An Ordinal Probability Scale for Synthetic Phenomenology

David Gamez*
*Department of Computer Science
University of Essex
Colchester
C04 3SQ, UK
daogam@essex.ac.uk

Abstract

Synthetic phenomenology can be broken down into three areas: (1) the determination whether a system is capable of phenomenal states, (2) the identification of the mental content of the machine (the machine's conceptual and non-conceptual representations), and (3) the analysis of a particular structure of mental content to identify the parts that are phenomenally conscious. This paper proposes that an ordinal probability scale could be used to address the first of these problems and sets out a proposal for such a scale that ranks machines according to the likelihood that they are capable of experiencing phenomenal states. The overall approach suggested here will be used to describe the synthetic phenomenology of Holland's and Troscianko's 'conscious' robot that is currently under development at the University of Essex and the University of Bristol.

1 Introduction

Research on machine consciousness aims to develop machines that exhibit conscious behaviour and might be capable of phenomenal states. A description of machines' phenomenal states is provided by synthetic phenomenology, which attempts to discover whether robots can have conscious experiences and to articulate them when they occur.

One of the challenges of synthetic phenomenology is that it seems possible that a zombie robot could perfectly mimic human behaviour without experiencing anything at all, and so external behaviour does not seem to be a reliable guide to conscious states. It might be thought that we could solve this by looking at the internal structure of the robot. If the robot contains structures that are correlated with or cause consciousness in humans, then it is likely to experience phenomenal states. The first half of this paper covers the problem with this approach: With only behavioural evidence to go on it is impossible to empirically identify a necessary and sufficient set of the correlates or causes of consciousness. Without this, we cannot identify what needs to be included in a machine to make it conscious and cannot tell whether a machine that exhibits conscious behaviour is likely to be experiencing phenomenal states.

To address this difficulty, the second half of this paper sets out a proposal for an ordinal probability scale that ranks machines according to the likelihood that they are capable of sustaining phenomenal consciousness, based on their proximity to our own case. The factors that are used in this scale are the rate of information processing, the size of the machine, the way in which its sub-functions are assembled to produce the global functions (functional granularity), the machine's time-slicing and whether it is analogue or digital. Weightings are given to each of these factors and the combination of these weights is used to situate each machine on an ordinal scale.

This scale is put forward as a pragmatic tool that will enable us to proceed with synthetic phenomenology and address some of the ethical issues in machine consciousness. At this stage, the proposed factors, along with their weightings, should be seen as tentative first suggestions, which I hope will be criticised and develop in the longer term into a commonly agreed scale.

2 Synthetic Phenomenology

Whilst synthetic phenomenology can be used to refer to the *synthesizing* of phenomenal states (Jordan (1998) coined the term in this connection), it can also be used to describe the *phenomenology* of artificial systems that may or may not be experiencing conscious states. It is in the latter sense that I will be using it in this paper. Husserl's phenomenological project was the description of human consciousness (without any commitments to the natural attitude); the synthetic phenomenological project is the description of machine consciousness. Synthetic phenomenology is a way in which people working on machine consciousness can measure the

extent to which they have succeeded in realising consciousness in a machine.

The phenomenology of artificial systems can be broken down into three stages: (1) the determination whether a system is capable of phenomenal states, (2) the identification of the 'mental content' of the machine[1] (the machine's conceptual and non-conceptual representations), and (3) the analysis of a particular structure of mental content to identify the phenomenally conscious parts. The first of these stages is the main focus of this paper, but before examining it in detail I will give a brief overview of all three stages in order to clarify the relationship between them.

2.1 Can a machine experience phenomenal states?

The external behaviour of a robot could be taken to indicate that it is experiencing phenomenal states, or it could just be the result of unconscious automatic processing. Even if the robot could master language and pass the Turing test, this would not guarantee that it is experiencing phenomenal qualia. It is not even inconsistent or wrong to see animals or other human beings as automatons.

We avoid solipsism and attribute consciousness to other humans and some animals because we share a similar biology. In the case of machines, this common underlying substrate is missing and so we need to find some other way to decide whether they are capable of phenomenal states or not. The most promising line of approach would be to work out what it is about our biology that makes us conscious – its proteins, neurons, functions or representations, for example - and then look inside the robot to see if these consciousness-producing properties are present as well. Alternatively, we could look for the *correlates* of phenomenal states in the brain. This weaker approach looks for the factors that are present whenever consciousness is present without trying to explain how these factors actually lead to conscious states. When we have identified either the mechanisms in the brain that produces phenomenal states or the correlates of phenomenal states, we can see if these mechanisms or correlates are present in the machine. If they are, then it seems reasonable to conclude that the machine may be capable of phenomenal states.

