1. INTRODUCTION

> The interdisciplinary project of consciousness research, now experiencing such an impressive renaissance with the turn of the century, faces two fundamental problems. First, there is yet no single, unified and *paradigmatic* theory of consciousness in existence which could serve as an object for constructive criticism and as a backdrop against which new attempts could be formulated. Consciousness research is still in a preparadigmatic stage. Second, there is no systematic and comprehensive catalogue of *explananda*. Although philosophers have done considerable work on the *analysanda,* the interdisciplinary community has nothing remotely resembling an agenda for research. We do not as yet have a precisely formulated list of explanatory targets which could be used in the construction of systematic research programs.
>
> (Metzinger 2003, pp. 116-7)

1.1 Overview

This PhD was carried out as part of Owen Holland's and Tom Troscianko's EPSRC-funded CRONOS project to build a conscious robot (GR/S47946/01), which took place at the Department of Computing and Electronic Systems, University of Essex and at the Department of Experimental Psychology, University of Bristol. One of the main contributions at Essex was the development of the CRONOS and SIMNOS robots, which are described in Section 1.2. This thesis documents my contribution to this project, which includes the construction of a spiking neural network to control SIMNOS's eye movements and the development of a new way of analyzing systems for consciousness that was used to make predictions about this network's phenomenal states. A summary of the thesis given in Section 1.3 and Section 1.4 describes the supplementary data files and other supporting materials.

1.2 The CRONOS Project [1]

1.2.1 Introduction

CRONOS is one of the few large projects that has been explicitly funded to work on machine consciousness. One of the motivations behind this project was the belief that embodied human-like systems carrying out tasks in the real world (or a reasonably realistic copy) are the best starting point for understanding how our brains operate and how consciousness emerges in the brain. Guided by this approach, Owen Holland, Rob Knight and Richard Newcombe developed CRONOS, a hardware robot closely based on the human musculoskeletal system (see Figure 1.1), and a soft real time physics-based simulation of this robot in its environment, known as SIMNOS (see Figure 1.2). More information about the CRONOS project is available at www.cronosproject.net.

1.2.2 CRONOS Robot

Most humanoid robots are essentially conventional robots that fit within the morphological envelope of a human. However, robots that can help us to understand human cognition and action might need to have a much higher level of biological inspiration, which imitates biological structures and functions as well as the human form. The CRONOS robot was developed to address this challenge and it has a body based on the human musculoskeletal system and senses that are as biologically inspired as possible.[2] This level of biological realism is important to machine consciousness because a more biological body is more likely to develop a human style of consciousness, and it also provides more realistic training data for biologically inspired neural networks.

---

[1] All of the work described in this section was carried out by Owen Holland, Rob Knight and Richard Newcombe at the University of Essex.

[2] Holland and Knight (2006) have proposed the term “anthropomimetic” as a label for humanoid robots that attempt to copy the physical structure of a human.

Figure 1.1. CRONOS Robot

To create the skeleton of CRONOS, the human skeleton was copied as accurately as possible at life size.[3] The bones were constructed from a new type of thermoplastic known in the UK as Polymorph and in the US as Friendly Plastic, which softens and fuses at 60 degrees and can be freely hand moulded until it resets at 30 degrees. This enabled bone like elements to be created and fitted together by hand and other materials can be embedded, such as a metal sphere mounted on a rod to make a ball and socket joint. The muscles of CRONOS were constructed using a motor and marine grade shock cord terminated at each end by 3mm braided Dyneema kite line. This cord was wound around the motor spindle, so that the rotation of the motor increased or decreased the tension in the elastic shock cord, mimicking the contraction and relaxation of a biological muscle.[4]

[3] To compensate for anticipated difficulties with the fine manual manipulation of grasped objects, the neck vertebrae were extended to allow a greater range of head movements during visual inspection of such objects.

[4] Videos of CRONOS are available at www.cronosproject.net.

