4. SYNTHETIC PHENOMENOLOGY[1]

> At present we are completely unequipped to think about the subjective character of experience without relying on the imagination - without taking up the point of view of the experiential subject. This should be regarded as a challenge to form new concepts and devise a new method - an objective phenomenology not dependent on empathy or the imagination. Though presumably it would not capture everything, its goal would be to describe, at least in part, the subjective character of experiences in a form comprehensible to beings incapable of having those experiences.
>
> Nagel (1974, p. 449)

4.1 Introduction

Synthetic phenomenology is a new area of research that has emerged out of work on machine consciousness. The term was first coined by Jordan (1998), who used it to refer to the *synthesizing* of phenomenal states and a second interpretation was suggested by Chrisley and Parthemore (2007), who interpret synthetic phenomenology as the "attempt to use the states, interactions and capacities of an artificial agent for the purpose of specifying the contents of conscious experience." (p. 44). In this usage, an artificial system is being employed to describe the phenomenology of a second system, which could be human, in order to overcome the limitations of natural language. Synthetic phenomenology can also refer to the *determination whether artificial systems are capable of conscious states and the description of these states if they occur*, and it is in this sense that I will be using it in this thesis. This approach to synthetic phenomenology is similar to that put forward by Aleksander and Morton (2007a) and it is close to the philosophical tradition of phenomenology, with the word "synthetic" being added to indicate that it is the phenomenology of artificial systems that is being described. Husserl's (1960) phenomenological project was the description of human consciousness; the synthetic

[1] Earlier versions of parts of this chapter were published as Gamez (2005) and Gamez (2006).

phenomenological project is the description of machine consciousness - a way in which people working on machine consciousness can measure the extent to which they have succeeded in realizing consciousness in a machine.[2]

It is impossible to describe the phenomenology of a system that is not *capable* of consciousness, and so the first challenge faced by synthetic phenomenology is to identify the systems that are capable of phenomenal states. In Chapter 2 it was argued that we do not have a viable metaphysical theory of consciousness, and so we can only tell if a system is conscious by looking at its type I and type II potential correlates of consciousness (PCCs). Setting aside the problem that some correlates of consciousness may be probabilistic and multifactorial, the behaviour-neutrality of type I PCCs means that we cannot identify a list of the necessary and sufficient correlates of consciousness. This prevents us from ever knowing for *certain* whether biological neurons, for example, are necessary for consciousness, or if they are just one of the mechanisms by which consciousness happens to be implemented in human beings. Since it is indeterminable whether silicon-based robotic systems are conscious or not, a major obstacle lies in the way of any attempt to describe the *phenomenology* of such systems.

One approach to this problem is to follow Prinz (2003) and suspend judgement about whether robots are capable of phenomenal states. However, one problem with this approach is that many people have a strong intuition that machines built in a similar way to humans are likely to be phenomenally conscious, and so it may be necessary to take the idea that certain types of machines have conscious experiences seriously. Second, as machine consciousness progresses we are likely to start developing machines that exhibit more complex behaviour and spend a lot of time confused and potentially in pain. This has been somewhat dramatically compared by Metzinger (2003, p. 621) to the development of a race of retarded infants for

[2] Traditional and synthetic phenomenology have different objectives: traditional phenomenology was trying to increase our understanding of the world; synthetic phenomenology is describing the phenomenal states of machines in order to monitor their consciousness and change their behaviour.

experimentation. To address these ethical worries without stifling research a way needs to be found to evaluate the likelihood that a robot is capable of phenomenal states. A third problem with suspending judgement is that as more sophisticated robots emerge, people are inevitably going to attribute more and more consciousness to them. People are already prepared to attribute emotions to robots as simple as Braitenberg's vehicles (Dautenhahn 2007), and a systematic way of evaluating phenomenal states in a system needs to be in place before this becomes a live public issue. The general public is very interested in the question whether something is *really* conscious and it would be helpful if the machine consciousness community could formulate some kind of answer, even if this is based on analogy with human beings. To address these issues and provide a framework within which the more detailed work of synthetic phenomenology can proceed, Section 4.2 outlines a scale that orders machines according to the degree to which their type I PCCs match human type I PCCs.

The next part of this chapter suggests how type II theories of consciousness can be used to generate a description of a machine's phenomenal states. This approach is based around concepts of a mental state and a representational mental state, which are defined in Section 4.3 along with some methods for identifying them in artificial systems. Once we have identified the system's representational and non-representational mental states and made predictions about their association with phenomenal states, we need to find a way of moving from the physical description of the mental states to a description of the system's phenomenology. Section 4.4 outlines some of the reasons why human language is unsuitable for the description of non-human mental states, and puts forward an alternative approach that uses a markup language to combine human and physical descriptions with other information about the system. Finally, the last part of this chapter covers some of the previous work that has been carried out in synthetic phenomenology.

It is worth noting that this approach to synthetic phenomenology makes no assumptions about whether any particular machine is capable of supporting conscious states: robots, stones and human beings all have internal states and all three can be analysed using this approach.[3]

4.2 Ordinal Machine Consciousness (OMC) Scale

> … we may say that measurement, in the broadest sense, is defined as the assignment of numerals or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.
>
> Stevens (1946, p. 677)

4.2.1 Introduction

The discussion of the brain-chip replacement experiment showed that it is impossible to establish whether the behaviour-neutral type I aspects of a system, such as the material it is made from, are correlated with consciousness or not (see Section 2.5.6). The presence of biological neurons *might* be necessary for consciousness or it *might* not, and the introduction to this chapter put forward a number of reasons why we need to make a decision about this, even if we cannot judge with certainty. To address this issue, this section sets out a proposal for an ordinal[4] machine consciousness (OMC) scale that makes predictions about what people would say about the consciousness of non-human systems based solely on their type I PCCs. Type II PCCs do not need to be included in the OMC scale because their correlation with consciousness can be empirically assessed.

---

[3] Stones have few of the human type I PCCs, but it is an open and empirical question whether any of the type II theories of consciousness would predict that they have phenomenal states.

[4] See Stevens (1946) for the difference between nominal, ordinal, interval and ratio scales. It was decided to make the scale ordinal because it was anticipated that it would only be possible to measure people's assessment about whether one system is more or less conscious than another. In the future it may be possible to develop an interval or ratio scale.

The OMC scale is a model of our subjective judgement about the consciousness of artificial systems, and although it might initially seem counterintuitive to use a numerical scale to rank our judgements about the consciousness of systems, there has been a lot of psychophysical work on the measurement of other subjective qualities, such as brightness, loudness, the hardness of minerals, beauty or the desirability of automobiles (Baird and Noma 1978). The OMC scale is a logical extension of this work that attempts to predict the degree to which a system's type I PCCs are judged by us to be relevant to consciousness. As Stevens (1946) points out, measurement scales are possible when there is an isomorphism between certain properties of objects and the properties of numerical series, and this isomorphism enables the series to model relevant aspects of the empirical world. In this thesis the OMC scale is a proposed ordering of systems that is predicted to match people's judgments about systems' consciousness based on their type I PCCs.

This project did not have the resources to base the OMC scale on empirical measurements of people's judgements about the link between type I PCCs and consciousness, and so the current version is put forward as a model of how people would make this type of judgement. This use of models in psychophysics is summarised by Baird and Noma (1978):

> In brief, a psychophysical theory is a set of statements (assumptions) that describes how an organism processes stimulus information under carefully specified conditions. The assumptions usually concern hypothetical processes that are difficult or impossible to observe directly. Once these assumptions are made explicit, however, formal models can be devised. The validity of the theory can be tested by comparing observations against the predictions of the model. In other words, a theory represents a set of "reasonable" guesses about exactly *how* a person behaves as a measuring instrument when asked to judge properties of stimuli.
>
> Detailed predictions of what a person will actually *do* in an experiment are based on models especially designed to test one or more theories. Although in recent years the terms "model" and "theory" have often been used interchangeably, a model is thought to be a concrete synthesis of the assumptions of a theory. This synthesis specifies the interrelationships among the postulated primitives of the theory. Often these statements are in the form of mathematical formulas, computer programs, or logical truisms. In this way they are both

> more specific and yet more general than the theory giving rise to them – more specific in that the theory, through its models, is now amenable to laboratory test, and more general in that an abstract model may be used to quantify theories in many areas of study.
>
> Baird and Noma (1978, pp. 2-3)

The current OMC model enables predictions to be made about what people would say about the consciousness of non-human systems based solely on their type I PCCs, and it is used in this thesis to demonstrate a new approach to synthetic phenomenology.

This description of the OMC scale starts with an overview of the systems that are covered by it. After explaining the factors and the way in which they are combined, some examples are given to illustrate how it works. How this model might be validated and improved using real data is discussed in Section 4.2.7.

4.2.2 Systems Covered by the OMC Scale

In order to focus on the *behaviour-neutral* aspects of each system, the systems ranked by the OMC scale need to have their behaviour held constant in some way, which can be done by specifying that all of the systems ranked by the OMC scale must conform to the behaviour set of a system that is generally acknowledged to be conscious. This ensures that a system's type I PCCs are the only factors that affect its position on the scale.

