8. CONCLUSIONS

> I shall certainly admit a system as empirical or scientific only if it is capable of being tested by experience. These considerations suggest that not the verifiability but the falsifiability of a system is to be taken as a criterion of demarcation. In other words: I shall not require of a scientific system that it shall be capable of being singled out, once and for all, in a positive sense; but I shall require that its logical form shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical system to be refuted by experience.
>
> Popper (2002, p.18)

8.1 Achievements

One of the key achievements of this thesis was the development and demonstration of a synthetic phenomenology framework that provides a way of predicting and describing the conscious states of artificial systems using different theories of consciousness. This methodology works entirely from a third person perspective and it does not rely on implicit assumptions about biological neurons being necessary for consciousness. Systematic falsifiable predictions about artificial conscious states could help machine consciousness to become more scientific, and this methodology may also contribute to the science of consciousness more generally since it enables predictions to be made about the consciousness of biological systems. The work on synthetic phenomenology also offered a number of significant innovations:

- An OMC scale that models our intuitions about the consciousness of artificial systems.
- A clear definition of mental states and representational mental states.
- A method for the identification of representational mental states that uses noise injection and mutual information.

- New approximation methods for measuring the information integration of systems with more than a few dozen elements.
- The use of a markup language to describe artificial phenomenal states.
- Detailed predictions about the distribution of consciousness in a neural network according to Tononi's, Metzinger's and Aleksander's theories.

A second achievement of this project was the development of a neural network that used some of the cognitive characteristics associated with consciousness to control the eye movements of the SIMNOS virtual robot. This network is a novel contribution to the field and differs from the networks developed by Aleksander (2005), Shanahan (2006, 2008), Dehaene et al. (1998, 2003, 2005) and Krichmar et al. (2005). This network exhibited a limited form of conscious behaviour (MC1), had cognitive characteristics associated with consciousness (MC2) and was predicted to be phenomenally conscious (MC4) according to three theories of consciousness, and so this thesis can lay reasonable claim to have created an extremely limited form of consciousness for SIMNOS, and thus to have fulfilled one of the key aims of the CRONOS project.

A further significant achievement of this project was the development of the SpikeStream neural simulator. This has good performance and its simulation features and graphical interface were a substantial advance over Delorme and Thorpes's (2003) implementation of the SpikeNET architecture. The source code of SpikeStream is fully documented and SpikeStream has been released both as source code and pre-installed on a VMWare virtual machine running SUSE Linux. The close integration between SpikeStream and SIMNOS makes them an extremely powerful toolset for carrying out research into all aspects of perception, muscle control, machine consciousness and spiking neural networks.

Finally, this thesis makes a number of theoretical contributions to the study of natural and artificial consciousness, which include the discussion of the relationship between the

phenomenal and physical, the distinction between type I and type II potential correlates of consciousness, and the analysis of conscious will and conscious control. The distinction between the different MC1-4 areas of machine consciousness was also original and the review of work in machine consciousness, published as Gamez (2007a), received a positive response from other people working in the field.

8.2 General Discussion and Future Work

This thesis has emphasized the importance of scientific experimentation in machine consciousness research. Whilst theoretical discussion is needed to establish a framework within which empirical work can take place, machine consciousness will only become fully scientific when it can make falsifiable predictions about the consciousness of artificial systems.[1] Key requirements for this are more formal definitions of each theory that can be used to make predictions about the consciousness associated with different systems. These definitions can be mathematical equations, algorithms or pieces of code – their only requirement is that they take the states of an arbitrary system as input and generate predictions about its phenomenal states. The work of Tononi (2004) is a good example of how a theory of consciousness can be formalized in this way, and the definitions offered in Section 7.6.2 and Section 7.7.3 were a first attempt at a formalization of Aleksander's and Metzinger's theories.

To compare predicted distributions of consciousness with first 'person' reports, more work needs to be done on how artificial systems can be given the ability to speak about their conscious states – perhaps using the work of Steels (2001, 2003). More theoretical work is also needed to understand how the reporting of conscious states fits into the framework of conscious control and how this works at a phenomenal and physical level. Formalized theories of consciousness could also be used to make predictions about the consciousness of biological

[1] This view is shared by Crick and Koch (2000) – see the quotation in Section 2.6.1.

systems that can report their conscious states, which could be tested through collaborations with people working in experimental psychology and neuroscience. The current lack of low level access to biological systems' states means that this work is not likely to progress very fast until scanning technologies experience breakthroughs in their temporal and spatial resolution.

Many parts of the approach to synthetic phenomenology in this thesis are based on numerical methods that need to be tested and calibrated on real data. To begin with, the OMC scale could be tested by using psychophysical methods to establish how accurately it models our subjective assessment about the link between type I PCCs and consciousness. Second, we need to measure how much mutual information is necessary for a state to become representational in a real biological system, and the link between mutual information and depiction needs to be validated and calibrated by estimating the amount of depiction in humans. Finally, the information integration of real biological systems needs to be measured to establish a connection between information integration and consciousness. This process faces many problems, such as the size of real biological neural networks, the fact that noise injection cannot be practiced on humans and the low spatial and/ or temporal resolution of scanning data.

The neural network developed by this project was very basic and could be improved in many ways. One direction of improvement would be to use SIMNOS's visual pre-processing to add layers sensitive to movement, edges and other data, which could work in a similar way to the visual input layers in the network developed by Krichmar et al. (2005). A reactive layer could also be included to improve the performance of the network and to make it capable of conscious will. In this thesis the lack of a viable software interface for CRONOS and delays in the production of the final robot meant that it was not possible to test the network on a real system, and this is something that could be attempted in future work. The learning of the network could also be improved and more research needs to be done on how learning can be implemented on different time scales.

In the future a well documented biologically inspired test network could be developed that would enable people to validate their predictions about consciousness on a commonly agreed standard and compare different methods of measuring functional and effective connectivity. Although a common project or series of meetings might be needed to design such a network, the previous work on machine consciousness (Chapter 3) and on the simulation of biologically inspired neural networks (Section 5.6) suggests that enough work has been done to design an initial test system.

The interpretation of consciousness put forward in Chapter 2 will not be popular with people who believe that some kind of reduction of the phenomenal to the physical is the only way in which a science of consciousness can proceed. However, if a non-reductive interpretation is correct, then it could provide a more secure framework for a science of consciousness, and in the future more work needs to be done to clarify this approach and work through 'use cases' that examine the relationship between the phenomenal and the physical in as much detail as possible. One major problem is how independent causal chains within the phenomenal and the physical should be understood, and it may need some reworking of the concept of causation to deal with the crossover that occurs when a conscious decision leads to changes in the physical world.[2]

The main focus of this thesis was on the development of a systematic framework for analyzing systems for conscious states. Since current theories could be used to illustrate this approach, it was not necessary to develop a new type II theory of consciousness in this thesis, and little attempt was made to criticize or improve existing theories. As robots and scanning technologies improve, we will be able to make more accurate comparisons between predictions about consciousness and reports of conscious states, which should enable us to develop better type II theories of consciousness.

[2] Hume's (1983) interpretation of causation as a constant conjunction between cause and effect might be applicable here.