Unfortunately it does not look as if either of these approaches will be able to answer the question about the consciousness of non-biological entities. At present we have no idea about how the brain produces conscious states and there are some potentially irresolvable difficulties with empirically separating out the correlates of consciousness, which may prevent us from making any progress at all in this direction. Section 3 covers these problems in detail. If a solution cannot be found, we may only be able to answer the questions raised by this first proposed stage of synthetic phenomenology with a probabilistic assessment of the likelihood that a machine can support phenomenal states. This will be discussed in the second half of this paper.

2.2 Identification of the mental content of a machine

As a machine interacts with its real or virtual environment it processes sense data into some kind of representation of its world – that there is a green apple five metres from its hand, that its hand is clenched, that it is feeling sad, and so on. These internal representations are the mental content of the machine.

The identification of conscious mental content in humans is imperfect, but relatively straightforward since we share language and a common biological base. When people describe their phenomenal world our common biology leads us to assume that it is very similar to our own. However, this is not the whole story, because people also have a great deal of unconscious and non-conceptual mental content that can only be identified indirectly (Chrisley, 1995).

The identification of mental content in robots is more challenging because they generally have rudimentary language and are built in a very different way from humans. This means that we cannot simplistically assume that their phenomenal experiences are in any way similar to our own. When the robot's information-processing is based around neural networks, we can try to identify the robot's mental content by exposing it to stimuli and interpreting the active parts of the network as representations of the external stimulus. By systematically varying the stimuli, a map of the representations in the network can be worked out and used to describe the robot's mental content in novel situations.

A second technique was suggested by Holland and Goodman (2003). In their experiments a simple robot was programmed to move around its environment and build up 'concepts' corresponding to combinations of sensory input and motor output. Once the robot's concept formation was complete, its mental content could be identified by a process of 'inversion'. Each concept was a combination of data about the distance of environmental features and information about how the robot moved during the sample period. The inversion consisted in plotting the movements and distances recorded by each

[1] I will be using the term 'mental content' to refer to a machine's representations and the non-conceptual content of its 'mind'.

concept to generate a map of the robot's representation of its environment.

2.3 Separation of conscious from unconscious mental content

Machines that can support phenomenal states present a third problem for synthetic phenomenology. If they are anything like humans, it is likely that at any point in time some of their mental content is conscious and the rest unconscious. In the well-known example of driving home from work, a person can avoid obstacles in the road, stop at traffic lights, navigate perfectly and yet see and remember nothing of the journey because they are preoccupied with something else. The sense data from the road, steering wheel and pedals must be mental content in some form, but it is not conscious mental content. Some feature of the mental content that the person is attending to must differentiate it from information about the road, steering wheel and pedals.

Although robots might be *capable* of consciousness, at any point in time it is possible that none of their mental content is actually conscious, and one of the tasks of synthetic phenomenology is to distinguish the conscious from the unconscious mental content. Different theories about human consciousness are gradually converging around the idea that it is the structure and interconnection of information that makes the difference between conscious and unconscious content (See Dehaene (2001) for an overview). To carry out this part of its task, synthetic phenomenology can use these theories about consciousness (for example, the global workspace model put forward by Baars (1988) or Metzinger's (2003) constraints) to identify the phenomenal mental content of the machine.

3 Inferring Consciousness from Behaviour

3.1 Local and global functions

The brain carries out a wide variety of functions, ranging from information and language-processing to low level functions in the neurons' ion channels. Some of these functions are *local* to regions of the brain, for example transforming retinal data entering V5 into movement information, whereas other functions are *global* to the whole brain, for example transforming incoming sensory data about an apple into an output instructing the arm to pick it up.

It might be thought that we can easily determine which of the brain's local functions are necessary and sufficient for consciousness. For example, if we damage the function of V5, then the person loses consciousness of movement information (see Zihl et al. (1983) for a case study and also Zeki and Bartels (1998) for the notion that micro-consciousnesses are distributed throughout the brain). However, this type of experiment only identifies a link between a local function and consciousness *indirectly* through its impact on the brain's global functions. If the person's global functions were not affected by damage to V5, we would have no idea whether the local function carried out by it had any effect on consciousness.