Since humans are our paradigmatic conscious systems, the functions of the human brain can be used to specify a set of behaviours that systems on the OMC scale would have to match.[5] This notion of approximating the functions of the human brain could be defined using Harnad's (1994) extended T3 version of the Turing test. A machine that could pass this test would be able to control a human or artificial body in a way that was functionally indistinguishable from a

[5] This way of specifying the behaviour of systems covered by the OMC scale sets aside the whole question of the body. In theory a computer could approximate the behaviour of the human brain without needing a body at all. However, such a system would be almost impossible to develop and there might be a critical link between the body and consciousness that would be missed by a purely brain-based approach – see, for example, Damasio (1999) for more on the link between the body and consciousness.

human for 70 years or more. Such a system could hold down a job, create works of art and have relationships with other human beings. Machines that were in a persistent vegetative state or interned in an asylum for strange behaviour would not be considered functionally identical to a human being according to this measure.

Whilst the T3 version of the Turing Test defines the behaviour of a paradigmatically conscious system, it has the disadvantage that our current machines are very far from passing it – if T3 was used as the definition of behaviour, then the OMC scale could only be applied to our current systems by treating them *as if* they had developed to the point at which they were capable of passing it. A second way of defining the behaviour set of a conscious system would be to look at humans who exhibit far less complex behaviour. For example, since we attribute consciousness to locked-in patients who are limited to the movement of a single eyelid,[6] the symbolic T2 version of the Turing Test might be enough for behaviour neutrality. Many other brain damaged people are also examples of systems that are attributed consciousness, but might not be able to pass the T3 Turing Test, and their behaviour could also be used as a common standard for systems ranked by the scale.

A third possibility is that our knowledge about animal consciousness might develop to the point at which an animal's brain could be used to specify a set of conscious behaviours. Systems that conformed to this behaviour set would have to approximate the behaviour of the brains of animals that are known to be conscious by controlling a body similar to the animal's for the lifetime of the animal (systems that imitated one or two simple behaviours, such as flying or swimming, would be attributed less consciousness than the animal on behavioural grounds). Whichever definition of behaviour is used, it is not the behaviour per se that is important, but the fact that it approximates the behaviour of a system that is agreed to be conscious, so that only the type I attributes of the system affect our judgement about its potential for conscious states.

---

[6] For example, see Baubey (2002).

4.2.3 OMC Factors and Weights

The scale is built in a modular fashion so that factors can easily be added, removed or adjusted to match data gathered by psychophysical experiments. Each system is assigned a weight, ω, for each of its type I PCCs, and these weights are combined according to the rules set out in Section 4.2.4 to generate the scale. The working assumption behind the OMC scale is that people's attribution of consciousness to a system is largely based on similarities between the system and the human brain, and so it was decided to set ω to 1.0 when the system was the same as the human brain for a particular PCC. When the system deviates from the human brain on a particular factor, it is given a weight less than 1.0, and to preserve the modularity of the scale the minimum value of ω was limited to 0.1. So, for example, Table 4.1 shows how the system is assigned a weight of 1.0 if it runs at approximately the same speed as the human brain, a weight of 0.55 if it runs ten times faster or slower than the human brain, and a weight of 0.1 if it runs over a hundred times faster or slower than the human brain.

The current version of the OMC scale only covers a very small selection of the type I PCCs that have turned up in discussions of consciousness in artificial systems by Block (1978), Searle (1980), Kent (1981) and others, and the assignment of weights has been done in a subjective and somewhat arbitrary fashion. In the future it is hoped that psychophysical methods could be used to test and improve the scale, and some suggestions about how this could be done are given in Section 4.2.7. An outline of the factors that I have selected for version 0.6 of the OMC scale now follows.

Rate

Machines can operate much faster or slower than the human brain and we are more likely to attribute consciousness to a machine that runs at approximately the same speed. If we were forced to say whether the economy of Bolivia or the Earth's crust is more likely to be conscious,

we would probably choose the economy of Bolivia. This is not because it is more complex or has more states, but because its states change more rapidly.

	Rate	***ω***
R1	Approximately the same speed as the human brain	1.0
R2	Ten times faster or slower than the human brain	0.55
R3	Over a hundred times faster or slower than the human brain	0.1

Table 4.1. Rate factors

Size

We are more likely to attribute consciousness to a system that fits inside a person's head, than to a system that is the size of the population of China.

	Size	***ω***
S1	Approximately the same size as the human brain	1.0
S2	A thousand times larger or smaller than the human brain	0.55
S3	More than a million times larger or smaller than the human brain	0.1

Table 4.2. Size factors

Function Implementation

There are a wide variety of ways in which the functions of a system can be implemented, some of which are closer to human biology than others. This factor weights machines according to the degree to which the implementation of their functions matches that of the human brain. I have gone down to the atomic level to take account of claims by Hameroff and Penrose (1996) that consciousness depends on quantum functions.

This factor is complicated by the fact that neurons can be used to implement functions in a biological and non-biological way. For example, a function can be implemented by a neural network trained by back propagation or by a more biological structure of neurons. Since neurons can themselves be simulated using neurons there is potential for infinite self-recursion, which I

have limited by introducing a restriction on the number of levels. To keep things simple I have also set aside the possibility that glia play an information-processing role (Haydon 2000).

The way in which these three tables are combined is fairly self-evident. If the functions are implemented by a biological structure of neurons (F1 in Table 4.3), then the way in which the function of the neurons is implemented has to be specified as well (for example, FN1 in Table 4.4). No further levels are required if a system's functions are implemented in a non-neural way (F3 in Table 4.3).

Since all computer simulations are physical systems consisting of a certain combination of molecules, atoms and ions, the purpose of the function implementation factor is not to determine whether the system is simulated or not, but to capture the level of detail at which the system's functions match the functions of the human brain. For example, if a system's functions are carried out using a large lookup table, then this might be stored as voltages in the computer's RAM, which is a physical thing, but we are more likely to attribute consciousness to a system that implements the brain's functions using simulated neural networks. We attribute maximum consciousness to systems that match the human brain all the way down to the level of molecules, atoms and ions and implement the molecules, atoms and ions using real biological molecules, atoms and ions.

	Function implementation	ω
F1	Produced by a biological structure of neurons	1.0
F2	Produced by a non-biological structure of neurons	0.55
F3	Produced using mathematical algorithms, computer code or some other method	0.1

Table 4.3. Function implementation

	Function of neurons	ω
FN1	Produced by a biological structure of molecules, atoms and ions	1.0
FN2	Produced by a non-biological structure of molecules, atoms and ions (silicon chemistry, for example)	0.7
FN3	Produced by a non-biological structure of neurons	0.4
FN4	Produced using mathematical algorithms, computer code or some other method	0.1

Table 4.4. Neuron implementation

	Function of molecules, atoms and ions	ω
FMAI1	Produced by real subatomic phenomena, such as protons, neutrons and electrons	1.0
FMAI2	Produced by a non-biological structure of neurons	0.55
FMAI3	Produced using mathematical algorithms, computer code or some other method	0.1

Table 4.5. Molecule, atom and ion implementation

Time Slicing

The processing of functions can be carried out in parallel with all of them operating simultaneously on dedicated hardware. On the other hand a single processor can emulate the parallel operation of many functions by time-slicing. This scale follows Kent (1981) in ranking time-sliced systems, which approximate the time complexity of the brain, as being less likely to be phenomenally conscious than systems with the same moment to moment space complexity as the brain.

	Time slicing	ω
TS1	All functions are dynamically changing and co-present at any point in time	1.0
TS2	Some functions are dynamically changing and co-present at any point in time	0.55
TS3	A single function is dynamically changing and present at any point in time	0.1

Table 4.6. Time slicing

Analogue / Digital

Although spiking neurons have a digital aspect, the brain includes many analogue processes that may be more faithfully captured by an analogue system.[7]

	Analogue / digital	ω
AD1	Analogue system	1.0
AD2	Mixture of analogue and digital	0.55
AD3	Digital system	0.1

Table 4.7. Analogue / digital systems

4.2.4 Putting it All Together

To obtain the final OMC scale, a complete list of all the possible machines is extracted from the factor tables. The weights applicable to each are then multiplied together to give total weightings for each of the possible machines, which are used to situate them on an ordinal scale. Since many of the machines have the same total weighting, this scale is much shorter than the number of possible combinations. A couple of extra rules were also introduced for the combination of factors:

1. Since neurons can be used to simulate the behaviour of neurons or the molecules, atoms and ions that neurons are composed of, the function implementation is potentially infinitely self-recursive. To prevent this I have stipulated that if non-biological structures of neurons are used to implement the functions of neurons or the functions of molecules, atoms and ions, then the neurons that are used for this cannot themselves have their functions implemented using non-biological structures of neurons.

