing experiments have generated a great deal of controversy and there is not space to go into the details here. Many criticisms of the voluntary action experiments can be found in the commentary following Libet (1985) and a fairly comprehensive review can be found in Gomes (1998).

constitutes an action has been broken down into the decision that selects the action, the initiation of the action and the sensory feedback from our bodies and the world as the action is carried out. Each of these stages can be carried out consciously or unconsciously, as shown in Table 2.1.

	Conscious	**Unconscious**
Decision	Using imagination and the emotions I reason about the different courses of action and select one. I might imagine eating at different hours of the day and decide that 1.00 is the optimum time for lunch.	Unconscious decisions are either hardwired into our nervous system - for example, reflexes - or reached through unconscious processes that are largely unknown at the present time.
Initiation	The initiation of the action occurs immediately after a conscious decision to start the action. For example, I decide to go to the shop, and then I get up and go to the shop.	The initiation of the action occurs unconsciously. For example, I am lying in bed and suddenly find myself in the act of getting up.
Execution	We are conscious of the action as we carry it out. For example, as I walk down the street, I look around me at the people and cars without entering into a state of imagination or memory.	We are unconscious when the action is being carried out - for example, cases of epileptic automatism or sleep walking.

Table 2.1. Different aspects of conscious and unconscious actions

These conscious and unconscious aspects of an action can be combined in different ways - for example I might consciously decide to eat my lunch at 1.00, and then make a second conscious decision to carry out the action of eating my lunch. Alternatively, I might have made a conscious decision several years ago to eat my lunch at 1.00 whenever possible, and start preparing my lunch automatically when I glance at the clock without a second conscious decision. Other combinations are also possible – for example, actions can be planned, initiated and executed completely unconsciously. The only intuitively implausible combination is the conscious initiation of an unconsciously chosen action, since it is hard to see how we could decide to execute a decision that we are not aware of.

Two combinations from Table 2.1 will now be used to develop models of conscious control and conscious will. With conscious control, the action is decided consciously, initiated consciously (because the action is immediately carried out) and the person is conscious of

sensory feedback from their body and the world as the action is executed. With conscious will, the action is decided consciously, initiated automatically in response to an environmental trigger and executed with the person conscious that they are carrying it out.

2.7.4 Conscious Control

In conscious control actions are decided consciously, initiated immediately and consciously carried out. One of the most plausible models of conscious decision making is offered by Damasio's (1995) somatic marker hypothesis, which gives a good account of the way in which the imagination and emotions work together to reach decisions.[47] Within this framework we make decisions by running through a number of imagined scenarios that trigger bodily feelings associated with them, and eventually settle on the one that feels best. To make this process more efficient there also has to be some mechanism for remembering which scenarios have already been evaluated. This process can be summarised as follows:

1. Generate imaginary scenario that has not been generated before or revisit previous scenario because all other options are exhausted.
2. Evaluate how scenario feels.
3. If scenario feels bad, remember that scenario felt bad and go back to 1.
4. Else if scenario feels right, carry out action immediately.

In *discrete* conscious control we carry out a single action and the conscious imagination of the action *precedes* the action. Since the conscious decision making process is quite slow, this type of conscious control does not happen very often – we believe that conscious control is more common than it is because in many cases the unconscious initiation of an action generates a conscious representation of the action just before it takes place (Libet 1985).[48] However, there

[47] The relationship between the emotions and judgement is discussed by Clore (1992).

[48] See Figure 2.5 for an illustration.

might be circumstances in which we consciously decide to do something and then immediately execute our decision, and a neural model of discrete conscious control has been developed as part of this thesis.

In *continuous* conscious control an action takes place under the guidance of a conscious model that determines the evolution of the action over time. Whilst the decision and the initiation of the action might be automatic, the management of the action is closely linked to consciousness. For example, if my friend asks me what I dreamt last night, then I will probably start my answer automatically without making a conscious decision about whether to reply or not. However, my narration is continuously guided by my conscious memory of the dream, and without this conscious recollection it is hard to see how the dream could be described.[49] Although many of our day to day actions, such as driving or diagnosing lung patients, do not need to be carried out under conscious control, there are numerous daily occasions when we do seem to be consciously controlling continuous actions. Continuous conscious control is likely to be more common than discrete conscious control, but it is often ignored because it is harder to measure experimentally.