It is even harder to decide whether the way in which a global or local function is implemented affects consciousness. The brain's global or local functions could be carried out by neurons or the population of China, but as long as its global functions remain constant, it will always describe its conscious experiences in the same way. A function that was carried out consciously when there were biological neurons present might be carried out unconsciously when there are no biological neurons present, but the function is still carried out, and so in both cases the person will continue to respond to the input: "Are you conscious?" with the output "Yes!" even if there is no longer any consciousness present. To make this point clearer I will consider a thought experiment that is often discussed in the literature, in which part of a person's brain is replaced by a chip that carries out the same functions as the brain part that is replaced.

3.2 Silicon brain functions

At first glance the replacement of a brain part by a chip seems to hold out the prospect of identifying whether the way in which the brain's functions are implemented affects consciousness. If we replaced V5 with a functionally equivalent chip and lost consciousness of movement information, then we could conclude that the brain's biological substrate and functions are *both* necessary for consciousness. However, as Moor (1988) and Prinz (2003) point out, since the global behaviour of the person would not be changed by the operation, neither an external observer nor the person who received the chip implant could observe any effect of the replacement on consciousness.

An outside observer would not detect the replaced part because the function of V5 would still be carried out by the chip. The person would still report movement information that is processed by affected area, even though there may not be any consciousness of movement present. From an outside point of view, this will not even seem like a confabulation because the visual system will be working perfectly.

A first-person perspective does not help matters either. Since the chip is functionally connected to

the rest of the brain in the same way that V5 was before the operation, our language centres will report phenomenal movement in the same way that they did before and it has already been established that the external behaviour of the person will remain unchanged. Searle (1992, 66-7) thinks that we might feel forced to say that we experience movement even though we do not experience any movement. However, if this distinction between inner thought and outer behaviour was conscious and could be remembered, it could be reported at a later time, and so there would be a change in the subject's behaviour, which is ruled out by this experiment. Furthermore, as Moor points out, the chip must also give the person the *belief* that they are conscious of the functions processed by the chip and so Searle cannot experience one thing and believe another. It seems that even a first-person perspective cannot be used to decide whether consciousness is affected by the replacement of biological neurons with a functionally equivalent chip.

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

3.3 Correlates and causes of consciousness

It might be objected that a great deal of progress has been made with identifying the correlates of consciousness, which in the longer term may enable us to work out what its causes are. Crick and Koch (2003) give a nice overview of this work and a more specific example would be Aleksander's (2000) connection between gaze-locked cells (identified by Galletti and Battaglini (1989)) and our experience of stable objective space. Eventually external observers may be able to use a brain scan to make a precise description of a person's conscious mental content.

However, whilst this work shows that certain patterns of firing neurons or synchronization between them are necessary and perhaps sufficient correlates of consciousness *in the human case*, they do not show that these are necessary and sufficient correlates of consciousness *in general*. All of the experiments on the correlates of consciousness have been carried out on biological subjects and so it is not clear whether the brain's functions are correlated with consciousness by themselves or whether a biological substance is also necessary. Without systematic separation of the factors it is impossible to say whether a robot with the same global functions as a human would experience phenomenal states.

Taken together, the arguments in this section force us to the conclusion that no test can separate out necessary and sufficient correlates or causes of consciousness. We can vary the ways in which the global functions of the brain are implemented in a vast number of ways, but since these will always lead to the same behavioural output, any impact of these changes on consciousness cannot be measured and we will never know for certain whether a functionally (and thus behaviourally) identical robot has conscious states or not.

4 Ordinal Probability Scale

Faced with these difficulties we could follow Prinz (2003) and suspend judgement about whether robots built from different principles are capable of supporting phenomenal states. However, there are three problems with this mysterianism. To begin with, we have a strong intuition that machines built along similar lines to human beings are likely to be phenomenally conscious. The more similar a system is to human beings, the more likely we are to believe that it experiences conscious states of some kind. Second, as machine consciousness develops we will be developing machines that exhibit increasingly complex behaviour and spend a lot of time in confused states and potentially in pain. This has been somewhat dramatically compared by Metzinger (2003) to the development of a race of retarded infants for experimentation. To address these ethical worries without stifling research a way needs to be found to evaluate the probability that a robot is experiencing phenomenal states. A third problem with mysterianism is that as more sophisticated robots emerge, people are inevitably going to attribute more and more conscious states to them. People already attribute feelings to Kismet or AIBO, and a systematic way of evaluating phenomenal probability 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 reply, 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, I propose the construction of a probability scale that orders machines according to the probability that their architecture is capable of supporting conscious states. This says nothing about whether a machine is *at present* conscious (this is the task of the second and third stages of synthetic phenomenology outlined in section 2); only whether it is likely that this kind of system can support conscious states.