[7] See Roberts and Bush (1981) for examples of analogue processing in the brain, and Shu et al. (2006) for experimental work on the hybrid analogue and digital nature of spike transmission.

2. When machines have less levels of functional implementation than the brain some kind of penalty needs to be imposed on machines that deviate from the human structure – for example, when functions are implemented by a lookup table instead of using biologically structured neurons implemented with molecules, atoms and ions. In the present scale there are three levels of functional implementation and I have used 0.1 as the weighting for each missing level.

The position, omc_{pos}, of an actual machine on a scale with omc_{tot} possible positions is found by calculating its total weighting, and looking for this value in the complete list of possible machines. To facilitate some kind of comparison between different versions of the scale, omc_{pos} is normalised to a value between 0 and 1 to give the final OMC rating, omc_{rat}, using Equation 4.1:

$$omc_{rat} = 1 + \frac{1 - omc_{pos}}{omc_{tot}}, \tag{4.1}$$

which gives a rating of 1 for human brains and a rating close to zero for the last system on the list. The closer this OMC rating is to 1 the more human-like are its type I potential correlates of consciousness. Citations of a system's OMC rating should include the version of the scale, since it is anticipated that it will evolve over time.

When all of a machine's functions are implemented in the same way, this scale provides the OMC rating for the complete system. However, some machines include components that have different OMC ratings – for example, a human brain with a silicon hippocampus. In this case, the OMC rating should be calculated for each part of the system.

The current version of the OMC scale starts with human beings and finishes with digital single-processor simulations based on non-biological principles that are much larger or smaller than the human brain and process at a much slower or faster rate. There is not space in this

chapter to list the OMC ratings of all of the possible machines (the complete list has several hundred thousand combinations), and so I have integrated everything together on a webpage,[8] which can be used to calculate the position of a machine on the scale and its OMC rating. Some examples are given in the next section.

4.2.5 Examples

To illustrate the operation of the OMC scale, this section gives some examples of the position and rating of different types of system. At present none of these are even close to reproducing all of the functions of the human brain for 70 years, and so this evaluation would only apply to them after they have developed to the point at which they can pass the T3 version of the Turing test or could match one of the less complex behaviour sets discussed in Section 4.2.2.

Neurally Controlled Animat

This is a system developed by DeMarse et al. (2001) that uses biological neurons to control a computer-generated animal in a virtual world. The biological neurons start off in a disorganised state and then self-assemble in response to stimulation from their environment. Since the organisation of the neurons is not determined by the many factors present in embryological development, this system produces the functions of the whole brain from a non-biological structure of neurons. The factors are thus: R1, S1 F2, FN1, FMAI1, TS1 and AD1, giving a total weighting of 0.55, an OMC position of 3 out of 192 and an OMC rating of. 0.990.

Lucy

Lucy is a robot developed by Grand (2003) that is controlled by a multi-processor simulation of neurons arranged in a biological structure. The factors are thus R1, S1, F1, FN4, TS2 and AD3, giving a total weighting of 5.5×10^{-3}. This needs to be multiplied by 0.1 to compensate for the

[8] The OMC scale webpage is included in the Supporting Materials along with the code that was used to generate it.

fact that Lucy's functions are not implemented at the level of molecules, atoms and ions, making the total weighting 5.5 x 10^{-4}. This gives Lucy an OMC position of 96 out of 192 and an OMC rating of 0.505.

IDA

IDA is a naval dispatching system created by Franklin et. al. (2003) that is based on Baars' (1988) global workspace model of consciousness.[9] The solutions used to implement this system are all non-biological, and so the factors are R1, S1, F3, TS2 and AD3. This gives a total weighting of 5.5 x 10^{-3}, but since the functions are not implemented at the level of neurons or molecules, atoms and ions, this needs to be multiplied by 2 x 0.1, to give a total weighting of 5.5 x 10^{-5}, which results in an OMC position of 146 out of 192 and an OMC rating of.0.245.

The Population of China

This is a thought experiment suggested by Block (1978) in which the functions of a human brain are carried out by the population of China interconnected by two-way radios and satellites. The population of China is approximately 1.3 billion and so this 'machine' is very much larger than the human brain. It is also likely to work at a much slower rate. This 'machine' contains both biological neurons and other hardware, and so the OMC rating has to be calculated separately for the different parts of the system.

The biological parts are implemented using a non-biological structure of neurons whose function is in turn implemented using a biological structure of molecules, atoms and ions, giving the factors R3, S3, F2, FN1, FMAI1, TS1 and AD1, which works out as a total weighting of 5.5 x 10^{-3}, an OMC position of 50 out of 192 and an OMC rating of 0.745. The rest of the system, consisting of the two-way radios, satellites, etc., has factors R3, S3, F3, TS2 and AD3, which gives a total weighting of 5.5 x 10^{-5} that needs to be multiplied by 2 x 0.1 to compensate for the

[9] IDA is covered in more detail in Section 3.5.6.

missing levels of functional implementation. This gives a total weighting of 5.5×10^{-7}, an OMC position of 188 out of 192 and an OMC rating of 0.02604.

4.2.6 OMC Scale Discussion

It is possible that consciousness decreases gradually as we move away from the human machine, or there may be a cut off point at which it simply vanishes. For example, there might be less consciousness when neurons are simulated using time slicing, or no phenomenal states at all when this is used in a system. We cannot empirically establish whether consciousness cuts off or not, but this does lead to two different interpretations of the OMC scale. If consciousness cuts off abruptly, then the OMC rating can be interpreted as our evaluation of the likelihood that consciousness is present in a machine that is built in a particular way. On the other hand, if consciousness decreases gradually as the factors become less human, then the OMC scale ranks machines according to our judgement about their degree of consciousness.

This is an extremely anthropocentric scale in which the great chain of machines is a kind of fall from grace from perfectly conscious man. This is an epistemological necessity – we only know for sure that we are conscious – but it is quite possible, although empirically undeterminable, that robots at the far end of the scale are more conscious than ourselves.[10]

The final OMC rating expresses an *ordering* of machines according to our subjective judgement about the relationship between their type I attributes and consciousness, so a system with an OMC rating of 0.8 is judged to be more conscious (or more likely to be conscious) than a system with a rating of 0.6. However, because successive intervals on the scale are not necessarily equal, it is incorrect to say that a system with an OMC rating of 0.8 is judged to be twice as conscious (or twice as likely to be conscious) as one with an OMC rating of 0.4.

[10] If we judged machines to more conscious than humans, then we could assign them an OMC rating greater than 1.

This scale only covers type I PCCs that cannot be empirically tested and affect our *a priori* judgement about a machine's potential for phenomenal states. For this reason the scale excludes many of the factors that have been put forward as PCCs, such as synchronization between neurons, a global workspace architecture, a model of the self, and so on. The correlation between these factors and consciousness can be assessed empirically and it is hoped that we will eventually come up with a list of type II correlates that are *necessary* in a conscious system. Any machine that lacked one of these necessary conditions would not be deemed to be conscious, regardless of its position on the OMC scale. However, a list of type II correlates will never be *sufficient* for the prediction of consciousness because one or a number of type I correlates might be necessary as well. Final judgements about a system's potential phenomenal states should combine the OMC scale's *a priori* evaluation about its capability for consciousness with an empirical assessment using a type II theory.

Many type I PCCs, such as the size of a system or its material, do not substantially change from moment to moment and the OMC rating can be calculated once for the entire lifetime of the system. When a system's type I PCCs change over time, its OMC rating may have to be recalculated each time its phenomenology is described.

It must be emphasized that a high OMC rating does not indicate that a system is actually conscious – for example, living humans have an OMC rating of 1.0 and yet they are only conscious for up to 16 hours per day. A high OMC rating only indicates that the system is judged to completely or approximately match humans on all of the type I PCCs that are judged to be relevant to consciousness; this OMC rating has to be combined with a type II theory to make predictions about whether the system is actually conscious at any point in time.

Although the current OMC scale has many limitations, the most important question is not whether this particular version makes sense, but whether the problems raised by the brain-chip replacement experiment require us to use this type of scale. If the type I/ II distinction outlined in

sections 2.5.7 and 2.5.8 is valid, then something like this OMC scale is likely to become an essential tool in machine consciousness research, and the question becomes which is the best possible scale for this purpose. On the other hand, if it can be shown that the distinction between type I and type II PCCs is mistaken, then there is no need for the OMC scale at all.

Finally, as technology and culture develops, people's intuitions will change, and a revised version of the scale will have to be produced every few years. As we get closer to achieving machine consciousness, this scale might eventually become superfluous: when we talk to robots every day, work with robots that display conscious behaviour and perhaps even marry robots, we might cease to worry about whether they *really* have phenomenal states, just as we rarely see other people as automatons.