2.7.5 Conscious Will

The time scale of discrete conscious control make it implausible that this is the main way in which our conscious decisions influence our actions, and it is much more likely that actions are decided consciously and then initiated unconsciously in response to conscious and unconscious perceptions. In this thesis I will use the term "conscious will" to refer to the process whereby actions are chosen consciously, initiated unconsciously and then consciously carried out.[50] The

[49] Without conscious control, the situation would be a bit like blindsight in which I might be able to guess accurately about the contents of my dream, but would not be able to offer a fluid and natural description.

[50] "Conscious will" could also plausibly be used to refer to actions that are consciously decided, unconsciously initiated and unconsciously executed. Since this does not appear to be a common situation, it has been set aside in this thesis because it would serve only to complicate the discussion.

decision process in conscious will is carried out in the same way as conscious control, but in conscious will, we *remember* the decision and execute it *automatically* in response to environmental or internal triggers (perhaps with the possibility of veto - see Libet (1985, 1999)). For example, at midnight I decide to get up at eight tomorrow morning and set my alarm clock; when the alarm goes off I lie in bed feeling reluctant and then suddenly find myself in the act of getting up. The stages in this model of conscious will can be summarized as follows:

1. Generate imaginary scenario that has not been generated before or revisit previous scenario because all other options are exhausted.
2. Evaluate how scenario feels.
3. If scenario feels bad, remember that scenario felt bad and go back to 1.
4. Else if scenario feels right, remember future action and an associational trigger that will release the action.
5. Continue acting in world.
6. When associational trigger is reached, perform action unconsciously.

This distinction between conscious decisions and automatic execution provides a way out of the problems thrown up by Libet's (1985) timing experiments on the will. Within the framework that I am presenting here, the subject's conscious decision to flex their wrist was taken when they decided to participate in the experiment minutes or hours before the actual action (a fact highlighted by some of the commentators following Libet's (1985) paper). As they randomly flexed their wrist they were not making conscious decisions, but automatically executing a decision that had already been made, and so it is not surprising that the readiness potential preceded the subjects' awareness of their decision to act. To test the timing of *conscious* decisions, the experiment would have to present a number of options to the subjects that would require internal simulation to choose an appropriate action. The timing relationships would then be between the internal modelling of the situation, the activation or simulation of

different body states, the memorization of the conscious decision and the onset of the readiness potential that precedes the execution of the action. It would be very surprising if the readiness potential preceded all of these events, which are likely to take at least one or two seconds. This interpretation of Libet is similar to that put forward by Aleksander et al. (2005).[51]

2.7.6 The Experience of Conscious Will

Our feeling of having willed something could be interpreted as the best evidence that we have for a link between consciousness and action. However, Wegner and Wheatley (1999) and Wegner (2002, 2003, 2004) claim that our experience of willing is actually an *inference* that we make about the relationship between a thought and a subsequent action, and we do not directly experience an actual causal process. This claim is supported by Wegner and Wheatley's (1999) *I Spy* experiment in which two people used a board mounted on top of a mouse to move a cursor to point to one of fifty tiny toy pictures taken from an *I Spy* book. One of the people was a genuine participant who heard words on his or her headset and was cued by music to bring the mouse to a stop. After each trial this participant was asked to rate each stop for the degree of intentionality that they felt when they made it. The second person in the experiment was a confederate pretending to be a participant, who was given instructions to stop the mouse on a particular picture or to allow the participant to stop the cursor wherever he or she liked. On some of the trials the participant heard words that matched the forced stop on a particular picture – for example, they might have heard the word 'swan' prior to the confederate bringing the cursor to rest on the swan.

