3. MACHINE CONSCIOUSNESS[1]

3.1 Introduction

This chapter tackles some of the theoretical issues surrounding machine consciousness and reviews some of the previous work in this field.[2] Machine consciousness is currently a heterogeneous research topic that includes a number of different research programs, with some people working on the behaviours associated with consciousness, some people modelling the cognitive characteristics of consciousness and some people interested in creating phenomenal states in machines. To make sense of this diverse subject, the first part of this chapter identifies four different areas of machine consciousness research:

MC1. Machines with the external behaviour associated with consciousness.

MC2. Machines with the cognitive characteristics associated with consciousness.

MC3. Machines with an architecture that is claimed to be a cause or correlate of human consciousness.

MC4. Phenomenally conscious machines.

This classification starts with systems that replicate aspects of human[3] behaviour and moves on to systems that are attempting to create real artificial consciousness. Although there is a certain amount of overlap between these categories, they are a useful way of understanding work on machine consciousness and will be used to identify different aspects of it throughout this chapter.

[1] An earlier version of this chapter was published as Gamez (2007a).

[2] I will be using the term "machine consciousness" to refer to this field, although "artificial consciousness" and occasionally "digital sentience" (Anon, 2006) have also been used to describe it. Each of these terms has their own merits, but the growing number of meetings dedicated to "machine consciousness" suggests that this is likely to become the standard name for the field.

[3] In this chapter discussion generally focuses on *human* behaviour, cognitive characteristics and architectures associated with consciousness because humans are generally taken as paradigmatic examples of conscious entities. However, any work on the replication of animal behaviour, cognitive characteristics and architectures associated with consciousness would also be part of machine consciousness research.

The first application of these categories is to clarify the relationship between machine consciousness and other fields. The interdisciplinary nature of machine consciousness is often a source of confusion because it takes inspiration from philosophy, psychology, and neuroscience and shares many of the objectives of strong AI and artificial general intelligence. These relationships between machine consciousness and other fields become much clearer once machine consciousness has been separated into MC1-4. For example, artificial general intelligence has a certain amount in common with MC1, but little overlap with MC2-4. On the other hand, neuroscientists, such as Dehaene et al. (1998, 2003, 2005), are creating computer models of the neural correlates of consciousness (MC3), but have little interest in MC1, MC2 or MC4. This classification is also very useful for dealing with some of the criticisms that have been raised against machine consciousness, which often only apply to one or two aspects of its research. For example Dreyfus' (1992) claims about what computers still can't do mainly apply to MC1 and many of them could be answered by work on MC2 and MC3. On the other hand, Searle's Chinese Room argument is directed against MC4 and leaves work on MC1-3 unaffected.

The second half of this chapter surveys some of the research projects that are taking place in machine consciousness and uses MC1-4 to unpack the different objectives of this work. This research includes theoretical approaches, models of consciousness, and systems designed to actually be phenomenally conscious. The last two sections cover some of the ethical issues linked to machine consciousness and explore its potential benefits.

3.2 Areas of Machine Consciousness Research

Machine consciousness is not a unified field with a set of clearly defined goals. At present a heterogeneous network of researchers are working on different aspects of the problem, and this

can often make it difficult to understand how everything fits together. This section clarifies machine consciousness research by dividing it into four different areas.

3.2.1 Machines with the External Behaviour Associated with Consciousness (MC1)

A lot of our waking behaviours are carried out unconsciously in response to stimulation from the environment. For example, the detailed muscle contractions involved in walking are rarely under conscious control and we can perform relatively complex behaviours, such as driving home from work, with our attention on other things.[4] Other examples of unconscious behaviour include patients in a persistent vegetative state, who commonly produce stereotyped responses to external stimuli, such as crying, grimacing or occasional vocalisation (Laureys et al., 2004), blindsight patients who have a limited ability to respond visually to objects that they cannot consciously see, and actions carried out under the influence of an epileptic seizure. A dramatic example of the latter is given by Damasio (1999):

> Suddenly the man stopped, in midsentence, and his face lost animation; his mouth froze, still open, and his eyes became vacuously fixed on some point on the wall behind me. For a few seconds he remained motionless. I spoke his name but there was no reply. Then he began to move a little, he smacked his lips, his eyes shifted to the table between us, he seemed to see a cup of coffee and a small metal vase of flowers; he must have because he picked up the cup and drank from it. I spoke to him again and again he did not reply. He touched the vase. I asked him what was going on and he did not reply, his face had no expression. … Now he turned around and walked slowly to the door. I got up and called him again. He stopped, he looked at me, and some expression returned to his face – he looked perplexed. I called him again and he said, "What?"
>
> (Damasio, 1999, p. 6).

These examples show that a limited amount of behaviour can be carried out unconsciously by humans. However, the stereotypical nature of this behaviour suggests that

[4] For another view on this issue see Franklin et. al. (2005).

more complex activities, such as interpersonal dialogue, can only be carried out consciously and many new behaviours have to be learnt when consciousness is present. This leads to a distinction between human behaviours associated with consciousness and those carried out automatically without consciousness.[5]

One research area in machine consciousness is on systems that replicate conscious human behaviour. Although this type of research can be based on cognitive models (MC2) or on an architecture associated with consciousness (MC3), this is not necessary to work on MC1, which could also use a large lookup table or first-order logic to generate the behaviour. Although certain external behaviours are associated with phenomenal states *in humans*, this is not necessarily important to people working on MC1, since it has often been claimed that a zombie robot could replicate conscious human behaviour without experiencing phenomenal states. However, the boundary between MC1 and MC4 might start to become blurred when robots can reproduce most human behaviours. In this case, Harnad (2003) argues that we will have to attribute phenomenal experiences to MC1 machines because our only guide to phenomenal states is a system's external behaviour. Supporting this point, Moor (1988) suggests that we will need to ascribe qualia to such systems in order to understand them.

Any attempt to pass the Turing Test has to replicate behaviours that are carried out consciously in humans, and so people working on this challenge[6] can be considered to be part of MC1. Research on artificial general intelligence (see Section 3.3.2) also falls within this area.

3.2.2 Machines with the Cognitive Characteristics Associated with Consciousness (MC2)

A number of connections have been made between consciousness and cognitive characteristics, such as imagination, emotions and a self - for example, Aleksander's (2005) axioms and

[5] See Section 2.7.2 for a more detailed discussion of this issue.

[6] For example, the contestants in the annual Loebner prize: http://www.loebner.net/Prizef/loebner-prize.html.

Metzinger's (2003) constraints (see sections 2.6.3 and 2.6.4). Detailed descriptions of conscious states have also been put forward by phenomenologists, such as Husserl (1964), Heidegger (1995a) and Merleau-Ponty (1995).

The modelling of the cognitive characteristics associated with consciousness has been a strong theme in machine consciousness, where it has been carried out in a wide variety of ways, ranging from simple computer programs to systems based on simulated neurons. Cognitive characteristics that are frequently covered by this work include imagination, emotions, and internal models of the system's body and environment. In some cases the modelling of cognitive states has aimed at more realistic conscious behaviour (MC1) or used an architecture associated with consciousness (MC3), but MC2 systems can also be created without MC1 or MC3 – for example, a computer model of emotions or imagination that does not have external behaviour. There is also no necessary connection between MC2 and MC4 since the simulation of fear, for example, can be very different from real phenomenological fear - just as the price of gold can be modelled in a computer without the program, CPU or RAM containing any real gold.

