Information Integration Based Predictions about the Conscious States of a Spiking Neural Network

David Gamez

Department of Electrical and Electronic Engineering,
Imperial College, London, SW7 2AZ, UK

david@davidgamez.eu

+44 (0) 20 8533 6559 (work) / +44 (0) 7790 803 368 (mobile)

NOTICE: this is the author's version of a work that was accepted for publication in Consciousness and Cognition. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version is forthcoming in Consciousness and Cognition, (2010), doi:10.1016/j.concog.2009.11.001.

Abstract

This paper describes how Tononi's information integration theory of consciousness was used to make detailed predictions about the distribution of phenomenal states in a spiking neural network. This network had approximately 18,000 neurons and 700,000 connections and it used models of emotion and imagination to control the eye movements of a virtual robot and avoid 'negative' stimuli. The first stage in the analysis was the development of a formal definition of Tononi's theory of consciousness. The network was then analyzed for information integration and detailed predictions were made about the distribution of consciousness for each time step of recorded activity. This work demonstrates how an artificial system can be analyzed for consciousness using a particular theory and in the future this approach could be used to make predictions about the phenomenal states associated with biological systems.

Keywords: Prediction; spiking neural network; robot; machine consciousness; synthetic phenomenology; neurophenomenology; information integration; consciousness.

1. Introduction

Scientific theories provide general explanations of phenomena that can be used to predict the current and future states of the world. When several theories are put forward to explain a phenomenon, the theory that makes the most accurate predictions is typically preserved (all other factors being equal), and theories that make incorrect predictions are discarded. According to Popper (2002), this ability to make falsifiable predictions is one of the key characteristics of scientific theories.

Although the scientific study of consciousness has made a great deal of progress in recent years, there is still an extensive debate about what is meant by 'consciousness', and a large number of conflicting theories have been put forward. One source of these problems is that most consciousness research has been broadly inductive in character and there has been very little emphasis on falsifiable predictions that can be used to discriminate between different theories. As Crick and Koch (2000, p. 103) point out, if the science of consciousness is to move forward it needs to make predictions according to competing theories and use empirical measurements to eliminate theories that make bad predictions.

One of the main reasons why there has been little work on detailed predictions about consciousness is that we have very poor access to neurons in the living brain. Using non-invasive scanning technologies, such as fMRI, it is possible to achieve a temporal resolution of the order of 1 second and a spatial resolution of about $1mm^3$, which represents the average activity of approximately 50,000 neurons (Witelson et al., 1995). Electrodes give access at much finer temporal and spatial scales, but only a few hundred neurons can be monitored at a time and there are few circumstances in which this type of experiment can be carried out on human subjects. A second problem with making predictions about the brain's consciousness is that the computational cost of analysing large numbers of neurons can be extremely high. For example, it has been calculated that it could take up to 10^{9000} years to complete a full analysis of an 18,000 neuron network using Tononi's (2004) information integration theory (Gamez, 2008b), and whilst optimizations can be found, we are very unlikely be able to analyze the entire human brain according to this theory in the foreseeable future.

Given the debate about consciousness in artificial systems and non-human animals (Baars, 2000; Crook, 1983), theories of consciousness can only really be tested on the human brain, which is generally acknowledged to be the gold standard of a conscious system. If we had better access to the human brain, we could use different theories to make detailed

predictions about its consciousness, and these could be confirmed or refuted by first person reports. Until access to the brain improves, artificial neural networks offer a way in which we can develop new ways of making predictions according to different theories of consciousness. Since artificial networks can be considerably smaller than biological brains, the computational cost of the analysis can be controlled, and artificial systems have the great advantage that it is possible to obtain full access their internal states. This type of work on artificial networks is part of what Seth (2009) calls weak artificial consciousness: artificial systems are being used to develop ways of making predictions about consciousness without any claims about the actual consciousness of the artificial system.

As access to the brain improves and computing power increases it will become possible to test theories of consciousness by making detailed predictions about human consciousness. Once a theory of consciousness has been validated, it can be used to make predictions about the consciousness of artificial systems, as part of work on strong artificial consciousness (Seth, 2009). Whilst some theories of consciousness predict zero consciousness in artificial systems with non-biological hardware (Searle, 1992), other theories, such as Tononi (2004), make positive predictions about the phenomenal states of artificial systems. If future experiments demonstrated a strong link between information integration and consciousness, then we would have grounds for believing that the predicted phenomenal states described in Section 6 were actually present in the artificial neural network described in Section 4.

To make predictions about a system's consciousness, the states of the system have to be recorded (typically the firing of neurons in a brain or network), and then a theory of consciousness is used to generate assertions about the phenomenal world associated with the physical system.[1] Since many theories of consciousness are set out in a fairly abstract high

1 Whilst the brain's physical structures and functions are often described as causing consciousness, conceptual problems surrounding consciousness, such as the putative hard problem (Chalmers, 1995), and a number of

level language, it might be necessary to transform a high level description into a more precise definition that can be implemented in computer code. This formal definition may depend on other analyses of the system – perhaps for mutual information or information integration – that have to be carried out before the final predictions about phenomenal states can be made.