I will start this description of the scale with an overview of the systems that are covered by it. After explaining the factors and the way in which they are combined, I will give a few specific examples to illustrate how it works.

4.1 Systems covered by this scale

This scale only covers systems that approximate the *global* functions of a human brain. By global functions I mean the functions that transform the brain's sensory inputs into motor outputs along the nerves connecting the brain to the body. Such a system could either be used to control a real human body or it could have its own real or virtual artificial body. In the latter case, the artificial body would have to have approximately the same number and type of sensors and effectors as the human body with approximately the same resolution.

The notion of approximating the global functions of the human brain is defined here using Harnad's (1994) extended T3 version of the Turing test. A machine that approximates the functions of a human brain by controlling a human or artificial body would have to be completely indistinguishable in external function from humans for 70 years or more. Such a robot could hold down a job, create works of art and have relationships with other human beings. Machines that were interned in an asylum for strange behaviour would not be considered functionally identical to a human being.

4.2 Factors affecting the probability of phenomenal consciousness

This scale is constructed in relation to humans, who are at present the benchmark example of conscious machines. The more similar a machine is to a human, the more likely it is to be phenomenally conscious. The factors within each group are assigned weightings (W) ranging from 1.0 to 0.1. These are arbitrary values and the way in which they are combined and converted into an ordinal scale is explained in section 4.3. An outline of the factors that I have selected for this first draft of the probability scale now follows.

4.2.1 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.

Table 1: Rate factors

	Rate	**W**
R1	Approximately the same speed as human brain	1.0
R2	10 times faster or slower than human brain	0.55
R3	Over 100 times faster or slower than human brain	0.1

4.2.2 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.

Table 2: Size factors

	Size	**W**
S1	Approximately the same size as human brain	1.0
S2	1000 times larger or smaller than human brain	0.55
S3	More than a million times larger or smaller than human brain	0.1

4.2.3 Functional granularity

This probability scale keeps the global functions of the brain constant. However, there is a wide variety of ways in which the global functions of the brain can be implemented by different collections of local functions, some of which are closer to the human brain than others. This factor weights machines according to the degree to which their functional granularity 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, the function of the whole brain could be implemented

by an vast neural network trained by back propagation, or it could be implemented by a more biological structure of neurons. Since neurons can themselves be simulated using neurons there is also potentially infinite self-recursion, which I have limited by a restriction introduced in section 4.3. To keep things simple I have set aside the possibility that glia play an information-processing role.

The way in which these four tables are combined is fairly self-evident. If the brain's global functions are implemented by a biological structure of modules, then the way in which the functions of the modules are implemented has to be specified as well. On the other hand, no further levels are required if the brain's global functions are implemented by a simulation that is not biologically structured.

Table 3: Whole brain function

	Function of whole brain	**W**
FW1	Produced by a biological structure of modules	1.0
FW2	Produced by a non-biological structure of modules	0.7
FW3	Produced by a non-biological structure of neurons	0.4
FW4	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 4: Module functions

	Function of modules	**W**
FM1	Produced by a biological structure of neurons	1.0
FM2	Produced by a non-biological structure of neurons	0.7
FM3	Produced by a mixture of methods	0.4
FM4	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 5: Neuron function

	Function of neurons	**W**
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	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 6: Function of molecules, atoms and ions

	Function of molecules, atoms and ions	**W**
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	Simulated using mathematical algorithms, computer code or some other method	0.1

4.2.4 Simulation time-slicing

The simulation of brain functions can be carried out in parallel with all the different functions working 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 simulations, which only have the same time complexity as the brain, as being less likely to be phenomenally conscious than simulations whose parts have the same moment-to-moment space complexity as the brain. In this first draft of this scale, I have placed all of the different types of simulation hardware together – such as a modern computer built from silicon and copper, a light computer, Searle's Chinese room or the economy of Bolivia. I have also set aside the potential question about the link between consciousness and virtual machines.