3.2.3 Machines with an Architecture that is Claimed to be a Cause or Correlate of Human Consciousness (MC3)

Many people are working on the simulation of architectures that have been linked to human consciousness, such as Baars' (1988) global workspace. This type of research often arises from the desire to model and test neural or cognitive theories of consciousness and it is one of the most characteristic areas of machine consciousness.

Work on MC3 overlaps with MC2 and MC1 when systems based on an architecture associated with consciousness are used to produce the cognitive characteristics of consciousness or conscious behaviour. It could also overlap with MC4 if it was thought that an implementation of an architecture associated with consciousness would be capable of phenomenal states.

However, simulating a 'conscious' architecture in a machine may not be enough for the machine to actually become conscious.

3.2.4 Phenomenally Conscious Machines (MC4)

The first three approaches to machine consciousness are all relatively uncontroversial, since they are modelling phenomena linked to consciousness without any claims about real phenomenal states. The fourth area of machine consciousness is more philosophically problematic, since it is concerned with machines that have real phenomenal experiences - machines that are not just tools in consciousness research, but actually conscious themselves.

As has already been indicated, this approach has some overlap with MC1-3, since in some cases it might be hypothesized that the reproduction of human behaviour, cognitive states, or internal architecture would lead to real phenomenal experiences. On the other hand, MC4 might be achievable independently of other approaches to machine consciousness. For example, it might be possible to create a system based on biological neurons that was capable of phenomenal states, but lacked the architecture of human consciousness and any of its associated cognitive states or behaviours.[7] Furthermore, it has been claimed by Chalmers (1996) that systems as simple as thermostats may have basic conscious states. If this is correct, the presence of phenomenal states in a machine will be largely independent of the higher level functions that it is carrying out.

Systems with real consciousness cannot be developed without methods for measuring phenomenal states, and so there is a close relationship between MC4 and synthetic phenomenology (see Chapter 4). The production of machines with real feelings also raises ethical questions, which are covered in Section 3.6.

[7] DeMarse et al.'s (2001) neural animat might be a system of this kind.

3.3 Relationship between Machine Consciousness and Other Areas

3.3.1 Strong and Weak AI

Work on artificial intelligence is often classified using Searle's (1980) distinction between strong and weak AI:

> According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind: rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.
>
> (Searle, 1980, p. 417)

According to Searle, strong AI is the attempt to create something that is a mind in the sense that I am a mind, whereas weak AI is the process of modelling the mind using human-interpretable symbols that work in the same way a mind works. This distinction is similar to that made by Franklin (2003) between phenomenal and functional consciousness and it also relates to the difference between the easy and the hard problems of consciousness (Chalmers, 1996). In all of these cases, a contrast is set up between the external manifestations of a mind and a real mind, which suggests a reasonably clear mapping between MC4 and strong AI, with MC1-3 being examples of weak AI in Searle's sense.

The problem with strict identity between MC4 and strong AI is that the notion of mind can be separated from phenomenal consciousness - suggesting that computers can really *be* minds without being conscious in the sense of MC4. For example, Carruthers claims that "The view that we have, or can have, notions of mentality which do not presuppose consciousness is

now widely accepted" (Carruthers 2000, p. xviii), and so it may be possible to build a strong AI machine that is not conscious in the sense of MC4. A robot that grounded its symbols in sensory data might be one example of a non-phenomenal mind that literally understands and has other cognitive states.

3.3.2 Artificial General Intelligence

Artificial general intelligence (AGI) is another area within AI that has similarities with machine consciousness. The aim of AGI is to replicate human intelligence completely and it is sometimes contrasted with a second interpretation of weak AI as the solving of computer science problems within a limited domain – for example, pattern recognition or chess playing.[8] AGI has a certain amount of overlap with MC1, with the difference that MC1 is focused on conscious human behaviour, whereas AGI is attempting to reproduce all human behaviours linked with intelligence. Which of these is the larger category depends to some extent on the definition of intelligence. Some behaviours linked to consciousness may be excluded by AGI's definition of intelligence, but it is also possible that AGI could use a broad interpretation of intelligence that includes all MC1 behaviours.[9]

How AGI could be implemented is a completely open question and some AGI systems may be produced by simulating the cognitive states associated with consciousness (MC2) or by copying an architecture linked to consciousness (MC3). It is also possible that AGI systems will have phenomenal states (MC4). The interpretation of weak AI as the solving of computer science problems within a limited domain does not have much in common with any of the definitions of machine consciousness.

---

[8] This interpretation of weak AI is also referred to as "narrow AI".

[9] More information about AGI can be found in Goertzel and Pennachin (2007) and in the proceedings from the 2006 AGIRI Workshop: http://www.agiri.org/forum/index.php?showtopic=23.

3.3.3 Psychology, Neuroscience and Philosophy

The empirical work carried out by experimental psychology and neuroscience often forms a starting point for the modelling work in machine consciousness, but there is generally little overlap between them. However, there are some exceptions to this trend, such as the research carried out by Krichmar and Edelman (2006) using the Darwin series of robots and Dehaene et al.'s (1998, 2003, 2005) modelling of neurons to test theories about attention and consciousness. Dehaene et al.'s work clearly fits within MC3 and will be covered in Section 3.5.6. On the other hand, although Krichmar and Edelman are modelling a reentrant neural architecture associated with consciousness, they do not explicitly link their Darwin work to consciousness, and so I have not included it in this chapter.[10]

Amongst the other disciplines, cognitive psychology and connectionism also build computer models of cognition, which leads to a substantial amount of overlap with MC2. However, this work is more general than that carried out by machine consciousness because it covers types of cognition that are not associated with conscious states. Although philosophy and AI have historically been linked through their common use of logic, this connection has declined in recent years with the atrophy of logic in both subject areas. The emergence of machine consciousness has changed this relationship and philosophy now provides a theoretical framework for MC1-4 and tackles ethical issues.

3.4 Criticisms of Machine Consciousness

3.4.1 The Chinese Room

The Chinese Room thought experiment consists of a person in a room who receives Chinese characters, processes them according to a set of rules and passes the result back out without

[10] Krichmar and Edelman's work is covered in the discussion of research on neural networks in Section 5.6.

understanding what the characters mean. This processing of characters could be used to create the external behaviour associated with consciousness, to simulate the cognitive characteristics of consciousness or to model a conscious architecture. However, Searle (1980) argues that in no case would the person processing characters in the room understand what is going on or have intentional states directed towards the objects represented by the Chinese characters. Although the Chinese Room might be able to *model* a mind successfully, it will never literally *be* a mind in the sense intended by MC4.

One response to this argument is based on the notion of symbol grounding. If the characters in the Chinese room could be linked to non-symbolic representations, such as images or sounds, then the system would understand what the symbols mean and have intentional states directed towards this meaning. According to Harnad "Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1)'iconic representations', which are analogs of the proximal sensory projections of distal objects and events, and (2)'categorical representations', which are learned and innate feature-detectors that pick out the invariant features of object and event categories from their sensory projections." (Harnad 1990, p. 335). Neural models have also been cited as a way of grounding higher level symbolic representations by connecting them to sensory inputs (Haikonen, 2003), and if the Chinese Room can be grounded effectively in some kind of non-symbolic lower level, then it can be said to understand the characters that it is manipulating.