This combination of bone-like elements and partially elastic ‘muscles’ gives the body of CRONOS a multi-degree-of-freedom structure that responds as a whole and transmits force and movement well beyond the point of contact. For example, when the arm is pushed down, the elbow flexes, the complex shoulder moves and the spine bends and twists. The disturbances due to the robot’s own movements are also propagated through the structure, producing what Holland et al. (2007) have called 'passive coordination'. Since different trajectories and finishing points are obtained with different loadings, any controllers that are developed for this robot will need feedforward compensation to anticipate and predictively cancel the effects of the load for any movement. This is interesting from the point of view of consciousness because feedforward control depends on the possession of forward models and the use of such models by the nervous system has been advanced by Grush (2004) and Cruse (1999) as one of the key factors underpinning consciousness.

CRONOS differs from humans in having only a single central eye. This approach was chosen because of the enormous simplification of visual processing that it brings about and it is justified by the observation that 2-4% of humans do not perform stereo fusion and their performance on other visual tasks is still within the normal range (Julesz, 1971). The high resolution colour camera has a 90 degree field of view and it can perform rapid saccades under the control of three servo motors that rotate, pan and tilt the eye. Each of the muscle motors has a potentiometer and touch sensors are being developed for the hands and stretch receptors for the tendons to give more realistic proprioceptive information An interface is also being developed that will allow CRONOS to stream its sensory data as spikes over the network and receive muscle commands as spikes from the network. This will be similar to the spike streaming between SIMNOS and SpikeStream that is described in Section 1.2.5.

1.2.3 SIMNOS Virtual Robot

SIMNOS is a model of CRONOS that was created to test Holland's (2007) theories about the link between consciousness and internal modeling and to accelerate the development of controllers for CRONOS. This model was created using physics-based rigid body modeling, implemented in Ageia PhysX,[5] in which the components of objects and surfaces are described in 3D by mathematical expressions in terms of their underlying physics, and the expressions are solved using extremely fast and efficient numerical techniques. This reliance on physics guarantees accuracy at all scales, and the efficiency of the computations allows thousands of complex objects interacting in real time to be modeled on a standard personal computer.