[51] There was not space in this thesis to examine how this concept of will relates to freedom of the will. The question about the freedom of the will is a complex topic that combines a number of conflicting intuitions (Honderich 1993, Double 1991). However, it is worth pointing out that this model of conscious will is broadly compatible with Hodgson's (2005) basic requirements for any account of free will and it is aligned with compatibilist accounts that balance psychological freedom with metaphysical determinism, such as Gomes (1999) and Clark (1999). It also largely agrees with Kane's (1996) libertarian concept of free will as "*the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes*" (p. 4).

This experiment showed that being cued with a word did not lead the participants to stop more frequently on the associated pictures. However, the participants did claim to have a higher amount of intentionality when they were cued 5 or 1 seconds before being forced to stop on a particular picture, which did not occur when they were cued 30 seconds before or 1 second after the forced stop. In other words, participants claimed that they had intended to stop on a picture associated with a word that they had heard 5 or 1 seconds before, even though they had no choice about where to stop and would not have stopped on the picture if the confederate had not moved the cursor to this position. This suggests that the participant's experience of will depended on an association between the cued words and actions, rather than on any actual causal link between their thoughts and actions. According to Wegner and Wheatley (1999), this experiment shows that the participant's experience of conscious will arises through an inferential process in which they reason about their actions and conclude whether they did them or not.

Three of the most important factors in this inferential process are the priority of the thought before the action, the consistency of the thought with the action and the exclusivity of the thought relative to the action. If we think of an action a short time before it happens, if our thought matches the action, and if no other causes can be put forward to explain the action, then we experience a feeling of intentionality relative to the action: an experience that we willed the action. Wegner (2003) supports his argument with other examples in which there is a disparity between the feeling of conscious will and the actual volition, such as alien hand syndrome, in which the person chooses the actions of the hand, but does not believe himself or herself to have willed them (Geschwind et al., 1995), schizophrenics' attribution of other people's actions to themselves (Daprati et al., 1997), and action projection in which a person performs a voluntary action that they attribute to someone else (Wegner, 2002).[52]

[52] Although Wegner and Wheatley (1999) and Wegner (2004) cite these as examples of wilful action, within the framework presented in Section 2.7.3, these are examples of unconscious decisions initiated unconsciously, which is quite different from the model of conscious will put forward in Section 2.7.5.

Although Wegner (2003) claims that the feeling of conscious will is an *illusion* because it does not reflect the underlying causal mechanisms, this should not be interpreted as the claim that there is *no* link between consciousness and action. Wegner's work convincingly demonstrates that our inference about our causal powers is fallible, but it does not show that it is always incorrect - a point that is made explicitly by Wegner:

> Does all this mean that conscious thought does not cause action? It does not mean this at all. The task of determining the causal relations between conscious representations and actions is a matter of inspection through scientific inquiry, and reliable connections between conscious thought and action can potentially be discerned by this process. The point made here is that the mind's own system for computing these relations provides the person with an experience of conscious will that is no more than a rough-and-ready guide to such causation, one that can be misled by any number of circumstances that render invalid inferences…
>
> (Wegner, 2003, p. 68)

Some people, such as Claxton (1999), have attempted to use arguments similar to Wegner's to virtually eliminate the relationship between conscious will and action. The problem with this position is that a complete break between consciousness and action makes consciousness epiphenomenal and eliminates any sense in which we can claim to *speak* about consciousness - a position that was discussed in detail in Section 2.4.3.

Wegner's account of our experience of conscious will fits in naturally with the models of conscious control and conscious will that were put forward in sections 2.7.4 and 2.7.5. In both conscious control and conscious will, the imagination and emotion that are involved in the decision making process have a completely different phenomenology from the feeling of intention, and it is perfectly plausible that our experience of will is the outcome of an inferential process that takes place after the action has been executed. This is particularly apparent in the model of conscious will, where there might be a delay of years between a conscious decision and the unconscious initiation of the action. In this case it is hardly plausible to claim that we experience the will in operation, and much more likely we find ourselves engaged in an action

and then experience a feeling of conscious will when we remember the earlier decision that led us to act in this way.

Although a connection between consciousness and action is essential to any theory of consciousness that is not epiphenomenal, it is important to remember that our inferences about this link are fallible and get the connection wrong in many cases. This is particularly apparent when the unconscious initiation of an action presents an image of the action in consciousness just before it is carried out, as shown in Figure 2.5.