Table 7: Simulation time slicing

	Simulation time slicing	**W**
STS1	Complete hardware simulation in which all parts of the model are dynamically changing and co-present at any point in time	1.0
STS2	Multi-processor time-sliced simulation in which only parts of the model are dynamically changing and co-present at any point in time	0.55
STS3	Single processor time-sliced simulation in which only a single part of the model is dynamically changing and present at any point in time	0.1

4.2.5 Analogue / digital

With an analogue simulation there is an infinity of possible states, which can only be approximated by a digital simulation. It is possible that some nonlinear properties of the brain are more faithfully captured by an analogue simulation.

Table 8: Analogue / digital simulation

	Analogue / digital	**W**
AD1	Analogue simulation	1.0
AD2	Mixture of analogue and digital	0.55
AD3	Digital simulation	0.1

4.3 Putting it all together

To obtain the final ordinal probability scale, a complete list of all the possible machines is extracted from the factor tables. The weightings that are applicable to each machine are then multiplied together to give a total weighting for each machine. These are then used to situate all of the different machines in an ordinal scale. Since many of the machines have the same total weighting, this scale is much shorter than the total number of possible combinations. I have also had to introduce a couple of extra rules for the combination of factors:

1. Since neurons can be used to simulate the behaviour of neurons, or the molecules/atoms/ions that neurons are composed of, the functional granularity 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 molecules/atoms/ions, then the neurons that are used for this cannot themselves have their functions implemented using non-biological structures of neurons.
2. When machines have less functional granularity than the brain some kind of penalty needs to be imposed on machines that deviate from the human structure – for example, when the function of whole brain is implemented by a complex lookup table. In the present implementation, the number of levels of a human brain is 4 and so I will use 0.1 as the weighting for each missing level of functional granularity.

This 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 paper to list all the possible combinations of factors in a single ordinal scale – the complete list has over a million combinations. Instead, I have integrated everything together on a webpage,[2] which can be used to calculate the position of a machine on the scale. Some examples are given in the next section.

4.4 Examples

None of the systems discussed in this section are even close to reproducing the global functions of the human brain. However, to illustrate how this scale could work in practice, I will assume that these examples have developed to the point at which they could pass the T3 version of the Turing test.

4.3.1 Neurally Controlled Animat

This is a system developed by DeMarse et al. (2001) that uses biological neurons to control a simulated body in a virtual world. The biological neurons are initially disassociated 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: R1, S1 FW3, FN1 and FMAI1, giving a total weighting of 0.4, This needs to be multiplied by 0.1 to compensate for the lack of functional granularity at the level of modules and so the total weighting is 0.04, which works out as an ordinal ranking of 48 out of 812.

4.3.2 Lucy

Lucy is a robot developed by Grand (2003) that is controlled by a multi-processor simulation of neurons arranged into a biological structure. The factors are thus R1, S1, FW1, FM1, FN4, STS2 and AD3 giving a total weighting of 5.5×10^{-3}. This needs to be multiplied by 0.1 to compensate for the lack of functional granularity at the level of molecules, atoms and ions, and so the total weighting becomes 5.5×10^{-4}. This gives Lucy an ordinal ranking of 285 out of 812.

4.3.3 IDA

IDA is a naval dispatching system created by Franklin et. al. (1998). This system is based on Baars (1988) global workspace model of consciousness and so its modules could be said to be biologically structured. However the solutions that are used to implement the different modules are non-biological. The factors are R1, S1, FW1, FM4, STS2 and AD3. This gives a total weighting of 5.5×10^{-3}, but since the functional granularity is less than

[2] http://www.syntheticphenomenology.net

the human brain by two levels, this weighting needs to be multiplied by 0.01, to give a total weighting of 5.5×10^{-5}, which is an ordinal ranking of 461 out of 812.

4.3.4 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. This is a non-biological structure in which modules assembled from biological neurons are combined with modules built with other hardware. 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. One problem with Block's thought experiment is that the details about the functional implementation are left very vague and so I have classified it as multi-processor hardware combined with modules assembled from neurons simulated using biological neurons. The factors are: R3, S3, FW2, FM3, MSTS2, MAD3, FN3, FNN1 and FMA1, which gives a total weighting of 6.16×10^{-5}. This works out as an ordinal ranking of 445 out of 812. Although this seems surprisingly high, it is the presence of biological hardware (organised in a non-biological way) that elevates it above systems that are purely based on simulation. 1.3 billion computers networked together to produce the same result would have a ranking of 786 out of 812.