A second reason why the Chinese Room argument is not fatal to MC4 is that brains and computers are both physical systems assembled from protons, neutrons and flows of electrons. Searle (2002) is happy to claim that consciousness is a causal outcome of the physical brain and so the question becomes whether the physical computer and the physical brain are different in a way that is relevant to consciousness. This can only be answered after we have done a lot more research on the correlates of consciousness. Until this work has been carried out, the Chinese

Room argument does not offer any *a priori* reason why the arrangement of protons, neutrons and electrons in a physical computer is less capable of consciousness than the arrangement of protons, neutrons and electrons in a physical brain.

3.4.2 Consciousness is Non-algorithmic

Machine consciousness has also been criticised by Penrose (1990, 1995), who claims that the processing of an algorithm is not enough to evoke phenomenal awareness because subtle and largely unknown physical principles are needed to perform the non-computational actions that lie at the root of consciousness: "Electronic computers have their undoubted importance in clarifying many of the issues that relate to mental phenomena (perhaps, to a large extent, by teaching us what genuine mental phenomena are *not*) … Computers, we conclude, do something very different from what *we* are doing when we bring our awareness to bear upon some problem." (Penrose 1995, p. 393). If consciousness does something that 'mere' computation cannot, then MC1-3 cannot be simulated by a computer and MC4 cannot be created in a computer.

The most straightforward response to Penrose is to reject his theory of consciousness, which is far from convincing and has been heavily criticised by Grush and Churchland (1995) among others. However, even if Penrose's theories about consciousness are correct, MC1-4 would continue to be viable research projects if they could develop an approach to machine consciousness that fits within his framework:

> I am by no means arguing that it would be necessarily impossible to build a genuinely intelligent *device*, so long as such a device were not a 'machine' in the specific sense of being computationally controlled. Instead it would have to incorporate the same kind of physical action that is responsible for evoking our own awareness. Since we do not yet have any physical theory of that action, it is certainly premature to speculate on when or whether such a putative device might be constructed. Nevertheless, its construction can still be contemplated

> within the viewpoint … that I am espousing …, which allows that mentality can eventually be understood in scientific though non-computational terms.
>
> (Penrose 1995, p. 393).

If Penrose is right, we may not be able to use algorithms to construct MC1-4 machines, but it might be possible to create some kind of quantum computer, which incorporates the physical mechanisms that are linked by Penrose to human consciousness.

3.4.3 What Computers Still Can't Do

Dreyfus (1992) put forward a number of arguments against artificial intelligence projects that attempted to reduce human intelligence to a large number of rules.[11] According to Dreyfus, this can never work because human intelligence depends on skills, a body, emotions, imagination and other attributes that cannot be encoded into long lists of facts. Dreyfus also criticises some of the approaches to AI that have emerged as alternatives to fact-based systems, such as interactive AI, neural networks with supervised learning and reinforcement learning.

These arguments affect work on the development of systems that are as intelligent as humans in real world situations. However, there is no reason why MC1-4 could not be pursued in a more limited way independently of this objective. For example, some of the behaviours that require consciousness in humans (MC1) could be created in a simple and non-general way, and imagination and emotion could be simulated (MC2) without the expectation that they will be able to work as effectively as human cognitive processes.[12] The modelling of architectures associated with consciousness (MC3) is largely independent of Dreyfus' objections and phenomenal consciousness (MC4) may be possible without the generality and complexity of human behaviour.

---

[11] Lenat's Cyc is a good example of this kind of system (Matuszek et al. 2006). More recently Bringsjord has been using logic-based artificial intelligence to control a four year old child in Second Life: http://www.sciencedaily.com/releases/2008/03/080310112704.htm.

[12] This is the case with the simple Khepera models described in Section 3.5.5.

It can also be argued that the work being carried out on imagination, emotions and embodiment in machine consciousness addresses some of the areas that Dreyfus claims to be lacking in current artificial intelligence. Furthermore, the human brain is itself a machine, and so biologically-inspired research on machine consciousness might eventually be able to solve Dreyfus' problems. However, all of this work is still at an early stage and it is far from clear whether MC1-4 devices will ever become intelligent enough to act and learn like humans in the real world.

3.5 Research on Machine Consciousness

The last few sections have outlined the different areas of machine consciousness, its relationship to other fields and the criticisms that could be raised against it. I will now move on to some of the research that has been carried out on MC1-4. In order to focus on the unique aspects of machine consciousness, this chapter will not include the large number of simulations that have been done as part of AI, connectionism and brain modelling, and theoretical work on consciousness will only be included if it deals explicitly with MC1-4. Although some of the projects have been organised under sub-headings to highlight general areas of machine consciousness research, it should be borne in mind that some systems could have been included in several sections – for example, IDA has a global workspace architecture and is also a software agent.

3.5.1 Aleksander's Kernel Architecture

Aleksander (2005) and Aleksander and Morton (2007c) have developed a kernel architecture that includes all five of Aleksander's axioms (see Section 2.6.3). This includes a perceptual module that depicts sensory input, a memory module that implements non-perceptual thought for planning and recall of experience, an emotion module that evaluates the 'thoughts' in the

memory module, and an action module that causes the best plan to be carried out (see Figure 3.1).

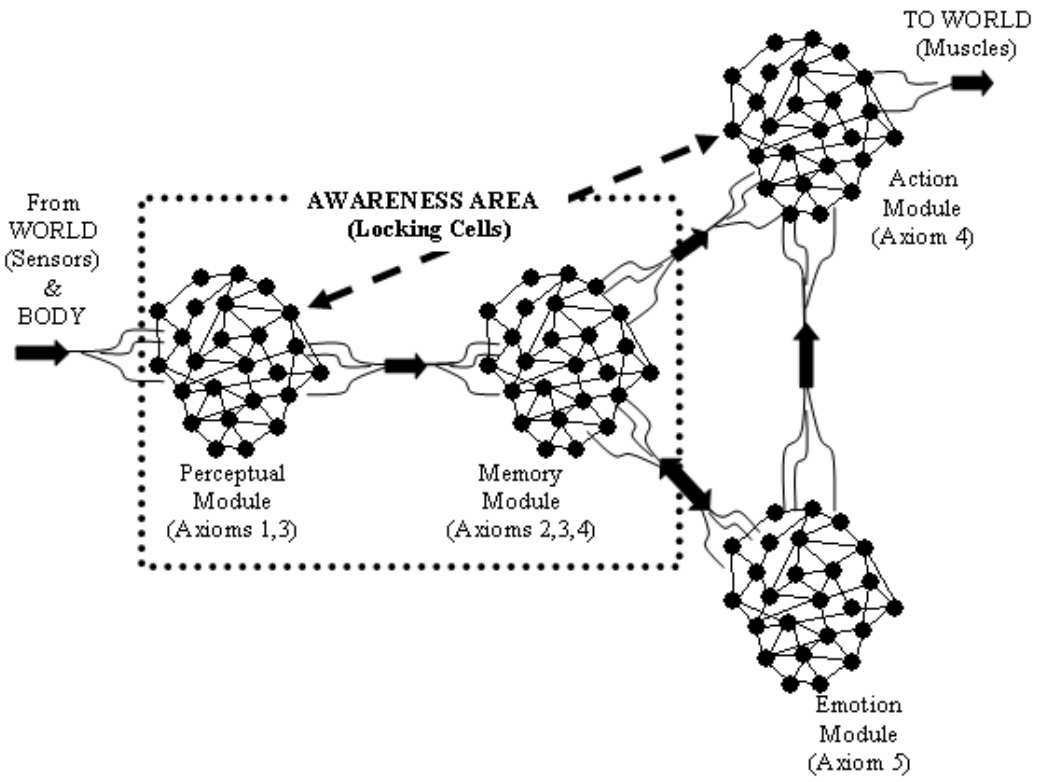

Figure 3.1. Aleksander's kernel architecture[13]

Aleksander and Morton (2007b) have built a number of brain-inspired implementations of this kernel architecture (MC3) using the Neural Representation Modeller (NRM) software,[14] which uses weightless neurons containing lookup tables that match input patterns to an output response. During training, these neurons store the link between each input pattern and the specified output; during testing, the neurons produce the output of the closest match to a known input pattern or a random sequence of 1s and 0s when there is more than one match. These neurons are assembled into large recurrent networks and trained using the graphical and scripting abilities of NRM.