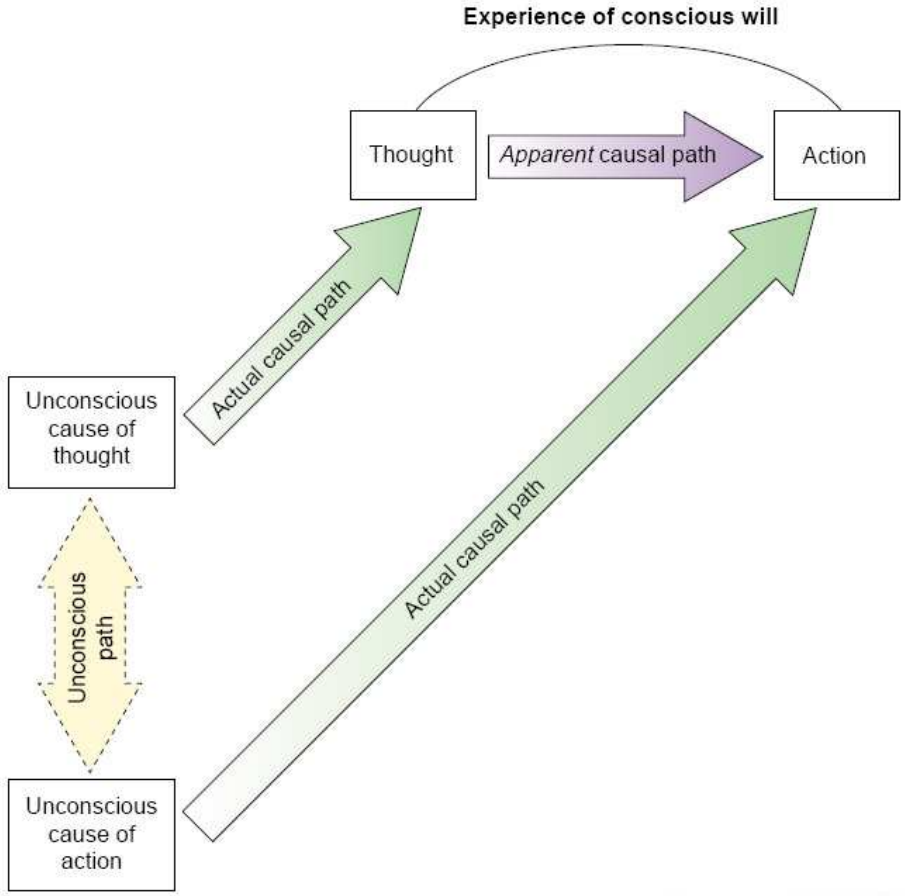

Figure 2.5. Unconscious cause of thought and action. Although the thought appears just before the action, both thought and action have the same unconscious cause. Reproduced from Wegner (2003).

Although the appearance of a thought prior to the action might enable the organism to veto the action (Libet, 1999), Libet's (1985) experiments have shown that the thought often occurs after the action has been unconsciously initiated, when there is only an apparent causal link between

the thought and the action. However, with conscious control and conscious will, it is the timing of the decision about the action that is important and detailed studies are needed to explore the timing relationship between conscious decisions and the conscious or unconscious initiation of actions.

2.8 Conclusions

This chapter has set out an interpretation of consciousness that will be applied in the rest of this thesis. A distinction between the phenomenal and the physical was used to define consciousness and to reject the hard problem of consciousness in favour of the real problem of consciousness, which can only be addressed through work on the correlates of consciousness. This led to the distinction between type I behaviour-neutral correlates of consciousness, which cannot be identified, and type II correlates of consciousness, which can be separated out through their influence on behaviour. This chapter then outlined three type II theories of consciousness and models of conscious control and conscious will.

The approach to consciousness in this chapter will be used to develop a new methodology for describing the consciousness of artificial systems in Chapter 4. The next chapter summarizes some of the previous work that has been carried out in machine consciousness.