5 Discussion

A number of issues arise in connection with this probability scale:

1) To begin with, it is at present unclear whether consciousness decreases gradually as we move away from the human machine, or whether there is a cut off point at which consciousness simply vanishes. Consciousness may simply cease to exist in a system unless neurons are simulated at the molecular level, or a cluster of factors may interact in a critical way such that phenomenal states cannot be produced without one of the factors. If consciousness cuts off abruptly, then this ordinal probability scale expresses the likelihood that consciousness is present in a machine built in a particular way. On the other hand, if consciousness decreases gradually as the factors become less human, then this ordinal scale ranks machines according to their level of consciousness.

2) This is an extremely anthropocentric probability scale. 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 probability scale are more conscious than ourselves. This scale is a probabilistic rating based on our guess that machines built along lines similar to our own (such as other people) are more likely to experience phenomenal states than machines built along lines very different from our own. I believe that such a scale could be useful, but it should not be taken as anything more than the systematisation of an intuition.

3) This scale does not explicitly list many of the factors that have been put forward as potential correlates of consciousness. However, many of these are implicit possibilities within the available architectures. For example, re-entrant connections are assumed to be possible within any of the machines that have biologically structured neurons. However, there will inevitably be some factors that are not included within this version of the scale, which can be added to subsequent versions.

4) This scale only applies to machines whose global functions approximate those of the human brain. A perpendicular scale could be added that orders people, machines and animals according to the *degree* to which their global functions approximate those of the human brain. Machines with functions processing visual data about faces might be ranked higher on this scale than machines that analyse banking details. The more the system's global functions match those of the human brain, the more likely it would be to possess phenomenal consciousness (or the more phenomenal consciousness it would possess).

5) It is worth noting that I have set aside the whole question of the body here. In theory a computer could approximate the global input and output functions of the brain without inhabiting a body at all. However, such a system would be almost impossible to develop and, according to Damasio (1995), there may be a critical link between the body and consciousness.

6) Finally, this scale is likely to become superfluous when we eventually achieve machine consciousness. When we talk to robots every day, work with robots that display conscious behaviour and perhaps even marry robots with emotional functions, we will cease to worry about whether they *really* have phenomenal states; just as we rarely think that other people are automatons.

6 Previous Work

Synthetic phenomenology is an area that is only just starting to receive detailed attention. According to Chrisley (2004), the term first made its appearance in the machine consciousness community at the Models of Consciousness Workshop (held in Birmingham 2003) and was independently coined by Scott Jordan (1998). It is related to *synthetic epistemology*, which is defined by Chrisley and Holland (1994, p. 1) as the "creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world." Since this area is so new, relatively little research has actually been carried out on it. The work that has been done includes Chrisley's (1995) analysis of non-conceptual content and Holland and Goodman's (2003) use of inversion to map out a robot's internal representations.

The question about phenomenal states in robots has been extensively discussed in the literature on consciousness. The contributions roughly divide into those who accept the difficulties with behaviour-based attribution of phenomenal states, and those with 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 will need to attribute qualia to robots in order to understand their actions. A similar position is set out by Harnad (2003), who accepts the behaviour-based arguments set out by Moor and Prinz, but claims that the other minds problem means that we can only ever attribute 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 paper 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 ordinal probability scale set out in this paper 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 limited 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 levels of functional granularity could be simulated and 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 Dunmall (2003), who suggests five necessary conditions or axioms for consciousness. According to Aleksander and Dunmall, machines can only be conscious if they have depiction, imagination, attention, planning and emotion.

7 Conclusion

In this paper I have set out a proposal for an ordinal probability scale, which can be used to assess the likelihood that a machine is capable of experiencing phenomenal states. This scale only applies to machines that can pass the T3 version of the Turing test by controlling a human or artificial body. This scale can help us to evaluate the ethical significance of machine consciousness experiments and in some cases it could be used to select a machine implementation that has less probability of phenomenal suffering. The scale put forward in this paper is only a first draft with some of the factors that may be correlates of consciousness. If it is found useful, I hope that it will be improved by other people and perhaps develop into a standard as we get closer to realising conscious machines.