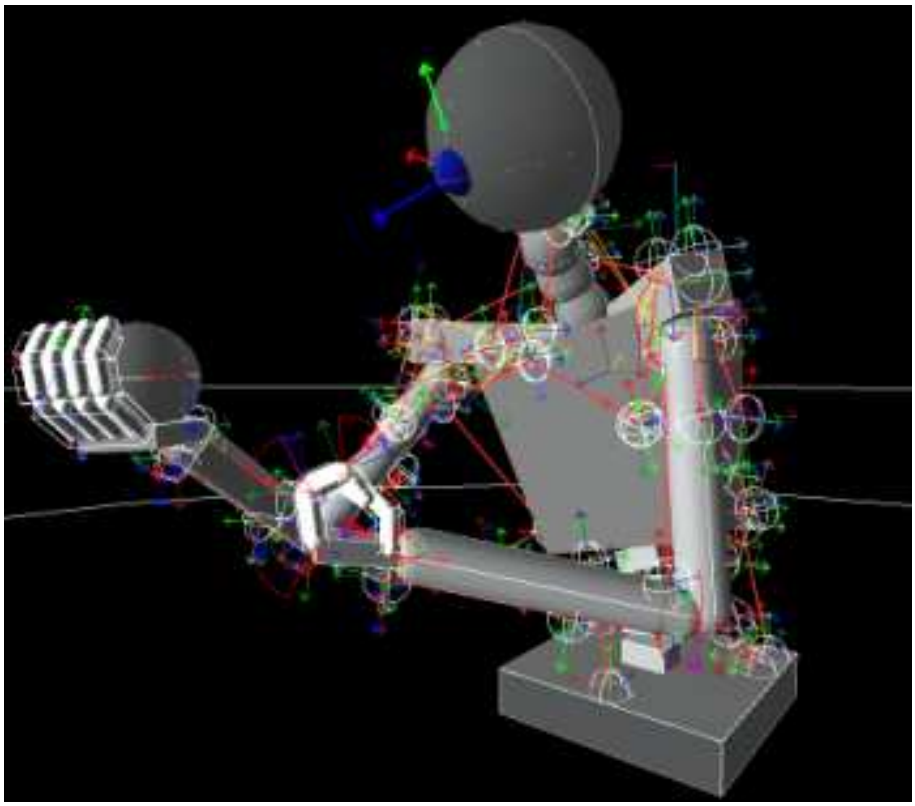

Figure 1.2. SIMNOS virtual robot. The red lines are the virtual muscles; the outlines of spheres with arrows are the joints. The length of the virtual muscles and the angles of the joints are encoded into spikes and sent to the SpikeStream neural simulator.

[5] Ageia PhysX: http://www.ageia.com/developers/api.html.

The individual components of CRONOS are modeled in SIMNOS using appropriate sizes and masses, but the shapes were simplified where possible – for example, detailed bone shapes were approximated by cylinders with the same dimensions and distribution of mass. The elastic actuators were created using springs of appropriate lengths connected to matching points on the modeled skeleton, and sufficient damping was added to produce the slight degree of under-damping seen on CRONOS. The virtual robot's environment contains rigid bodies that are either simple geometrical shapes or triangular meshes and new objects can be created using 3D simulation packages, such as Maya or Blender, and imported into SIMNOS using the COLLADA format.[6] In the future it will be possible to add cloth- and fluid-based objects to SIMNOS's virtual environment.

The SIMNOS model of CRONOS is convincing at the physical level and displays a similar quality of movement. The fluidity, load sharing and passive coordination in CRONOS are also seen in SIMNOS, which presents comparable control problems.

1.2.4 SIMNOS Performance

A simple virtual world was developed to test SIMNOS's computation time. The simulated scene was started with random parameters for every muscle and all of the sensory and motor data was calculated to ensure the maximum computational load. The simulator was then run for 3000 time steps and at each step a newly created sphere was dropped onto the surface of the table where the robot was fixed. As the objects fell onto the table and floor they interacted with the robot, the environment and each other.

The computation times for this virtual world were recorded for a number of different time step values and plotted in Figure 1.3. These results show that soft real time simulation of the robot in an environment with 300 objects, with full scene rendering for user output, is

[6] COLLADA format: www.collada.org.

possible for time step values greater than 1/50th second and this performance will improve substantially as more cheap physics processing hardware for the PhysX engine becomes available.