4.2.7 Future Development of the Scale

The current version of the scale is a model that predicts the subjective judgements that people will make about the link between type I PCCs and consciousness. In the future this model needs to be tested on real data by surveying people's judgments about the consciousness of systems with different type I PCCs. One way in which this could be done would be to show people short films of a humanoid body controlled by brains, computers and other artefacts with different type I PCCs, and ask participants to order them according to their potential for consciousness. To begin with each factor could be varied individually and people could be asked about whether system A was more or less conscious than system B to get an ordinal scale for each factor. The factors could also be varied in combination and factors would have to be tested that were not on the current version of the scale. One potential problem with carrying out these experiments on the general public is that their judgements are likely to be based on an amalgam of what they have seen in science fiction films and read in the media - although it could be argued that these popular representations reflect our underlying beliefs as well as alter them. Expert opinion has

the opposite problem that it can be too linked to particular theories, and so it would be best to obtain sample data from both groups.

The first application of this data would be to revise the lists of factors that are used to construct the scale. For example, if people systematically believed that green objects were less likely to be conscious than red, then colour could be added as a factor. The weightings within each list of factors would also have to be fine tuned, and it is anticipated that many of them will approximate Fechner's logarithmic law, which is given in Equation 4.2:[11]

$$\varphi = k \log \phi , \tag{4.2}$$

where φ is the sensation magnitude, k is a constant and ϕ is the intensity of the stimulus in units above an absolute threshold.[12]

A second application of this data would be to look at different ways of integrating the factor scales. It might turn out that the current approach makes good predictions about the data, but if it is not a good fit, then it would be worth experimenting with different methods of combining the weights. One possibility would be to add the weights, and it might be necessary to weight the factors to accommodate the fact that people attribute different importance to different PCCs. Another option would be to use Shepard-Kruskal multidimensional scaling to combine the different ordinal rankings into a single Euclidean space and use the normalized distance from the most conscious system as the OMC rating (Shepard 1962a,b, Kruskal 1964).

Another direction of future work would be to use psychophysical methods to establish thresholds for the subjective assignment of consciousness and it might be possible to obtain an interval scale by including equisection or category scaling in the survey of people's judgements – see Gescheider (1997) for an overview of these methods. To obtain a more mathematically sophisticated scale, nonmetric scaling could be used to convert the ordinal scale into an interval

[11] More details about Fechner's law can be found in Gescheider (1997).

[12] This logarithmic relationship has already been incorporated into the size and rate factors of the current scale.

scale (Shepard 1966).[13] A ratio scale would be more difficult to achieve since it depends on an absolute zero, which might be difficult to agree upon in consciousness research – for example, some people might be prepared to assign consciousness to a vacuum, thinking, perhaps, that it could contain a spiritual non-material substance.

4.3 Mental and Representational States

4.3.1 Human and Synthetic Phenomenology

To clarify the relationship between synthetic and traditional phenomenology,[14] I will give a couple of examples from Husserl's phenomenology of time consciousness and Merleau-Ponty's phenomenology of the body and the senses:

> In the "perception of a melody," we distinguish the tone *given now*, which we term the "perceived," from those which *have gone by*, which we say are "not perceived." On the other hand, we call the *whole melody* one that is *perceived*, although only the now-point actually is. We follow this procedure because not only is the extension of the melody given point for point in an extension of the act of perception but also the unity of retentional consciousness still "holds" the expired tones themselves in consciousness and continuously establishes the unity of consciousness with reference to the homogeneous temporal Object, i.e., the melody. An Objectivity such as a melody cannot itself be originarily given except as "perceived" in this form.
>
> Husserl (1964, p. 60)

> Already in the "touch" we have just found three distinct experiences which subtend one another, three dimensions which overlap but are distinct: a touching of the sleek and of the rough, a touching of the things – a

[13] This would only work if the rank ordering of the intervals exhibited certain properties, such as weak transitivity of the ordering and monotonicity.

[14] I am using "traditional phenomenology" to refer to the phenomenological tradition that started with Husserl and Brentano and attempted to describe human experience. I have left Dennett's (1992) heterophenomenology out of this discussion, which is a third person method for gathering the phenomenological descriptions of subjects: "It involves extracting and purifying texts from (apparently) speaking *subjects*, and using those texts to generate a theorist's fiction, the subject's *heterophenomenological world*. This fictional world is populated with all the images, events, sounds, smells, hunches, presentiments, and feelings that the subject (apparently) sincerely believes to exist in his or her consciousness. Maximally extended, it is a neutral portrayal of exactly *what it is like to be* that subject – in the subject's own terms, given the best interpretation we can muster." (Dennett 1992, p. 98).

> passive sentiment of the body and of its space – and finally a veritable touching of the touch, when my right hand touches my left hand while it is palpitating the things, where "the touching subject" passes over into the rank of the touched, descends into the things, such that the touch is formed in the midst of the world and as it were in the things. Between the massive sentiment I have of the sack in which I am enclosed, and the control from without that my hand exercises over my hand there is as much difference as between the movements of my eyes and the changes they produce in the visible. And as, conversely, every experience of the visible has always been given to me within the context of the movements of the look, the visible spectacle belongs to the touch neither more nor less than do the "tactile qualities." We must habituate ourselves to think that every visible is cut out in the tangible, every tactile being in some manner promised to visibility, and that there is encroachment, infringement, not only between the touched and the touching, but also between the tangible and the visible, which is encrusted in it, as conversely, the tangible itself is not a nothingness of visibility, is not without visual existence. Since the same body sees and touches, visible and tangible belong to the same world. It is a marvel too little noticed that every movement of my eyes – even more, every displacement of my body – has its place in the same visible universe that I itemize and explore with them, as, conversely, every vision takes place somewhere in the tactile space. There is double and crossed situating of the visible in the tangible and of the tangible in the visible; the two maps are complete, and yet they do not merge into one. The two parts are total parts and yet are not superposable.
>
> Merleau-Ponty (1995, p. 134)

The first thing to note about these examples, is that they are based on *first-person* introspection, in which the phenomenologist examines his or her experiences and writes down a description in human language. At the current stage of development, artificial systems are fairly rudimentary and incapable of describing any phenomenal states that they might have. This forces synthetic phenomenology to start with *third-person* objective measurements, which can be combined with type II theories of consciousness to make predictions about the system's phenomenal states.[15] These objective measurements are generally carried out on a subset of the system, such as its artificial neural networks or the code implementing a global workspace

[15] This approach is similar to neurophenomenology (see Section 4.5), which attempts to make predictions about people's first person phenomenology on the basis of objective brain measurements.

architecture, which is analyzed *as if* it were a mind capable of representations and phenomenal consciousness. To clarify this transition from the physical to the mental, Section 4.3.2 sets out a definition of a mental state, which applies at the physical level and can be used to interpret artificial as well as natural systems.

A second feature of traditional phenomenology is that it is based on objective features of the world that can be physically measured and experienced by more than one person - for example, the sound waves in Husserl's melody can be recorded with scientific instruments and Merleau-Ponty's touching and touched hands are physical as well as phenomenal objects.[16] This suggests that phenomenal experiences can be interpreted as *representations* of objects that appear in other peoples' streams of experience, and these objects can be probed in a variety of different ways. This interpretation of phenomenal experiences as representations is very useful when we are describing the phenomenology of artificial systems, with the difference that we have to find a way of identifying representations from a third person perspective. To address this problem, a definition of a representational mental state is given in Section 4.3.3, and Section 4.3.4 discusses some of the ways in which representational mental states can be identified in artificial systems.

A third observation about these quotations is that Husserl and Merleau-Ponty are describing *conscious* mental states and do not consider the many unconscious mental states that are in their minds. Section 4.3.5 explains how a theory of consciousness (based on type II correlates of consciousness) can be used to make predictions about the association between mental states and phenomenal states. Finally, Husserl and Merleau-Ponty are describing states that are *integrated* together into a *single* consciousness, and this question about the relationships between mental states is briefly covered in Section 4.3.6. The outcome of this process is a set of physical descriptions of representational and non-representational mental states that are

[16] This notion of a physical world would be interpreted with caution by traditional phenomenology, which often claims that the phenomenal is more primordial than the physical – see Husserl (1960).

associated with phenomenal states and Section 4.4 suggests how these can be turned into a full phenomenological description.

4.3.2 Mental States

Homeric man believed that the seat of human consciousness was in the heart and lungs (Onians 1973) and over thousands of years people have gradually come to associate human consciousness with human brains. Although many philosophers would argue that mental states are conceptually distinct from physical states, the increase in our knowledge about the brain, and the constant reduction of our mental functions to brain functions has led Churchland (1989) to suggest that the term "mental state" will eventually become redundant and our use of mental terminology will be superseded by descriptions in terms of states of the brain – a position known as eliminative materialism.[17] In the human case, this may eventually occur because a clear link has been established between mental states and the brain. However, synthetic phenomenology is analysing systems without biological brains and it is far from clear which part of the system is the right place to look for phenomenal states. Within this context we need the concept of a mental state to specify the part of the system (or subset of the system's states) that we are analysing for consciousness. For this purpose I will use "mental state" according to the following definition:[18]

> *A mental state is a state of the part of the system that is being analysed for consciousness.* (4.1)

When people analyze humans for consciousness they generally focus on the brain and human mental states are usually taken to be states of human brains. Within the human brain,

[17] Rorty's (1979, p. 71) thought experiment in which Antipodeans use brain descriptions instead of mental terms to express their inner states is an example.