[13] This figure is reproduced from Aleksander (2007c).

[14] This used to be called Magnus. More information about NRM is available at Barry Dunmall's website: http://www.iis.ee.ic.ac.uk/eagle/barry_dunmall.htm.

These brain-inspired simulations of the kernel architecture are minimal implementations of Aleksander's five axioms, and so they have the potential for phenomenal consciousness (MC4) according to the axiomatic theory. Full details about how the kernel architecture implements the axioms can be found in Aleksander and Morton (2007c).

3.5.2 Internal Modelling with SIMNOS and CRONOS

The CRONOS project and its main components were outlined in Section 1.2 and this thesis covers one of the approaches to machine consciousness that was developed as part of this project. A different approach to machine consciousness in the CRONOS project was developed by Holland, who claims that internal models play an important role in our conscious cognitive states (MC2) and may be a cause or correlate of consciousness in humans (MC4) (Holland and Goodman 2003, Holland 2007).[15] Holland is particularly interested in internal models that include the agent's body and its relationship to the environment and the extent to which the connection between this type of internal model and consciousness may be supported by Metzinger's (2003) discussion of the phenomenal self model and Damasio's (1999) analysis of the origins of consciousness. To test these theories about internal modelling, SIMNOS is being employed as an internal model of CRONOS and the computational technique of simultaneous localization and mapping (SLAM) will be applied to the visual stream from CRONOS's eye to obtain information about the environment and the robot's movements in relation to it, which will be used to continually update SIMNOS and its virtual environment. The internal model will then be employed offline to 'imagine' potential actions with SIMNOS before the selected action is carried out by CRONOS.

[15] Some of the other work carried out by Holland on the link between internal models and consciousness is described in Section 3.5.5.

3.5.3 Cog

Cog was a humanoid robot developed by Brooks et al. (1998) that consisted of a torso, head and arms under the control of a heterogeneous network of programs written in L, a multithreaded version of Lisp (see Figure 3.2). Cog was equipped with four cameras providing stereo foveated vision, microphones on each side of its head, and a number of piezoelectric touch sensors. This robot also had a simple emotional system to guide learning and a number of number of hard wired 'innate' reflexes, which formed a starting point for the acquisition of more complex behaviours. The processors controlling Cog were organised into a control hierarchy, ranging from small microcontrollers for joint-level control to digital signal processor networks for audio and visual processing.

The development work on Cog was organised as a number of semi-independent projects that focused on different aspects of human cognition and behaviour, such as joint attention and theory of mind, social interaction, dynamic human-like arm motion and multi-modal coordination. Although Brooks et al. (1998) do not explicitly situate this work within machine consciousness, Dennett (1997) put forward a good case for Cog having the potential to develop phenomenal states (MC4). Some of the behaviours of Cog, such as joint attention and theory of mind, could also be said to be associated with consciousness in the sense of MC1, and Cog's emotional system is a cognitive characteristic associated with consciousness (MC2).

Although Cog could display many individual human behaviours, when the systems were active together, competition for actuators and unintended couplings through the world led to incoherence and interference. This made it difficult for Cog to achieve higher cognitive functions and coherent global behaviour, which may be one of the reasons why this project has now effectively stopped.

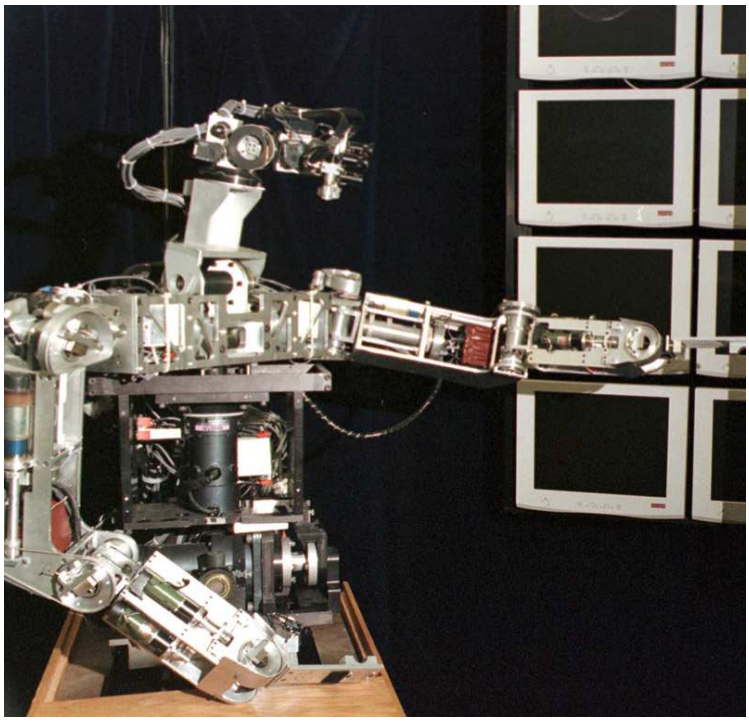

Figure 3.2. Cog robot[16]

3.5.4 CyberChild

CyberChild is a simulated infant controlled by a biologically-inspired neural system based on Cotterill's (2000) theory of consciousness. This virtual infant. (see Figure 3.3) has rudimentary muscles controlling the voice and limbs, a stomach, a bladder, pain receptors, touch receptors, sound receptors and muscle spindles. It also has a blood glucose measurement, which is depleted by energy expenditure and increased by consuming milk. As the consumed milk is metabolised, it is converted into simulated urine, which accumulates in the infant's bladder and increases its discomfort level. The simulated infant is deemed to have died when its blood glucose level reaches zero. CyberChild also has drives that direct it towards acquiring sustenance and avoiding discomfort and it is able to raise a feeding bottle to its mouth and control urination by tensing its bladder muscle. However, these mechanisms are not enough on their own to ensure the survival of the simulated infant, which ultimately depends on its ability to communicate its state to a human operator.

[16] Photograph taken by Donna Coveney.