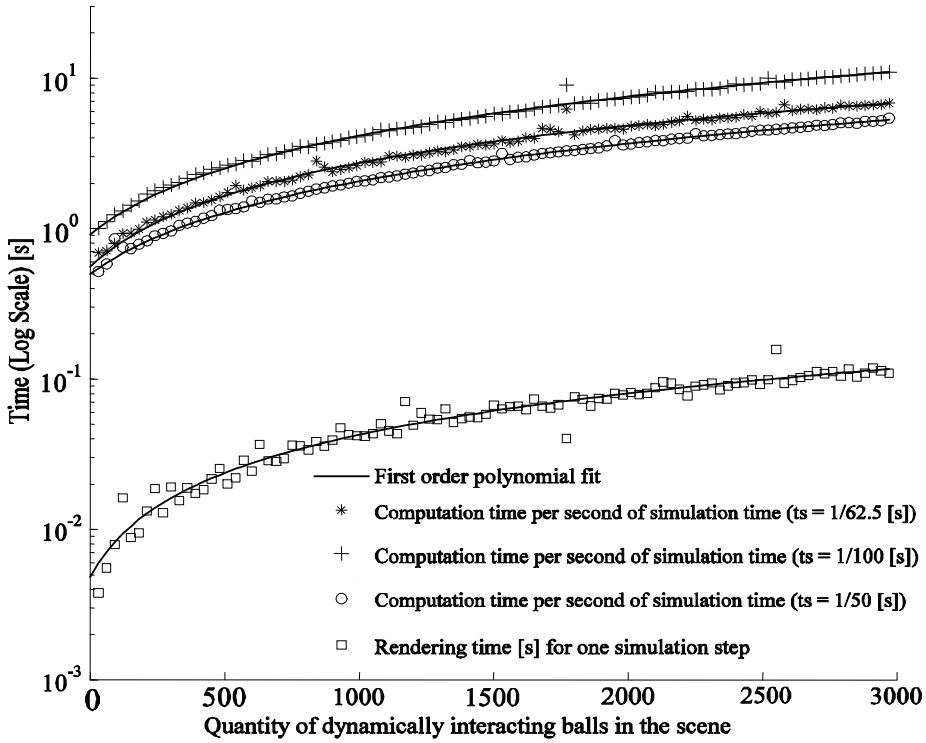

Figure 1.3. Performance of SIMNOS

1.2.5 Sensory Data and Spike Encoding

The sensory data generated by SIMNOS includes 25 Euler angle values that monitor the relative rotations of thorax-pelvis and head-thorax and every degree of freedom in each hand, arm and shoulder complex.[7] The robot is equipped with 41 muscles and the current length is available for each muscle, together with the control values that were issued to it: a total of 164 values per time step.[8] The virtual robot is configurable to have either one or two eyes, which provide a continuous visual stream from the virtual environment.

[7] These angles are indicated in Figure 1.2 by the positions of the arrows within the outlined spheres.

[8] The muscles are shown as red lines in Figure 1.2.

To interact with the SpikeStream simulator Richard Newcombe developed a simple model to convert the real valued sensor data into a time varying spike train. Current theories of neural coding fall under either rate or temporal encoding schemes (Bialek et al. 1991, Shadlen and Newsome 1994) and this model utilizes a hybrid, spatially distributed, average rate encoding method**.** This spans the range of a real valued variable with a set of *N* broadly tuned 'receptors'. Each receptor, $n \in \{0..N\}$, is modelled with a normalised Gaussian with mean μ_n and variance σ_n^2 (1.1) (1.2), with the values of μ_n computed to equally divide the variable range with a receptor mean at the minimum and maximum of the range.

$$\mu_n = \frac{n}{N-1} \tag{1.1}$$

$$\sigma_n = \frac{1}{3(N-1)} \tag{1.2}$$

Given a real valued variable at time *t,* ($v_t \in [0..1]$), the spiking output of each receptor ($r_n \in \{0,1\}$) is computed based on the probability, $p(n, v_t)$ of that receptor firing (equations 1.3 and 1.4), where *c* is a scaling factor used to control the maximum firing rate of a receptor and *rand* is drawn from a uniform distribution. The variance of a receptor is chosen to ensure that $p(n, v_t) = 1$ when $\mu_n = v_t$, with all other receptors having negligible probability.

$$r_n(t) = \begin{cases} 1 & iff \quad p(n, v_t) > k \\ 0 & \end{cases} \tag{1.3}$$

$$p(n, v_t) = e^{\frac{-(v_t - \mu_n)^2}{2\sigma_n^2}} \qquad k = c \cdot rand[0,1] \tag{1.4}$$

Given *N* spike trains the conversion back to a real value is performed by taking the average normalised firing rate $fr_n(t)$ for the current time step *t* within a given window of *w*

previous simulation steps for each of the *N* spiking signals. The approximated real value at this time, $\tilde{v}_n(t)$, is then the sum of the receptor means weighted by the firing probability (equations 1.5 and 1.6).

$$fr_n(t) = \frac{\sum_{i=t}^{t-w} r_n(i)}{w} \tag{1.5}$$

$$\tilde{v}_n(t) = \sum_{n \in N} \mu_n \cdot fr_n(t) \tag{1.6}$$

Such a spatially distributed rate encoding provides resilience to noisy signals, with the benefit that increased resolution in spiking representation can be achieved without altering the rate of firing of an individual neuron. The same sensory data scheme is being applied to the CRONOS hardware robot so that the two systems will have the same interface. Unfortunately this was not completed in time for this thesis, and so only the SIMNOS robot was used in this PhD.