[18] This differs from Metzinger's (2003) definition, which links mental states to phenomenal accessibility.

work on the neural correlates of consciousness has shown that neural activity is important for consciousness, and so mental states can be defined in terms of the neuron's firing rates, the timing of their spikes or other properties of the neurons. However, it is also possible to analyse other parts of the human body for phenomenal states. For example, we can examine the liver or blood for consciousness, and when we do this, different states of the liver or blood become mental states according to this definition.[19]

In artificial systems a mental state can be a pattern of firing activity in simulated neurons or a sequence of 1s and 0s in the computer's RAM - for example, mental states could be monitored in Franklin's IDA (see Section 3.5.6) by using a debugger to measure changes in the memory. Different ways of defining a system's mental states may lead to different predictions about its phenomenology, which can be tested by monitoring its behaviour.

In this thesis mental states will be described at the physical level, either in physical terms or in terms that can easily be mapped down to physical descriptions without any loss of meaning or information. These states of the physical world can be identified within our phenomenal world by making phenomenal measurements of some region of the physical system and defining any states that take place within this region as mental.[20]

4.3.3 Representational Mental States

Some mental states are systematically related to features of the world. "Representation" is a natural way of describing this relationship, but since it is a controversial word, I will use it in a very restricted way in this thesis according to the following definition:

[19] See Holcombe and Paton (1998) and Paton et al. (2003) for a discussion of the computations carried out by the liver and other tissues.

[20] Mental states can also be a particular class of states that are not physically distinct – for example, neurons firing at 40 Hz could be classified as mental states.

A representational mental state is functionally or effectively connected to other mental states or to the data that is entering or leaving the system. (4.2)

Within a neural network functional connectivity is defined by Sporns et. al. (2004) as a statistical relationship between two neurons that may or may not be due to a causal relationship between them - for example, two neurons that share mutual information are said to be functionally connected. Effective connectivity describes the set of causal effects that one neuron has on another and it can be inferred experimentally by perturbing one part of the system or by observing the temporal ordering of events. Whilst Sporns et al. (2004) apply their definition of functional and effective connectivity to neuronal units, in this thesis it will be applied to all mental states and to the data that is entering and leaving the system.[21] It is also important that a representational mental state is distinguished from the state that is being represented - or it would no longer be a representation, but the thing itself. Some of the ways in which representational mental states can be identified are discussed in the next section.

Representational mental states do not necessarily *resemble* what they represent, although this is not excluded by Definition 4.2.[22] They are also different from depiction in Aleksander's (2005) sense. Depictions are mental states that are systematically related to both motor and visual information, whereas the definition of representation that I am using here is much broader and includes all mental states that are functionally or effectively connected to other mental states or to features of the system's incoming and outgoing data. The relationship between language and representation is not covered by this definition, although it may be possible to analyse language using this approach.

---

[21] The outgoing data is included to cover cases in which the system is representing its own motor activity.

[22] The question about representation and resemblance is a large topic that is beyond the scope of this thesis. A discussion of resemblance can be found in Gamez (2007c, pp. 71-83) and it is also worth pointing out that the interpretation of the phenomenal and the physical that was presented in Chapter 2 provides a strong argument against the idea that phenomenal experiences associated with representational mental states resemble the physical world in any way.

This definition of representation is extremely broad and can be applied to any system. Even a stone sustains transient internal vibrations in response to a blow that can be interpreted as representational mental states. However, systems do exhibit substantial differences in the complexity of their representations. For example, humans have a vast repertoire of states linked to incoming light, whereas stones generate almost no internal states in response to light.

Many systems contain non-representational mental states. One candidate for a non-representational state was put forward by Block (1995), who claimed that the phenomenal content of orgasm is non-representational. This is not a particularly good example because the phenomenal content of orgasm can readily be interpreted as a representation of the internal states of a person's body, genitalia and emotion system. However, other human mental states are likely to be non-representational, such as the ones regulating breathing and the states corresponding to spontaneous neuron activity. The same is likely to be true of many artificial systems.

Mental states that represent other mental states can also respond to complex features of the world. For example a mental state that is functionally or effectively connected to mental states that respond to combinations of lines could become active when the system is presented with a cube. In this case the 'meta representation' is representing *both* the mental states responding to the lines *and* the presence of a cube in the world. Mental states that represent non-representational mental states lack this double level of representation.

Representations are most easily identified when the system is actively processing information from its environment. Under these conditions, the internal states can be measured and correlated with features of the data entering and leaving the system. At a later point in time these representational mental states might become activated when the stimulus is no longer present in a way that is analogous to imagination. Systems with language can be probed for these offline representations by asking them what they are imagining, but without this kind of first

person report it is difficult to identify unclassified representational mental states when the system is not actively processing the stimulus.

4.3.4 Identification of Representational Mental States

The general procedure for identifying representational mental states is to expose the system to a variety of different stimuli, record its responses, and look for functional or effective connections between the stimuli and the mental states.[23] To carry this out successfully, a comprehensive test suite needs to be designed that can probe a reasonable selection of the sensitivities of the system and specify them as precisely as possible. This could start with simple low level features, such as points, lines, and edges and work its way up to more abstract stimuli, such as faces and houses. All of these single modality tests would have to be combined with other modalities, such as audition, proprioception and sensation, and they would have to be carried out whilst the system is engaged in different activities, such as looking to the left, moving forward, and so on, to take account of sensorimotor contingencies. With even a moderately complex system this will soon escalate into an unmanageable number and complexity of tests. Some of these challenges could be met by appropriate subsampling of the test space and many tests can be automated by simulating input to the sensors. Common sense can also be used to prune the test suite down to a manageable size. This problem of scale will also appear in our animal and human tests as we improve our scanning and recording technologies.

The use of electrodes to identify representational mental states in animal and human subjects was pioneered by Hubel and Wiesel (1959), who inserted electrodes into the brains of cats and measured the activity of the neurons when different stimuli were presented in different

[23] One potential problem is that a system's representational mental states may change over time and it may have to be retested at regular intervals or have its adaptivity frozen whilst the description of its synthetic phenomenology is taking place.

parts of the visual field. Neurons whose activity changed[24] when the external stimulus was presented were judged to be representing the information in the stimulus. More recently a similar approach was pursued by Quian Quiroga et al (2005), who used electrodes to record from neurons in the medial temporal lobe in eight human subjects, who were presented with pictures of individuals, landmarks or objects. These experiments identified neurons that responded[25] to highly specific stimuli - for example one unit only responded to three completely different images of the ex US president Bill Clinton and another responded to pictures of the basketball player Michael Jordan.

The main limitation of using electrodes to identify representational mental states in human subjects is that simultaneous recording is only possible from a few hundred out of the billions of neurons in the brain. An alternative approach is to use scanning techniques, such as fMRI, PET, MEG and EEG to record the response of different brain areas as stimuli are presented. One example of this type of work is Haxby et al. (2001), who used fMRI to record the activity in the ventral temporal cortex while subjects viewed faces, cats, five categories of man-made objects and nonsense pictures. The distinct pattern of response that was identified for each category of object was linked by Haxby et al. (2001) to the presence of widely distributed and overlapping representations of faces and objects in the ventral temporal cortex. The main limitation of the scanning approach is that current procedures have limited spatial resolution – for example, fMRI measures the average activity within voxels of the order of 1 mm^3 - and so they can only be used to identify the general areas that hold representational mental states.

With artificial systems one generally has full access to their internal states and incoming/ outgoing data, and they can be probed precisely for all of their representations. Previous work in this area includes the backtracing method developed by Krichmar et. al. (2005), which finds

[24] This change could take several forms, such as an increase in firing rate, a decrease in firing rate or a burst of spikes in response to the onset or offset of the stimulus.

[25] A response was considered significant if it was larger than the mean plus 5 standard deviations of the baseline and had at least two spikes in the post-stimulus time interval (300–1000 ms).

functional pathways by choosing a reference neuronal unit at a specific time and then identifying the neuronal units connected to the reference unit that were active during the previous time step. This procedure is then repeated with the new list of neuronal units until the input neurons are reached that initiated the internal activity. Since the response characteristics of the input neurons are known, backtracing can be used to link internal states of the system to the stimuli presented to its sensors. Another way of identifying representational mental states in an artificial system is Granger causality, which is a method based on prediction that has been used by Seth (2007) to link a system's input to changes in its internal states. If a signal X_1 causes a signal X_2, then past values of X_1 should contain information that helps predict X_2 over and above the information contained in past values of X_2 alone. X_1 is said to Granger cause X_2 if the prediction errors in X_2 are reduced by the inclusion of X_1.[26] In this thesis representational mental states were identified using a method based on Tononi and Sporns (2003), in which noise was injected into the input or output layers and the mutual information that was shared between the input/ output and internal layers was measured. The full details of this procedure are given in Section 7.3.3.