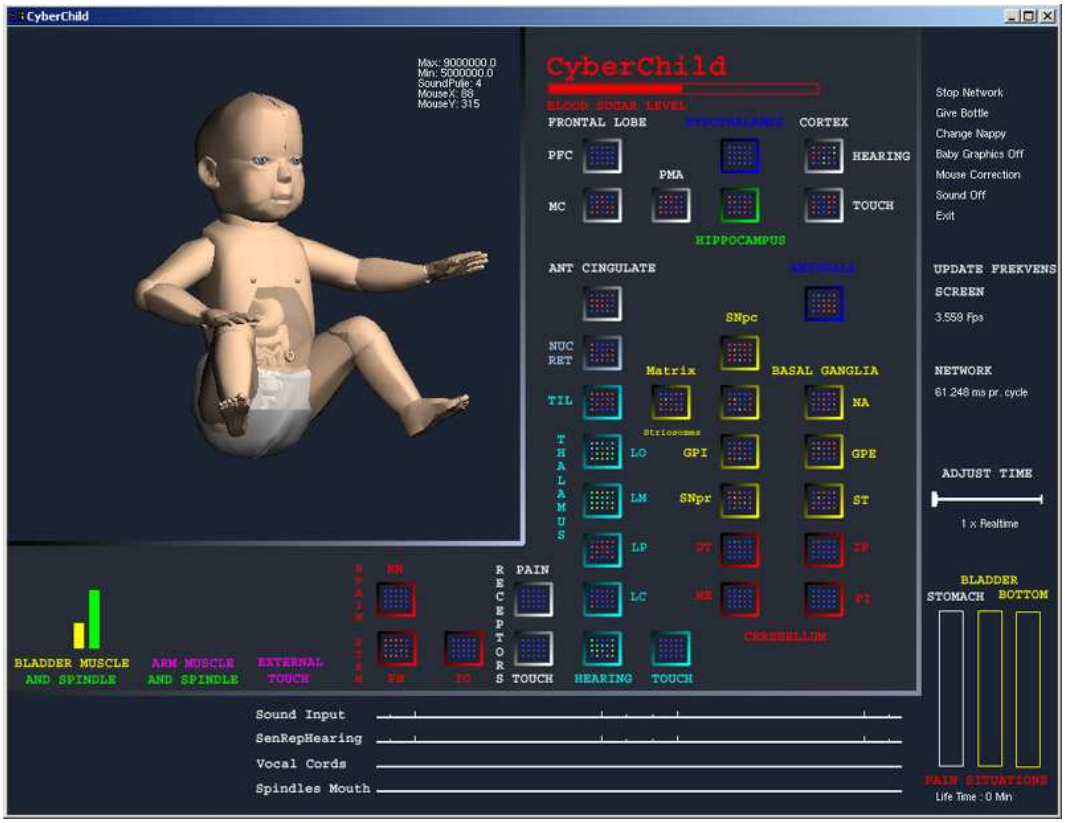

Figure 3.3. CyberChild

CyberChild is controlled by a simulated neural network containing a number of different areas based on the brain's neuroanatomy, including the premotor cortex, supplementary motor cortex, frontal eye fields, thalamic nuclei, hippocampus and amygdala. Each of these areas is modelled using twenty neuronal units and within each area about half of the units are active at any one time. Interconnection between the neural areas is based on the known anatomical connectivity of the brain and it includes efference copy connections from the premotor and supplementary motor cortices to sensory receiving areas, which Cotterill claims to be a vital feature of the neural processes underlying consciousness.

The overall aim of the CyberChild project was to use this detailed simulation to identify the neural correlates of consciousness (MC3) and perhaps even create phenomenal states (MC4). Cotterill (2003) planned to do this by looking for conscious behaviours (MC1), such as the

ability to modify communications with a human operator, which could be linked to the neural correlates of consciousness in the system.[17]

3.5.5 Simple Khepera Models

A number of researchers are using simulated or real Khepera robots (see Figure 3.4) to develop simple embodied systems containing analogues of the cognitive characteristics associated with consciousness. As these robots move around their environment they build up representations, which can easily be examined for internal models or imagination.

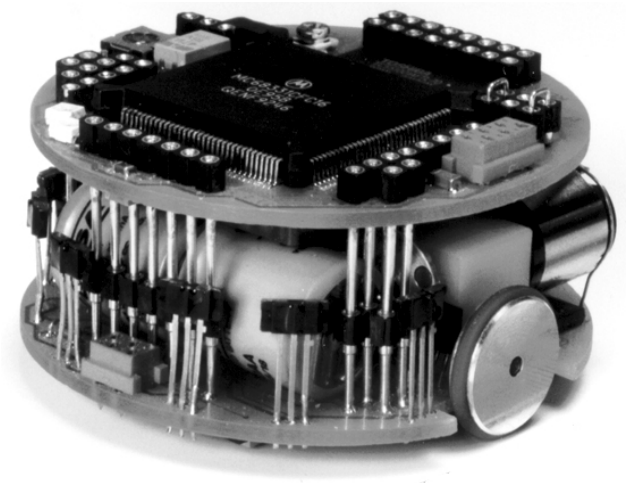

Figure 3.4. Khepera robot

Internal models

To test their ideas about the role of internal models in consciousness, Holland and Goodman (2003) used Linåker and Niklasson's (2000) Adaptive Resource-Allocating Vector Quantizer (ARAVQ) method to build models of the sensorimotor data from a Khepera robot. The ARAVQ approach is based on the observation that a robot's sensory input and motor output are often relatively stable over time - for example, when a robot is following a wall, its distance from the wall and speed remain approximately constant. Linåker and Niklasson's (2000) method takes advantage of this fact by regularly sampling a robot's sensory input and motor output and

[17] Sadly, Cotterill passed away in 2007 and it is unlikely that his work on CyberChild will be continued.

clustering this data using the ARAVQ on-line algorithm, which produces a small number of relatively stable and distinct combinations of sensory inputs and motor outputs called concepts. These concepts can be used to store long sequences of experiences very economically by labelling them and recording the number of times that each is repeated.

In their experiments, Holland and Goodman programmed a simulated Khepera with wall following and obstacle avoidance behaviour and allowed it to move around its environment while the ARAVQ method built up concepts corresponding to combinations of sensory input and motor output. Each concept represented the environmental features that activated the Khepera's rangefinders and how the robot moved in response to this stimulus, and so it was possible to plot the movements step by step along with the range finder data to produce the map of the environment that was stored inside the robot – a process that Linåker and Niklasson call inversion. By inverting the Khepera's concepts in this way Holland and Goodman produced a graphical representation of the Khepera's internal model and then examined how it could be used to control the simulated robot. They discovered that an internal model formed by concepts could accurately control the robot, process novel or incomplete data, detect anomalies and inform decisions.

These experiments showed that internal models can be developed and studied in a simple system and that they have the potential to play a useful role in the behaviour of an organism. Some of the internal models in humans are integrated into conscious cognitive states, and so this work is an example of MC2. Although Holland and Goodman do not claim that their simple system was conscious, more complex systems with internal models could contain phenomenal states (MC4) if their theories about the link between internal modelling and consciousness are correct.

Imagination

Ziemke et al. (2005) carried out a number of experiments on imagination using a simulated Khepera robot. This robot was controlled by a simple neural network that was based around a sensorimotor module, which mapped sensory input to motor output, and a prediction module. An evolutionary algorithm was used to train the weights on the two modules, with the sensorimotor module being evolved first to avoid obstacles and perform fast straightforward motion, and the prediction module evolved to predict the sensory input of the next time step. When the robot received real sensory input it was controlled by the sensorimotor module alone; when the robot was 'blindfolded' so that it received no external sensory input, it was controlled by feeding the prediction module's predictions about the next sensory input into the sensorimotor module. During the testing phase, it was found that 'imagined' sensory inputs produced very similar behaviour to real sensory input, although the pattern of activation of the internal units was very different in the two cases. These experiments demonstrated that the cognitive characteristics associated with consciousness (MC2) could improve the performance of a robot.