1.3 Thesis Summary

The overall aim of this PhD was to develop a neural network to control the SIMNOS robot (Chapter 5) and to analyze this network for consciousness (Chapter 7). This analysis required a consistent interpretation of consciousness (Chapter 2) and I had to develop a way of analyzing systems for phenomenal states (Chapter 4). A new spiking neural simulator called SpikeStream was developed to model the neural network (Chapter 6) and a considerable amount of background research was also carried out (Chapter 3).

Chapter 2: Consciousness

Machine consciousness is a relatively new research area that is highly cross-disciplinary and takes elements from computer science, philosophy, neuroscience and experimental psychology. Although this thesis is primarily about computer science, a significant obstacle to progress in research on consciousness is the large number of conflicting theories and there is a general lack of consensus about what is meant by consciousness. These problems are highlighted by Metzinger (2003), who claims that consciousness is in a pre-paradigmatic state,[9] and Coward and Sun (2007, p. 947) argue that our understanding of consciousness suffers from "considerable meta-theoretical confusion". In order to develop a systematic way of analyzing machines for consciousness, it was necessary to carry out some philosophical work to clarify the concept of consciousness and outline a framework for its scientific study, which is used in the analysis work in later chapters. This examination of consciousness uses the neurophenomenological approach put forward by Varela (1996), in which phenomenological methods are used to shed light on work in the physical sciences.

The first part of this chapter develops an interpretation of consciousness that distinguishes between the phenomenal world of our experiences and the physical world described by science. This distinction between the phenomenal and the physical leads to a definition of consciousness that is compared with other definitions and linked to a correlates-based approach, which is becoming increasingly popular through research on the neural correlates of consciousness. The correlates of consciousness are examined in more detail and two types of potential correlates of consciousness (PCCs) are identified. Type I PCCs are behaviour-neutral, which makes it makes it impossible to prove their connection with consciousness empirically, whereas type II PCCs do affect behaviour and it is possible to establish if they are systematically linked to conscious states. This type I/ II distinction is used to classify different

[9] See the quotation at the beginning of this chapter.

theories of consciousness and it plays an important role in the approach to synthetic phenomenology that is developed in Chapter 4.

The last part of Chapter 2 sets out three theories of consciousness, which are used to analyse the network in Chapter 7, and it concludes with a discussion of the relationship between consciousness and action.

Chapter 3: Machine Consciousness

This chapter provides a context for the work in this thesis by summarizing some of the previous research on machine consciousness. To provide a more systematic interpretation of this work, the research on machine consciousness is divided into four different areas:

- *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.

In the first part of Chapter 3 this classification is used to examine the relationship between machine consciousness and other disciplines, and to interpret some of the criticisms that have been raised against work in this area. The central part of this chapter covers some of previous work on machine consciousness and the final part discusses the ethical issues surrounding this type of research and looks at the potential benefits.

Chapter 4: Synthetic Phenomenology

A systematic method for measuring the consciousness of an artificial system is essential if researchers want to prove that they have created a conscious machine, and feedback about the consciousness of a system is also useful if one wants to extend or enhance its consciousness.

Whilst it is reasonably easy to see how the behaviour, cognitive characteristics and architecture associated with consciousness can be identified using standard techniques, it is much harder to see how phenomenal consciousness can be measured. With humans, the presence of phenomenal states is generally established through verbal communication, but most of the systems that have been developed as part of research on machine consciousness are only capable of non-verbal behaviours. Since relatively little work had been carried out in this area, new techniques had to be created to identify and describe the phenomenal states of the artificial neural network that was developed by this thesis.

The correlates of consciousness can only be used to decide whether a machine is conscious when scientific experiments have identified a list of the necessary and sufficient correlates, and Chapter 2 argues that type I potential correlates of consciousness cannot be empirically separated out. To address this problem, Chapter 4 outlines an ordinal machine consciousness (OMC) scale that models the contribution that a system's type I correlates make to our belief that it is capable of phenomenal states. When a system's type I correlates match those of the human brain, it is given an OMC rating of one; when we believe that a system is unlikely to be conscious, its OMC rating is close to zero.