Acknowledgements

Many thanks to Owen Holland for feedback and comments about this paper. Thank you also to the EPSRC for funding this project.

References

Igor Aleksander and Barry Dunmall, An extension to the hypothesis of the asynchrony of visual consciousness. *Proceedings of the Royal Society of London B*, 267: 197-200, 2000.

Igor Aleksander and Barry Dunmall. Axioms and Tests for the Presence of Minimal Consciousness in Agents. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

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

Ned Block. Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, Volume IX, *Perception and Cognition Issues in the Foundations of Psychology*, edited by C. Wade

Savage, Minneapolis: University of Minnesota Press, 1978.

David Chalmers. *The Conscious Mind*. Oxford: Oxford University Press, 1996.

Ronald J. Chrisley and Andy Holland. Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity. *COGS CSRP*, 353: 1-21, 1994.

Ronald J. Chrisley. Taking Embodiment Seriously: Nonconceptual Content and Robotics. In Kenneth M. Ford, Clark Glymour, & Patrick J. Hayes (eds), *Android Epistemology*, Menlo Park/ Cambridge/ London: AAAI Press/ The MIT Press, 1995.

Ronald J. Chrisley. Synthetic Phenomenology. Talk at the Workshop on Machine Consciousness, Antwerp, 28th June 2004.

Francis Crick and Christof Koch. A framework for consciousness. *Nature Neuroscience*, 6(2): 119-26, 2003.

A. R. Damasio. *Descartes' Error: emotion, reason and the human brain*. London : Picador, 1995.

S. Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness : Basic evidence and a workspace framework. *Cognition*, 79: 1-37, 2001.

T. B. DeMarse, D. A. Wagenaar, A. W. Blau, and S. M. Potter. The Neurally Controlled Animat: Biological Brains Acting With Simulated Bodies. *Autonomous Robots,* 11(3): 305-310, 2001.

Stan Franklin, A. Kelemen and L. McCauley. IDA: a cognitive agent architecture. *IEEE International Conference on Systems, Man, and Cybernetics*, 3: 2646–2651, 1998.

Claudio Galletti and Piero Paolo Battaglini. Gaze-Dependent Visual Neurons in Area V3A of Monkey Prestriate Cortex. *The Journal of Neuroscience,* 9(4): 1112-1125, 1989.

Steve Grand. *Growing up with Lucy*. London: Weidenfeld & Nicolson, 2003.

Stuart Hameroff, and Roger Penrose. Orchestrated Reduction Of Quantum Coherence In Brain Microtubules: A Model For Consciousness? In S.R. Hameroff, , A.W. Kaszniak, and A.C. Scott (eds), *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, Cambridge, MA: MIT Press, 507-540, 1996.

Stevan Harnad. Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* 1(3): 1994.

Stevan Harnad. Can a Machine Be Conscious? How? In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

Owen Holland and Rod Goodman. Robots With Internal Models. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

J. S. Jordan. Synthetic phenomenology? Perhaps, but not via information processing. Talk given at the Max Planck Institute for Psychological Research, Munich, Germany, 1998.

Ernest W. Kent. *The Brains of Men and Machines*. Peterborough: BYTE/ McGraw Hill, 1981.

Thomas Metzinger. *Being No One*. Cambridge Massachusetts: The MIT Press, 2003.

J.H. Moor. Testing robots for qualia. In H.R. Otto and J.A. Tuedio (eds), *Perspectives on Mind*, Dordrecht/ Boston/ Lancaster/ Tokyo: D. Reidel Publishing Company, 1988.

Jesse J. Prinz. Level-Headed Mysterianism and Artificial Experience. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

V. S. Ramachandran and S. Blakeslee. *Phantoms in the Brain*. London: Fourth Estate, 1998.

J. Searle. Minds, Brains and Programs. *Behavioral and Brain Sciences,* 3: 417-57, 1980.

J. Searle. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press, 1992.

B. Van Heuveln, E. Dietrich and M. Oshima. Let's dance! The equivocation in Chalmers' dancing qualia argument. *Minds and Machines,* 8: 237-49, 1998.

S. Zeki and A. Bartels. The asynchrony of consciousness. *Proceedings of the Royal Society B*, 265: 1583-1585, 1998.

J. Zihl, D. Von Cramon and N. Mai. Selective Disturbance of Movement Vision after Bilateral Brain Damage. *Brain,* 106: 313-340, 1983.