Ziemke's approach was developed further by Stening et al. (2005), who replaced the low level neural networks used by Ziemke with Linåker and Niklasson's (2000) ARAVQ method,[18] which was used to identify combinations of sensory input and motor output that were relatively invariant over time. The concepts generated by this method were then fed into a neural network consisting of an input layer and a hidden layer that was trained to predict when the next concept would occur. During the experiments, the robot's behaviour was initially controlled by a pre-trained neural network that moved the simulated Khepera around its environment with simple right-hand following behaviour, whilst the ARAVQ method extracted the basic features of the environment. The neural network's predictions about the next concept were then fed back into its input layer, which enabled the neural network to internally simulate a sequence of concepts

[18] See the earlier discussion of ARAVQ for more information about this method.

without the need for external movement. Stening et. al. then 'inverted' this sequence of concepts to produce a graphical representation of the Khepera's 'imagination'. This work is an example of MC2 and also falls within synthetic phenomenology (see Chapter 4). Although Hesslow and Jirenhed (2007) discuss the potential consciousness of this type of system, it is not entirely clear whether they are referring to MC2 or MC4.

3.5.6 Global Workspace Models

Global workspace theory is an influential interpretation of consciousness that was developed by Baars (1988). The basic idea is that a number of separate parallel processes compete to place their information in the global workspace, which is broadcast to all the other processes. A number of different types of process are used to analyse information or carry out actions, and processes can also form coalitions that work towards a common goal. These mechanisms enable global workspace theory to account for the ability of consciousness to handle novel situations, its serial procession of states and the transition of information between consciousness and unconsciousness. A substantial amount of work has also been done connecting the global workspace architecture to the thalamo-cortical system in the brain (Newman et al., 1997).

IDA naval dispatching system

Franklin's (2003) IDA naval dispatching system was created to assign sailors to new billets at the end of their tour of duty. This task involves natural language conversation, interaction with databases, adherence to Navy policy and checks on job requirements, costs and sailors' job satisfaction. These functions are carried out using a large number of codelets[19] that are specialised for different tasks and organised using a global workspace architecture.

[19] A codelet is a special purpose, relatively independent mini agent that is typically implemented as a small piece of code running as a separate thread. These codelets correspond with processors in global workspace theory.

The apparatus for ‘consciousness’ consists of a coalition manager, a spotlight controller, a broadcast manager and a number of attention codelets. These attention codelets watch for an event that calls for conscious intervention, and when this occurs they form a coalition with codelets containing data about the situation and compete for the spotlight of consciousness. If the coalition wins, its contents are broadcast to the other codelets, which may eventually choose an action that resolves the issue. The selection of behaviours in IDA is controlled by drives that award activation to behaviours that satisfy them, with activation spreading from behaviour to behaviour along excitatory and inhibitory links until an action is chosen. A model of deliberation is also included, which explores different scenarios and selects the best, and the architecture contains emotions, such as guilt at not getting a sailor’s orders out on time, frustration at not understanding a message and anxiety at not be able to convince a sailor to accept a suitable job. A number of different learning mechanisms are also implemented.

IDA is an example of a system that produces behaviour requiring consciousness in humans (MC1) and its architecture has some of the cognitive characteristics associated with consciousness (MC2), such as attention, emotions and imagination. All of this is produced by an architecture linked to human consciousness (MC3), and although Franklin thinks that IDA is unlikely to be phenomenally conscious (MC4), he does not entirely rule this out.

Dehaene et. al.’s neural simulations of the global workspace

Dehaene et. al. (1998) created a neural simulation to study how a global workspace and specialised processes interact during the Stroop task.[20] Their neural model included input and response units, global workspace neurons and vigilance and reward systems that modulated the activity in the global workspace. This simulation demonstrated that tasks that were easy for the system could be accomplished by local specialised processes without sustained activation in the

[20] In the Stroop task a subject is presented with a series of cards and has to state either the colour name that is printed on the card or the colour of the ink. This task is harder when the ink’s colour does not match the colour name, for example when “red” is printed in blue ink.

global workspace. On the other hand, tasks that were difficult for the model to accomplish, such as naming the colour of the ink when this conflicted with the colour name, could only be done by activating the global workspace and using the reward and vigilance systems to correct errors. Dehaene et. al. (1998) used this model to make predictions about brain imaging patterns generated during a conscious effortful task and about the pharmacology and molecular biology of the brain.

More recent work by Dehaene et. al. (2003) studied the attentional blink,[21] which they explained using their theory about the implementation of a global workspace in the brain. When the first target is presented to the subject, it gains access to the brain's global workspace by generating long range activations between many different neural areas and when the brain is in this state it is much harder for the second target to globally broadcast its information. Although local areas continue to carry out low level sensory processing on the second target, this does not become conscious because it cannot access the brain areas that are responsible for memory and reporting. Dehaene et al. tested these ideas about global workspace theory using a detailed neural simulation and compared their results with human subjects tested on the same experiment. Dehaene and Changeux (2005) have also used neural simulations to explore the role of spontaneous activity in workspace neurons and how this affects phenomena related to consciousness, such as inattentional blindness and transitions between the awake state and sleep, anaesthesia or coma.

Although the main emphasis of this work is on neuroscience, it closely ties in with theories about consciousness and Dehaene et al.'s neural models of global workspace theory are examples of MC3, even if they are not explicitly situated within machine consciousness. Their

[21] An attentional blink occurs in human subjects when two targets are presented in succession with 100-500 ms between them. Under these conditions the subject's ability to detect the second target is reduced, as if their attention had blinked after processing the first target.

models also fall within MC2 since they capture the fact that conscious experiences move through a serial progression of states with a limited content.

Shanahan's brain-inspired global workspace models

Shanahan (2006) developed a brain-inspired cognitive architecture based on global workspace theory, which was built using components that are functionally analogous to structures in the brain. At the bottom level of this system a sensorimotor loop made an immediate motor response to its situation, and on top of this a higher-order loop modulated the behaviour of the first order loop by adjusting the saliency of its actions. The first-order loop was closed through its interactions with the world, whereas the second-order loop was internally closed through an association area, which simulated the sensory stimulus that followed from a motor output in a way that was analogous to imagination. This simulation function was carried out using a global workspace architecture in which association areas received information from the basal ganglia analogue and competed to pass their information back to the basal ganglia analogue, which selected the next set of information to be broadcast. This architecture enabled the system to follow chains of association and explore the potential consequences of its actions prior to carrying them out.

In his experimental setup Shanahan (2006) used NRM[22] to create the neural simulation and the robot simulator Webots to simulate a Khepera robot with a camera. This system was programmed with a small suite of low level actions and trained to have positive and negative preferences for cylinders with different colours. Using its global workspace architecture the robot could explore the consequences of potential actions and give a low weighting to actions that would bring about an aversive stimulus. This enabled it to select actions that were more 'pleasant' than the ones that it would have chosen using the simple sensorimotor loop. This system is an example of MC1-3 since it is using imagination and emotion (MC2) implemented in

[22] See the brief discussion of NRM in Section 3.5.1.

a global workspace architecture (MC3) to produce behaviour that requires consciousness in humans (MC1). Although Shanahan claims that his system respects all five of Aleksander's axioms, he is cautious about attributing real phenomenal consciousness to it.

In more recent work, Shanahan (2008) built a global workspace model using simulated spiking neurons, which was based on the work by Dehaene et. al. (1998, 2003, 2005). This showed how a biologically plausible implementation of the global workspace architecture could move through a serial progression of stable states, and it had the potential to carry out the same function as the core circuit described in Shanahan (2006). Unlike the earlier model, it did not exhibit external behaviour, and so it is an example of MC2-3.