The second half of Chapter 4 develops a new and systematic way of describing artificial conscious states. This approach formulates precise definitions of mental states and representational mental states, and suggests how representational mental states can be identified by exposing the system to different test stimuli and measuring its response. Problems with the description of representational mental states in human language led to the use of a markup language for the final phenomenological description, which makes less assumptions about the common ground between the consciousness of humans and artificial systems.

Chapter 5: Neural Network

Chapter 5 describes a spiking neural network with 17,544 neurons and 698,625 connections that controls the eye movements of the SIMNOS virtual robot and uses its 'imagination' and 'emotions' to decide whether it looks at a red or blue cube. This network was designed to give SIMNOS the external behaviour associated with consciousness (MC1) using the cognitive characteristics associated with consciousness (MC2), and it was analyzed for phenomenal states (MC4) using the methodology set out in Chapter 4. As part of the testing of the network some visualizations of its 'imagination' were recorded and its behaviour was quantitatively measured.

Chapter 6: SpikeStream

Although it might have been easier to use an existing simulator to create the network described in Chapter 5, none of the available simulators were suitable, either because of the scale of the network, the type of modelling, or because they would have been difficult to modify to interface with the SIMNOS virtual robot. This led me to develop a new spiking neural simulator called SpikeStream, which is based on Delorme and Thorpe's (2003) SpikeNET architecture. Chapter 6 gives a brief high level summary of the architecture, features and performance of SpikeStream; much more detailed information is available in the SpikeStream manual, which is included as Appendix 1 in this thesis.

Chapter 7: Analysis

The final chapter documents the work that was done to establish whether the neural network created by this project was predicted to be conscious according to Tononi's (2004), Aleksander's (2005) and Metzinger's (2003) theories. The first stage in this process was the identification of representational mental states in the network. This was done by injecting noise into the input and output layers and mutual information was used to identify the parts of the system that responded to information in the input or output layers. The network was then examined for information

integration (Tononi and Sporns 2003), which was used to analyze the network according to Tononi's theory of consciousness, to support the analysis for Metzinger's theory of consciousness and to evaluate the integration between neurons in the network. This analysis for information integration was a considerable challenge because of a factorial relationship between the size of the network and the number of calculations that had to be carried out, and a number of different approximation strategies were used to complete the analysis in a reasonable time. The final part of the analysis was the generation of files containing a description of the predicted phenomenology of the network at each time step, and the predicted distribution of consciousness was plotted for Tononi's, Aleksander's and Metzinger's theories. These results showed that different parts of the network were predicted to be conscious according to the three theories, but it was not possible to predict the absolute amount of consciousness because the measures had not been calibrated on normal waking human subjects.

Appendix 1: SpikeStream

Appendix 1 is a manual documenting the installation and features of SpikeStream. This manual was included with the SpikeStream 0.1 release.

Appendix 2: Network Analyzer

This appendix summarizes the main features of the Network Analyzer software, which was developed for the analysis part of this thesis.

Appendix 3: Seed and Group Analyses

This appendix presents the detailed results from the seed and group information integration analyses.

Appendix 4: Gamez Publications Related to Machine Consciousness

A list of publications by David Gamez that are connected to the work in this thesis.

1.4 Supporting Materials

This thesis is accompanied by a number of supplementary materials, which are available on CD and at www.davidgamez.eu/mc-thesis/. These include:

- A copy of the thesis in Adobe's .pdf format.
- A website implementing the OMC scale.
- Java code for the OMC scale.
- SpikeStream code.
- SpikeStream source code documentation.
- Network Analyzer code.
- Results from the representational mental states analysis in XML format.
- Results from the validation on Tononi and Sporns' test networks in XML format.
- Results from the information integration analysis in XML format.
- The neural network developed by the project in SpikeStream format.
- Recordings of the network in SpikeStream format.
- Videos of the network.
- The final XML description of the synthetic phenomenology of the network.

These supporting materials are constructed as a website, which can be launched by double clicking the index.html file at the root directory of the CD.