4.3.5 Which Mental States are Phenomenally Conscious at Time *t* ?

At any point in time, many of a system's representational and non-representational mental states are *unconscious* (see Section 2.7.2), and to describe the *phenomenology* of the system a theory of consciousness is needed to predict which of the physically defined mental states are associated with phenomenal states. Since type I PCCs have been incorporated into the system's OMC rating, this separation between conscious and unconscious states is carried out using type II theories of consciousness. In this thesis I am using Tononi's theories about information integration, Aleksander's axioms and Metzinger's constraints (see sections 2.6.2 – 2.6.4) to make predictions about phenomenal states. Each of these theories can be used to predict which parts of

[26] More information about how Granger causality is calculated can be found in Seth (2007).

the system are conscious at time *t*, and these instantaneous predictions can be put together in a sequence to describe the evolution of the system's phenomenology over time. The details about how these theories are applied to the neural network developed by this project are given in Chapter 7.

Although I have decided to focus on the work of Tononi, Aleksander and Metzinger, the methodology described in this thesis is completely general and can be used with other type II theories of consciousness to make predictions about which mental states are associated with phenomenal states. It is highly likely that different theories of consciousness will make different predictions, and it may eventually be possible to discriminate between type II theories of consciousness by comparing their different predictions with first-person reports or the system's behaviour.

4.3.6 Integration Between Mental States

A description of the phenomenology of a system also has to identify the *relationships* between mental states, which determine how the mental states are integrated together into one or more consciousnesses. For example, consider a system that is looking at a red cube and has conscious representational mental states that respond to red information and conscious representational mental states that respond to shape information. If the colour and shape information is integrated or bound together, then it might be reasonable to claim that the system is conscious of a red cube. However, if the information is not integrated together, then it would be more accurate to say that there are two separate consciousnesses in the system: one that is conscious of redness, and another that is conscious of a cube. In humans, the importance of the integration between mental states is illustrated by the work on split brain patients (Gazzaniga, 1970), which suggests that two substantially independent consciousnesses are created when the corpus callosum is cut in the human brain, and the phenomenology of these two consciousnesses is likely to be very

different from that of a normal person. The integration between mental states can be identified using methods for measuring functional and effective connectivity, such as Granger causality (Seth 2007) and information integration (Tononi and Sporns 2003).

4.4 XML Description of the Phenomenology

4.4.1 Introduction

This section explains how information about a system's OMC rating and mental states can be integrated into a description of its phenomenology as it interacts with the world. A major problem with describing the phenomenology of artificial systems is that the words and structures of human languages are adapted to the description of human states. This problem is covered in Section 4.4.2 and Section 4.4.3 suggests why a markup language, such as XML, is more appropriate for synthetic phenomenology. Section 4.4.4 then outlines the structure of the XML that I will be using to describe the phenomenology of a neural network in this thesis. After a brief discussion of the use of XML to describe phenomenology, Section 4.4.6 looks at how this approach to synthetic phenomenology relates to the interpretation of the science of consciousness that was outlined in Section 2.4.5.

4.4.2 Problems Describing the Phenomenology of Non-Human Systems

Traditional phenomenology, especially in the work of Husserl (1960) and Heidegger (1995a), derives its significance from the claim that the phenomena we experience are as important and substantial as the physical world described by science, which is often portrayed as a secondary interpretation of our experiences. In this way traditional phenomenology sets itself up with an 'objective' field of phenomena that are assumed to be the same for everyone and can be unproblematically described in natural human language The problem with this approach is that these assumptions about common experience start to break down once phenomenology is applied

to the experiences of infants, animals and robots. To illustrate this problem, I will consider a short extract from Wordsworth (2004), which contains a fairly straightforward description of daffodils in natural human language:

> When all at once I saw a crowd,
> A host, of golden daffodils,
> Beside the lake, beneath the trees,
> Fluttering and dancing in the breeze.

Most people have had the experience of daffodils fluttering and dancing in the breeze and when Wordsworth's description is read by humans, they can readily imagine a similar past experience and understand his words well enough. Although this description is reasonably straightforward, it is actually an extremely vague and imprecise way of communicating daffodil information, and each reader will imagine the flowers differently. More serious problems start to arise when we try to use ordinary language to describe the experiences of an infant placed in front of a field of daffodils. As Chrisley (1995) points out, we cannot simply say that the infant sees a host of golden daffodils because the infant has a preobjective mode of thought, which is unable to locate the daffodils within a single unified framework. Adults understand daffodils as something objectively located in three dimensional space, whereas infants do not necessarily continue to believe in the existence of the daffodils when they are occluded. In the adult and infant the word "daffodils" refers to two different concepts and experiences. As Chrisley puts it: "The infant's concepts are not fully objective and are therefore non-conceptual. To ascribe conceptual content to the infant in this case would mischaracterize its cognitive life and would not allow prediction or explanation of the infant's behavior." (Chrisley 1995, p. 145).

These problems become even more difficult when the attempt is made to describe the phenomenology of a non-human animal, such as Nagel's famous bat (Nagel 1974). When a bat flies over a field of daffodils it receives a complex pattern of returning ultrasound pulses, which

are processed into phenomenal experiences that are likely to be very different from our own. Sentences like "the bat is experiencing a host of golden daffodils" are at best an extremely misleading description of the bat's phenomenology.

The same difficulties are encountered by attempts to describe the phenomenal experiences of artificial systems. For example, a robot that is pointing its camera at a field of daffodils might have phenomenal states associated with mental states that are effectively connected to its camera's response to yellow light (independently of the location, movement or shape of the light). However, we would have no basis for believing that the robot would have the *human* phenomenal experience of yellow when the daffodils were placed in front of it, or even that two different robots would have the same experience of yellow as each other. This problem becomes even more acute when a system has phenomenal states that are systematically related to features of the world that are *invisible* to human beings - for example, we have no words at all to describe mental states that respond to X rays.

One approach to this problem would be to describe the scene in front of the robot in the language of physics – for example, we could talk about the system having a representation of 590 nm electromagnetic waves, instead of talking about it experiencing yellow light,[27] and use the language of chemistry, biology and geometry to describe the features of the daffodils that the system is sensitive to. The trouble with this approach is that it does not describe the *phenomenology* of the system and it has the limitation that the data coming out of a system does not always lend itself easily to an objective physical description. For example, to describe the motor output signals that control an eye or arm one would have to come up with a physical description of the eye or arm and specify its movement relative to a frame of reference that

[27] There is not a straightforward link between wavelength of light and perceived colour and it is possible to experience yellow when there are no 590 nm electromagnetic waves present. This problem has been set aside in this thesis - in the future a more accurate physical description could specify all of the physical conditions under which we would experience yellow.

would also have to be physically described. Whilst this can be done, it is much easier to interpret the motor output signals as an eye or arm movement.

The pragmatic solution that will be followed in this thesis is to use both human and physical descriptions to describe a system's representational mental states, when these are possible and appropriate. The human description should be interpreted with caution (the phenomenology of an artificial system that only responds to yellow is likely to be very different from our human experience of yellow) and the physical description should only be taken as a starting point for a phenomenological description. In the future, it may be possible to create closer links between phenomenological and physical descriptions - perhaps by using the information characteristics of mental states (Tononi 2004) or by applying O'Regan and Noë's (2001) theories about sensorimotor contingencies.

4.4.3 Markup Languages for Synthetic Phenomenology

A combination of human and physical descriptions enables something to be said about the contents of an artificial system's phenomenal states, but it does not capture the *relationships* between them. Furthermore, depending on how mental states are defined for the system, there could be millions or even billions of active mental states that are predicted to be associated with consciousness at any point in time. Even if it was possible to integrate all of these mental states into a natural language description, the resulting document would be so long and tedious that it would be almost impossible to read.

One way of solving these problems is to abandon the attempt to describe the phenomenology of artificial systems in natural human language and use a markup language, such as XML or LMNL, to structure the descriptions of the representational mental states and to indicate the relationships between them. There are a number of reasons why a markup language would be a good choice for the description of an artificial system's phenomenology:

- Markup languages are much more precise and tightly structured than natural language, which enables markup languages to describe complex nested hierarchies and represent some of the relationships between different pieces of information.

- Markup languages can describe low level details of a system's hardware, but they can also abstract from them, so that high level comparisons can be made between machines with different architectures and between humans and machines. Whilst two systems' lower levels might be different – perhaps using neurons or silicon - the higher levels are likely to be more similar, which would allow direct comparisons between them once everything was encoded into a markup language.[28]

- Markup languages can be written and read by both machines and humans. With simple small-scale analyses it is useful to be able to manually read and edit a description of a machine's mental states. However, it is also relatively easy to automatically generate and analyse the states of a machine using a markup language, for example by writing programs that look for phenomenal states using different type II theories of consciousness.