Neural schemas

The neural schema approach developed by McCauley (2002) is a neural and connectionist implementation of some aspects of global workspace theory. This system is based on a network of nodes that represent the state of the environment, actions, the effect of actions and the goals of the system, and the level of activation of these nodes can spread along the links between them. There is also a model of attention and consciousness based on global workspace theory, which allocates 'consciousness' to nodes based on their change in activation over time, their ability to accomplish current goals and their association with other nodes recently involved in 'consciousness'. This 'consciousness' of the nodes alters their behaviour and the information in them is broadcast across the network. This system is described by McCauley as an implementation of part of a psychological theory of consciousness (MC2-3), and not as something that displays true consciousness.

3.5.7 Language and Agency

Agent-based conscious architecture

Angel (1989) sets out a language- and agent-based architecture for a conscious machine centred around three attributes that must be possessed by any conscious system:

1. Independent purpose regardless of its contact with other agents.
2. The ability to make interagency attributions on a pure or natural basis.
3. The ability to learn from scratch significant portions of some natural language, and the ability to use these elements in satisfying its purposes and those of its interlocutors.

According to Angel, these behavioural attributes associated with consciousness (MC1) can only be used to infer real phenomenal states in a machine (MC4) if human consciousness is a physical phenomena that conforms to physical laws. If human consciousness can somehow pre-empt or transgress natural causes, then we cannot attribute consciousness to entities using these criteria.

Since Angel's attributes are based on language and agency, it is not difficult to produce formal models of them on a computer, and Angel suggests how a machine could be built that would actually be conscious (MC4) according to his criteria. This would lead to a minimally conscious system, which could be attributed more degrees of consciousness if it exhibited cognitive characteristics associated with consciousness (MC2), such as emotion, wakefulness, a sense of continuity with the past and an ego. As far as I am aware, there has not been any attempt to implement the architecture that Angel proposes, although the work of Steels (2003) points in this direction.

Inner Speech

According to Steels (2003), inner speech is linked to conscious experience through the role that it plays in our sense of self and agency. Steels' work on inner speech started with experiments in which two robotic heads watched scenes and played a language-game that evolved a lexicon or

grammar (Steels, 2001). In one language-game, a speaker chose an object in the scene and sought a verbal description so that the hearer could guess which object was chosen. In the early versions of these experiments it was relatively easy for the agents to develop a lexicon, but they could not evolve grammar until Steels applied the speaker's language system to its own utterances, either before transmitting them or after observing incomprehension in the listener. This model of inner speech enabled the agents to evolve case grammar and Steels (2003) suggests that it could be used outside of communication to rehearse future dialogue, submit thoughts to self criticism, and conceptualise and reaffirm memories of past experiences. All of these additional functions of inner speech could be the foundation of our sense of self and they could also play a role in our inter-agency relationships with others. Steel's modelling of inner speech is mainly directed towards reproducing important aspects of our conscious experience (MC2). Although Steels suggests that complex language production may have played a crucial role in the origin of consciousness, he leaves open the possibility that models of inner speech will lead to actual phenomenal states.

Other work on the link between inner speech and consciousness includes Clowes (2006, 2007), who argues that inner speech helps to organise conscious experience, direct attention and manage ongoing activities. These ideas were tested by Clowes and Morse (2005) in some simple experiments on the structuring of action by language. Haikonen (2006) also has a detailed discussion of the relationship between inner speech and consciousness.[23]

[23] Inner speech is an example of deliberation in the sense of Sloman (1999), which is implemented in Franklin's IDA naval dispatching system - see Franklin (2000) for more on the relationship between deliberation and IDA. Deliberation in the sense of a consciously evoked internal virtual reality is closely related to internal models and imagination, which appear in several of the projects covered by this chapter.

3.5.8 Cognitive Architectures

A Cognitive Approach to Conscious Machines

Haikonen (2003, 2006, 2007) is developing a system that is intended to display cognitive characteristics associated with consciousness, such as emotion, transparency, imagination and inner speech, using a detailed neural simulation. This cognitive architecture starts with sensory modules that process visual, auditory and tactile data into a large number of on/off signals that carry information about different features of the stimulus. Perceived entities are represented using combinations of these signals, which are transmitted by modulating a carrier signal (an important aspect of Haikonen's theory of consciousness). There is extensive feedback within the system and cross connections between different sensory modalities integrate qualitative characteristics carried by the signal with its location in motor space. Haikonen's architecture also includes emotions – for example, there is an analogue of pain, which uses information about physical damage to initiate withdrawal and redirect attention. In this architecture, language is part of the auditory system and the association of words with representations from other modalities enables sequences of percepts to be linguistically described. Haikonen (2006) claims that percepts become conscious when different modules cooperate in unison and focus on the same entity, which involves a wealth of cross-connections and the forming of associative memories.

If this system can be constructed, it will be an example of MC1-4 since it is attempting to produce behaviour and cognitive states linked to consciousness using an architecture theorized to be a cause or correlate of consciousness, which may actually become conscious. At the time of writing Haikonen is working on the implementation of his proposed architecture and it is not clear how much has been completed.

Schema-based model of the conscious self

Samsonovich and DeJong's (2005a,b) cognitive architecture is based around schemas that process data items, such as semantic knowledge, action primitives or sensory qualia. The behaviour of these schemas is constrained by a set of axioms that correspond to the system's 'conscious' self. These self axioms are beliefs that the agent holds about itself, such as the fact that the self is the only author of self-initiated acts, the self is indivisible, and so on. In Samsonovich and DeJong (2005b) this system was integrated using a dynamic multichart architecture, whereas in Samsonovich and DeJong (2005a) it was coordinated by contextual, conceptual and emotional maps based on the hippocampus. Samsonovich and DeJong (2005b) describe how this cognitive architecture was used to control a virtual robot that learnt to move in open space, navigate a maze and solve a simple push-push puzzle.

This cognitive model of the conscious self is an example of an MC2 system that is capable of behaviours that require consciousness in humans (MC1). Although Samsonovich and DeJong (2005a) map their architecture onto brain areas and functions, they do not explicitly link it to any of the architectures that have been put forward as a cause or correlate of human consciousness (MC3). Samsonovich and DeJong (2005a,b) do not comment on whether their system is capable of real phenomenal states (MC4).

Cicerobot

Cicerobot is a robot created by Chella and Macaluso (2006) and Chella (2007), which has sonar, a laser rangefinder and a video camera, and works as a museum tour guide in the Archaeological Museum of Agrigento (see Figure 3.5). The cognitive architecture of this robot is based around an internal 3D simulation, which is updated as the robot navigates around its environment. When the robot moves it sends a copy of its motor commands to the 3D simulator, which calculates expectations about the next location and camera image. Once the movement has been executed, the robot compares its expected image with the 2D output from its camera and uses discrepancies

between the real and expected images to update its 3D model. Cicerobot uses this 3D simulation to plan actions by exploring different scenarios in a way that is analogous to human imagination.

Figure 3.5. Cicerobot

This 'conscious' cognitive architecture (MC2) is used to control the robot in the unpredictable environment of a museum (MC1). Chella and Macaluso (2006) also link the robot's comparison between expected and actual perceptions to the presence of real phenomenological states (MC4).