- Data that has been structured using a markup language is typically stored in plain text files that can be shared between different operating systems and easily archived, either by converting them into a database or by storing them directly.

- The structure of some markup languages can be validated without prior knowledge of their form.

- Once you have a highly structured representation of a machine's mental states and a methodology for analysing them for phenomenal consciousness, it is possible to see how a machine's conscious states can be extended or enhanced.

[28] Coward and Sun (2007) claim that this type of hierarchical description is necessary for a science of consciousness.

- Markup languages are a good foundation for other techniques for representing non-conceptual mental states, such as the suggestions made by Chrisley (1995) about content realization, ability instantiation and self instantiation (see Section 4.5), which depend to some extent on a precise specification of states of oneself and the environment
- Markup languages can be very flexible. For example, in addition to tags and data, XML can contain references to external files, pieces of code and equations, which enables it to include features that cannot be precisely described in natural language.

Whilst a number of markup languages, such as JSON, LMNL, YAML and OGDL, would have been appropriate for synthetic phenomenology, the popularity of the eXtensible Markup Language (XML) and the availability of good parsers in most programming languages made it a good choice for illustrating this approach. In the future it might be necessary to change to a more sophisticated markup language, such as LMNL, which supports overlapping elements and structured attributes.[29]

4.4.4 Example XML Description

This section outlines the XML structure that will be used to describe the phenomenology of an artificial neural network in Chapter 7. This is only an example, rather than a fully fledged standard, because it is tailored to an approach in which individual neurons are interpreted as individual representational states, and the mutual information shared between each of the internal neurons and neurons in the input and output layers is calculated using the methodology described in Section 7.3.3. If XML is found to be a useful way describing the phenomenology of artificial systems, then it is hoped that a more general specification can be developed. This example does

[29] A good XML tutorial can be found at: http://www.w3schools. com/ xml/default.asp. More information about LMNL can be found here: http://lmnl.net/.

not include non-representational mental states and mental states that represent other mental states. As Chapter 7 shows, at the current stage of research it is hard enough to identify and describe mental states that are systematically related to states of the world, without trying to include mental states that are almost impossible to articulate in human language.

```
<!-- Standard XML header. -->[30]
<?xml version="1.0" encoding="ISO-8859-1"?>

<!-- Start of the analysis. -->
<analysis>

  <!-- General description of the contents of the file. -->
  <description>Synthetic phenomenology of the SIMNOS virtual robot.
                                                    </description>

  <!-- Author(s) of the file and date on which the analysis was generated. -->
  <author>David Gamez</author>
  <date>Mon Jan 28 14:44:27 2008</date>

  <!-- The system that is being analysed along with its version number. A full description of the
       system should be included in the source files. -->
  <system>SIMNOS version 1.0; SpikeStream version 0.1</system>

  <!-- Source files for the analysis. These include the files for the neural network (if there is
       one, since the system may not be neural) and the analysis files. Source files should always
       be included with the phenomenological description to enable other researchers to validate
       the predictions and generate their own description of the synthetic phenomenology. -->
  <source_files>
    <file>TrainedNeuralNetwork_version1.sql.tgz</file>
    <file>AnalysisRun1_NoiseRun1_NeuralArchive.sql.tar.gz</file>
  </source_files>

  <!--The archive that is being described. -->
  <archive>Analysis Run 1 [ 2007-12-18 20:42:55 ]</archive>

  <!-- The time step of the archive that is being analyzed or the time at which the data was
       captured from the system. -->
  <time_step>13194</time_step>

  <!-- Start of the phenomenological description. -->
  <phenomenology>
```

[30] XML comments start with "<!-- " and end with "-->".

```
<!-- The next part of the file lists the system's mental states. These may be representational
   and they may be predicted to be conscious according a type II theory of consciousness. -->

  <!-- A mental state of the system. -->
  <mental_state>

    <!-- The OMC rating of the part of the system in which this mental state is instantiated,
       along with the version of the scale that is being used. -->
    <omc_scale>
      <rating>0.427</rating>
      <version>0.6</version>
    </omc_scale>

    <!-- In this example mental states are active neurons. -->
    <physical_description>
      <firing_neuron>
        <id>120811</id>
      </firing_neuron>
    </physical_description>

    <!--The cluster tag is used to indicate the functional or effective connectivity between
       this mental state and other mental states. Different methods can be used to measure
       this, such as information integration (Tononi and Sporns 2003). -->
    <cluster>
      <id>200809</id>
      <type>phi</type>
      <amount>75.1173</amount>
    </cluster>

    <!-- List of the states of the world that are functionally or effectively connected to this
       mental state. In this example, representational states are identified using the mutual
       information that is shared with neurons in the input or output layers – see Section
       7.3.3. -->
    <representations>

      <!-- This mental state is effectively connected to data leaving the system. -->
      <output>
        <neuron>
          <id>127936</id>
        </neuron>
        <mutual_information>0.993765</mutual_information>
        <human_description>Proprioception / motor output
                                                  </human_description>
        <physical_description>N/A</physical_description>
      </output>

      <!-- This mental state is effectively connected to data entering the system. -->
      <input>
        <neuron>
```

```
            <id>104327</id>
          </neuron>
          <mutual_information>1.00854</mutual_information>
          <human_description>Red / blue visual input
                                                  </human_description>
          <physical_description>700/450 nm electromagnetic waves
                                                 </physical_description>
        </input>

        <!-- Further input and outputs can be added here. -->

      <!-- The end of the list of representations. -->
      </representations>

      <!-- Type II theories of consciousness are used to predict whether phenomenal
           consciousness is associated with this mental state. In this example, the predictions
           are made using Tononi's (2004), Aleksander's (2005) and Metzinger's (2003)
           theories. -->
      <phenomenal_predictions>

        <!-- Whether this mental state is part of the conscious part of the system according
             to Tononi's theory of consciousness (see Section 7.5 for the criteria for this). -->
        <tononi>0</tononi>

        <!-- Whether this mental state is part of the conscious part of the system according
             to Aleksander's theory of consciousness (see Section 7.6.2 for the criteria for
             this). -->
        <aleksander>0.993765</aleksander>

        <!-- Whether this mental state is part of the conscious part of the system according
             to Metzinger's theory of consciousness (see Section 7.7.3 for the criteria for this).
             -->
        <metzinger>75.1173</metzinger>

        <!-- Other phenomenal predictions can be added here. -->

      <!-- The closing tag of the phenomenal predictions. -->
      </phenomenal_predictions>

    <!-- The closing tag of the mental state. -->
    </mental_state>

    <!-- Any number of mental states can be added here. -->

<!-- The end of the description of the phenomenology of the system. -->
</phenomenology>
```

```
<!-- This final closing tag ends the analysis of the system. -->
</analysis>
```

4.4.5 A Description of the Synthetic Phenomenology?

Given the history of phenomenology, we might expect that the final outcome of synthetic phenomenology would be a natural language description. Even if we cannot achieve this at present, it might be thought that this should be the final goal of the procedures outlined in this chapter. Viewed from this perspective, the markup language would only be a preparatory stage that would help us to prepare a traditional phenomenological account of the experiences of COG, CRONOS or IDA. However, the problems discussed in Section 4.4.2 make it unlikely that we are ever going achieve fluid natural language descriptions of the phenomenology of non-human systems. Instead, it might be much better to treat the XML as the best description that we are going to get of the phenomenology of an artificial system. We don't have adequate words in human language to describe a system that can only experience vertical lines, but we can represent such a system accurately using XML, and by looking at the XML we can start to understand how much and how little we can imagine what it is like to be such a system. Some of the issues raised by the use of XML in synthetic phenomenology are covered in Section 7.9.9.

4.4.6 Synthetic Phenomenology and Science

This section takes a brief look at how this approach to synthetic phenomenology fits in with the approach to the science of consciousness that was put forward in Section 2.4.5. The main difference between the study of human consciousness and synthetic phenomenology is that robots are currently unable to describe their conscious states, and so we can only make *predictions* about their consciousness based on theories that have been developed using humans and animals.

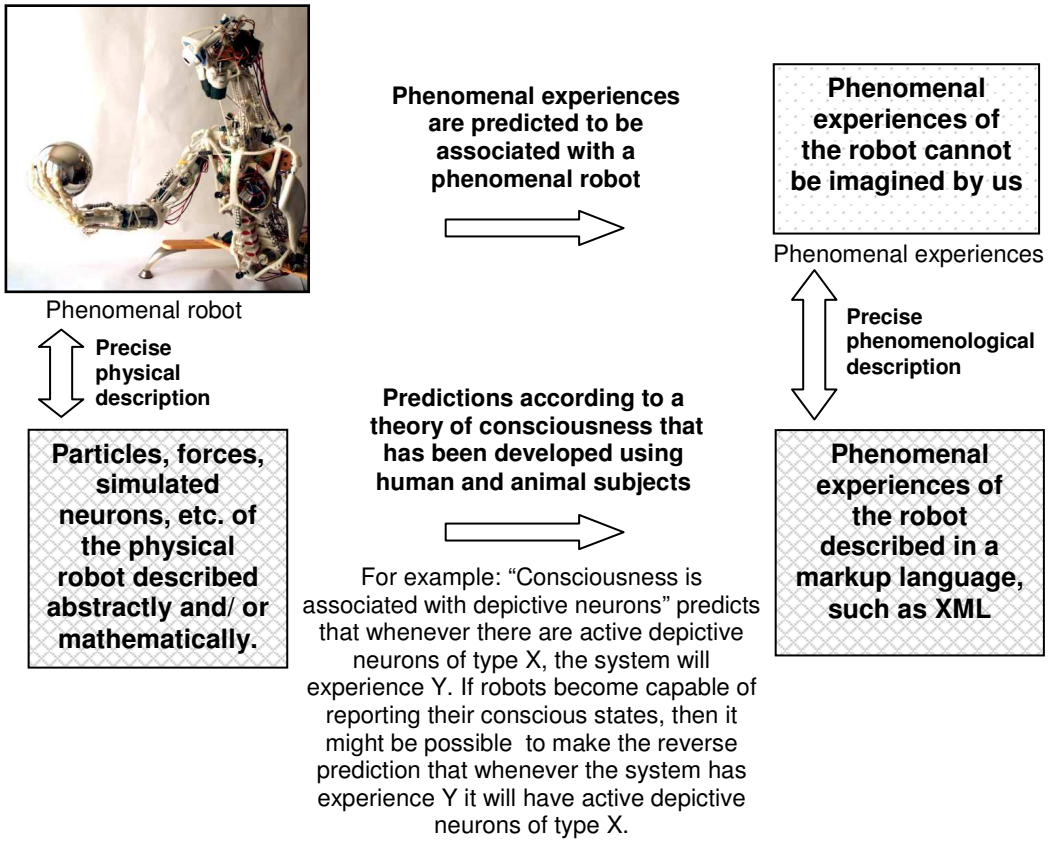

Figure 4.1. How synthetic phenomenology fits in with the approach to the science of consciousness that was put forward in Section 2.4.5. With artificial systems, it is only possible to make predictions about the phenomenal experiences that are associated with them, and so there are unidirectional arrows from the phenomenal robot to the robot's phenomenal experiences and from the description of the physical system to the description of the robot's phenomenology. This diagram should be contrasted with Figure 2.4 in Chapter 2, where the horizontal arrows are bidirectional because the association between phenomenal experiences and the phenomenal brain is the starting point for experiments on the correlates of consciousness and systematic relationships are being identified between the phenomenal and physical descriptions.

This situation is illustrated in Figure 4.1, in which the arrows between the robot and its phenomenal states and between the physical and phenomenal descriptions are only one way to indicate that phenomenal states are predicted to be associated with the robot. If we can develop robots that can report their conscious states, then it will be possible to validate these predictions and speak about an association between the phenomenal states and the robot.

4.5 Previous Work in Synthetic Phenomenology

This approach to synthetic phenomenology has been substantially influenced by the previous work in traditional phenomenology, such as Husserl (1960, 1964), Merleau-Ponty (1989, 1995) and Heidegger (1995a), which attempted to describe different aspects of human conscious experience from a first person perspective. These descriptions were carried out in natural language and generally took the position that the physical world is a secondary interpretation of our phenomenal experiences and not something to which our phenomenal experiences should be reduced. Although Heidegger (1995b) made some attempts to understand animal consciousness, the main emphasis of traditional phenomenology is on human phenomenal experience.

The question whether artificial systems are capable of conscious states has been extensively discussed in the literature on consciousness and the contributions roughly divide into those who accept the difficulties with behaviour-based attribution of phenomenal states, and those who have a theory of consciousness that enables them to make definite claims about which machines are phenomenally conscious. In the first group, Moor (1988) sets out the arguments against knowing for certain whether robots have qualia, but claims that we need to attribute qualia to robots in order to understand their actions. A similar position is set out by Harnad (2003), who claims that the other minds problem limits us to attributing consciousness on the basis of behaviour, and so any robot that passes the T3 version of the Turing test for a lifetime must be acknowledged to be conscious. Prinz (2003) is closest to the position of this thesis since he does not think that we can identify the necessary and sufficient conditions for consciousness and does not suggest other grounds for attributing consciousness to machines.

People who claim to know exactly what the causes or correlates of consciousness are can say precisely which machines are capable of phenomenal states - replacing the OMC scale set out in this chapter with a dividing line dictated by their theory of consciousness. One of the most liberal of these theories is Chalmers (1996), whose link between consciousness and information

leads him to attribute phenomenal states to machines as simple as thermostats. At the other extreme, Searle (1980) believes that his Chinese room argument excludes the possibility that any of the functional levels could be simulated and Searle (2002) rather vaguely ties consciousness to a causal property of matter, so that only biological humans, animals and possibly aliens could be conscious. In between these positions are people like Aleksander and Morton (2007a), who set out two criteria that a system must conform to if it is to be a candidate for synthetic phenomenology: "To be *synthetically phenomenological*, a system S must contain machinery that represents what the world and the system S within it *seem* like, from the point of view of S." (Aleksander and Morton 2007a, p. 110). An unpacked version of this definition is used by Aleksander and Morton to argue that their own kernel architecture is synthetically phenomenological, whereas the global workspace architecture is not.

Once it has been decided which artificial systems are capable of phenomenal states (if any) the second question faced by synthetic phenomenology is how artificial phenomenal states can be described. One approach to this was put forward by Chrisley (1995), who set out a number of techniques for representing non-conceptual content. These include content realization, in which content is referred to by listing "perceptual, computational, and/or robotic states and/or abilities that realize the possession of that content" (Chrisley, 1995, p. 156), ability instantiation, which involves the creation or demonstration of a system that instantiates the abilities involved in entertaining the concept, and two forms of self instantiation, in which the content is referred to by pointing to states of oneself or the environment that are linked to the presence of the content in oneself. More recently Chrisley and Parthemore (2007) used a SEER-3 robot to specify the non-conceptual content of a model of perception based on O'Regan and Noë's (2001) sensorimotor contingencies. Initially the robot had no expectations about what it was going to see and as it moved its eye around it built up expectations about what it would see if it were to move its eye to a particular position. These expectations were plotted for each position in visual

space to generate a graphical representation of the robot's visual experience. Chrisley and Parthemore used this representation to evaluate some aspects of O'Regan and Noë's (2001) theory, such as their interpretation of change blindness and how visual experience appears to be coloured at the periphery despite the lack of colour receptors outside the fovea. Other graphical representations of a robot's inner states have been produced by Holland and Goodman (2003) and Stening et al. (2005), who plotted the sensory and motor information stored in a Khepera's concepts. More details about this work are given in Section 3.5.5.

Synthetic phenomenology has a number of overlaps with the description of human phenomenology from a third person perspective. This type of research is commonly called "neurophenomenology", although this term is subject to two conflicting interpretations. The first interpretation of "neurophenomenology" was put forward by Varela (1996), who used it to describe a reciprocal dialogue between the accounts of the mind offered by science and phenomenology.[31] This type of neurophenomenology emphasises the first person human perspective and it has little in common with synthetic phenomenology. However, neurophenomenology can also be interpreted as the description of human phenomenology from a third person perspective using measurements of brain activity gathered using techniques, such as fMRI, EEG or electrodes. Good examples of this type of work are Kamitani and Tong (2005), Haynes and Rees (2005a,b) and Kay et al. (2008), who used the patterns of intensity in fMRI voxels to make predictions about the phenomenal states of their subjects. In some ways neurophenomenology is easier than synthetic phenomenology because it does not have to decide whether its subjects are capable of consciousness and the description of non-conceptual states is considerably simpler in humans. However, both disciplines are attempting to use external data to identify phenomenal states in a system and there is considerable potential for future collaboration between them.

[31] A review of this interpretation of neurophenomenology can be found in Thompson et al. (2005) and it had a substantial influence on the analysis of consciousness in Chapter 2.

4.6 Conclusions

This chapter has set out an approach to synthetic phenomenology that can be used to describe a machine's predicted phenomenal states. Since the link between type I PCCs and consciousness cannot be empirically established, the first part of this chapter outlined an OMC scale, which models our subjective judgement about the relationship between type I PCCs and consciousness. The next part of this chapter developed concepts of a mental state and a representational mental state and outlined how these could be identified in a system and used to make predictions about phenomenal states using type II theories of consciousness. Problems with the description of artificial phenomenal states in human language were then discussed and it was suggested how a markup language, such as XML, could be used to describe the phenomenal states of artificial systems.

The next chapter outlines the design and implementation of a neural network that is based on the some of the theories of consciousness set out in Chapter 2. The approach to synthetic phenomenology that has just been described is used to make predictions about the consciousness of this network in Chapter 7.