3.5.9 Other Work

Other work on machine consciousness includes Mulhauser (1998), who used physics, computer science and information theory to outline how consciousness and a conscious self model could

be implemented in a machine. There is also Duch (2005), who sets out an architecture for a conscious system that is inspired by brain-like computing principles. This proposed system's claims to be conscious would be based on its interpretation of variations in its internal states as different feelings or qualia associated with the perceived objects. Bosse et al. (2005) have carried out simulations of Damasio's core consciousness using the Temporal Trace Language (TTL) (Jonker and Treur 2002) and a simpler variation called *leads to*. In their simulations dynamic properties of the neural processes leading to emotion, feeling and core consciousness were expressed using statements in TTL and *leads to* and executed within a custom built simulation environment that enabled temporal dependencies between different parts of the model to be traced and visualised. Other neural network models of consciousness include the CODAM model that links consciousness to a copy of the signal that changes the focus of attention (Taylor 2007, Taylor and Fragopanagos 2007), Ikegami's (2007) work with a mobile agent equipped with a Fitz-Hugh-Nagumo neural network, and Cleeremans et al.'s (2007) networks inspired by Rosenthal's (1986) higher-order thought theory. More theoretical work on machine consciousness can be found in Holland (2003), Chrisley et al. (2007) and Chella and Manzotti (2007).

3.6 Social, Ethical and Legal Issues

Many people believe that work on machine consciousness will eventually lead to machines taking over and enslaving humans in a Terminator or Matrix style future world. This is the position of Kaczynski (1995) and Joy (2000), who believe that we will increasingly pass responsibility to intelligent machines until we are unable to do without them - in the same way that we are increasingly unable to live without the Internet today. This would eventually leave us at the mercy of potentially super-intelligent machines that may use their power against us. Against these apocalyptic visions, Asimov (1952) agrees with Kaczynski and Joy about how the

machines will take over, but suggests that computers will run the world better than ourselves and actually make humanity happier.[24] A similar position is put forward by Sloman (2006), who argues that "It is very unlikely that intelligent machines could possibly produce more dreadful behaviour towards humans than humans already produce towards each other, all round the world even in the supposedly most civilised and advanced countries, both at individual levels and at social or national levels."

At present our machines fall far short of many aspects of human intelligence, and we may have hundreds of years to consider the matter before either the apocalyptic or optimistic scenarios come to pass. It is also the case that science fiction predictions tell us more about our present concerns than about a future that is likely to happen, and our attitudes towards ourselves and machines will change substantially over the next century, as they have changed over the last. For example, Kurzweil (2000) argues that as machines become more human and humans become more machinic, the barriers will increasingly break down between them until the notion of a *takeover* by machines makes little sense. Furthermore, as machines develop, the safety regulations will increase and we may be able to build a version of Asimov's laws into them, or at least exclude intense negative emotions such as hate or envy. At present, work on machine consciousness has many benefits (see Section 3.7) and it is not justified to call a halt to the whole program because of scare stories and science fiction visions.[25]

A second ethical dimension to work on machine consciousness is how we should treat conscious machines. As Torrance (2005) points out, we will eventually be able to build systems that are not just instruments for us, but participants with us in our social existence. However, this can only be done through experiments that cause conscious machines a considerable amount of confusion and pain, which has led Metzinger (2003) to compare work on machine consciousness

[24] Moravec (1988) was also an early advocate of this view.

[25] These ethical issues were discussed at length at the 2006 AGIRI Workshop: http://www.agiri.org/forum/index.php?showtopic=23.

to the development of a race of retarded infants for experimentation. We want machines that exhibit behaviour associated with consciousness (MC1) and we want to model human cognitive states (MC2) and conscious architectures (MC3), but we may have to *prevent* our machines from becoming phenomenally conscious (MC4) if we want to avoid the controversy associated with animal experiments. This can only be done by developing systematic methods for evaluating the likelihood that a machine can experience phenomenal states.[26]

A final aspect of the social and ethical issues surrounding machine consciousness is the legal status of conscious machines. When traditional software fails, responsibility is usually allocated to the people who developed it, but the case is much less clear with autonomous systems that learn from their environment. A conscious machine might malfunction because it has been maltreated, and not because it was badly designed, and so its behaviour could be blamed on its carers or owners, rather than on its manufacturers. Conscious machines could also be held responsible for their own actions and punished appropriately.[27] A detailed discussion of these issues can be found in Calverley (2005).

3.7 Potential Benefits of Machine Consciousness

This final section takes a look at some of the positive outcomes that might be realised through research on machine consciousness. Although research on MC1 is still at an early stage, it could eventually help us to produce more plausible imitations of human behaviour. In the shorter term, this might appear as more sophisticated chatterbots that carry out simple conversations as part of a telephone or web application. Progress with MC1 is most likely to come from research on other aspects of machine consciousness, such as MC2 or MC3.

---

[26] The ethical treatment of conscious machines is also discussed by Stuart (2003).

[27] Punishment might have to be limited to machines with some kind of self awareness if we want to avoid the absurdities of the criminal prosecution of animals in the Middle Ages – see Evans (1987).

One of the main benefits of research on MC2 will be the development of machines that can connect emotions with objects and situations, attend to different aspects of their environment, and imagine themselves in non-present scenarios.[28] This could eventually lead to machines that can understand our human world and language in a human-like way, which would vastly increase their ability to assist us and interact with us. Work on MC2 might also open up intersubjective possibilities between humans and machines, enabling computers to imagine what people might be thinking, empathize with them and imitate them.

At present, MC3 research is mainly oriented towards modelling the architectures that have been associated with human consciousness, which is an excellent way to test ideas about how consciousness works in human beings. When this modelling involves simulated neural networks, it can advance our understanding of the neural correlates of consciousness, as seen in the work of Shanahan (2006, 2008) and Dehaene et al. (1998, 2003, 2005). This neural modelling could improve our diagnosis of coma and locked-in patients and help us to understand how the brain processes information, so that we can develop prosthetic interfaces to restore visual, auditory or limb functions. MC3 work can also help us to develop machines that tackle problems in a similar way to humans, such as Franklin's naval dispatching system.[29]

Although we often want to avoid phenomenal states in machines, work on MC4 does have a number of potential benefits. The most important of these is the development of systematic ways of examining systems for signs of consciousness and making predictions about their phenomenal states. By working hand in hand with neurophenomenology, this synthetic phenomenology could lead to more scientific theories about animal suffering and it will be discussed in detail in the next chapter. Up to this point it has always been a vague question about whether, for example, snails feel pain, but MC4 research may eventually be able to make detailed predictions about the phenomenal states of non human systems. This could also help us

[28] Part of the work on deliberation – see footnote 23.

[29] See Franklin (2001) for more on how IDA tackles problems in a similar way to humans.

to understand the phenomenal states of very young or brain-damaged people who are incapable of communicating their experiences in language.

3.8 Conclusions

Machine consciousness is a relatively new research area that has gained considerable momentum over the last few years, and there is a growing number of research projects in this field. Although it shares some common ground with philosophy, psychology, neuroscience, computer science and even physics, machine consciousness is rapidly developing an identity and problems of its own. The benefits of machine consciousness are only starting to be realised, but work on MC2-3 is already proving to be a promising way of producing more intelligent machines, testing theories about consciousness and cognition, and deepening our understanding of consciousness in the brain. As machine consciousness matures it is also starting to raise some novel social and ethical issues.

One of the challenges in MC4 work on machine consciousness is to establish whether a system is capable of phenomenal states and to describe these phenomenal states when they occur. This challenge is addressed by the emerging discipline of synthetic phenomenology, which is covered in Chapter 4. Chapter 5 describes the design and implementation of an MC1, MC2 and potentially MC4 neural network, whose phenomenal states are analyzed in detail in Chapter 7.