To illustrate this process, this paper describes how Tononi's (2004) theory was used to make detailed predictions about the phenomenal states of a spiking neural network. The first part of this paper describes previous work in this area, and then Tononi's theory of consciousness is summarized and given a formal definition in Section 3. Section 4 covers the spiking neural network that was developed to demonstrate this approach, which uses biologically-inspired models of potential correlates of consciousness, such as imagination and emotion (Aleksander, 2005), to control the eye movements of a virtual robot. Section 5 then outlines how the network was analyzed for information integration to support the predictions about consciousness. The main results of the analysis are presented in Section 6 and the paper concludes with a discussion of the results and some suggestions for future work.

The overall goal of this paper is to show how predictions can be made about the phenomenal states of a neural network using different theories of consciousness. To keep the paper to a reasonable length some of the details of the analysis have been omitted, and can be found in the Supplementary Material and Gamez (2008b).

2. Previous Work

A number of people working in neuroscience and experimental psychology have used fMRI data to make predictions about subjects' perceptions and thoughts – a process that is sometimes referred to as "brain reading". A recent example of this type of work is Kay et al.

theoretical considerations (Gamez 2008b) make talk about correlation or association more theoretically and empirically tractable. The presence of particular physical, neural or functional states can be used to predict the co-occurrence of phenomenal states.

(2008), who used fMRI data to predict the images that subjects were viewing with over 70% accuracy. In the first stage of these experiments Kay et al. scanned the subjects' brains while they looked at different test images. The images were processed into a set of Gabor wavelets with different sizes, positions, orientations, spatial frequencies and phases and a correlation was established between the wavelets present in each image and the response of each fMRI voxel. Once their brains' responses had been mapped out, subjects were exposed to novel images and Kay et al. attempted to predict the image that subjects were viewing by transforming the novel images into the Gabor wavelet representation and selecting the image whose predicted brain activity most closely matched the actual brain activity. A similar approach was used by Mitchell et al. (2008) to predict the words subjects were reading and Haynes et al. (2007) used fMRI data to predict the task that people were intending to perform. Whilst this type of work is highly relevant to the attempt to make predictions about phenomenal states using different theories of consciousness, its main emphasis is on the identification of patterns in fMRI data that can be correlated with perceptions or intentions, and none of the authors claimed that these patterns were correlates of consciousness or attempted to make predictions according to a particular theory of consciousness.

Within research on machine consciousness there has been some work on the analysis of artificial systems for consciousness, which forms part of the emerging discipline of synthetic phenomenology. One example of this type of research is Holland and Goodman (2003), who programmed a simulated Khepera with simple behaviours and used Linåker and Niklasson's (2000) Adaptive Resource-Allocating Vector Quantizer (ARAVQ) method to build up concepts that corresponded to a combination 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 Holland and Goodman used these concepts to produce a graphical representation of the Khepera's internal model and examined

how the model was used to control the robot. A similar approach was used by Stening et al. (2005) to graphically represent the 'imagination' of a Khepera robot. This type of work can be interpreted as a description of the robot's inner states based on a theory of consciousness, although neither set of authors claimed that the graphical representations were predictions about the systems' phenomenal states. Other related work in synthetic phenomenology was carried out by Chrisley and Parthemore (2007), who used a SEER-3 robot to specify the non-conceptual content of a model of perception based on O'Regan and Noë's (2001) sensorimotor contingencies. In Chrisley's and Parthemore's work the robot model was used to describe human phenomenal states that are difficult to articulate in natural language.

Other analysis work based on information integration has been carried out by Lee et al. (2009), who made multi-channel EEG recordings from eight sites in conscious and unconscious subjects and constructed a covariance matrix of the recordings on each frequency band that was used to identify the complexes within the 8 node network using Tononi and Sporns' (2003) method. This experiment found that the information integration capacity of the network in the gamma band was significantly higher when subjects were conscious. Theoretical work on information integration has been carried out by Seth et al (2006), who identified a number of weaknesses in Tononi and Sporns' (2003) method and criticized the link between information integration and consciousness.

A number of other measures of the information relationships between neurons have been put forward, including neural complexity (Tononi et al. 1994, 1998), transfer entropy (Schreiber, 2000) and causal density (Seth et. al., 2006). These measures have been used by a number of people to examine the anatomical, functional and effective connectivity of biological networks, either using scanning or electrode data, or large-scale models of the brain. For example, Honey et al. (2007) used transfer entropy to study the relationship between anatomical and functional connections on a large-scale model of the macaque cortex,

and demonstrated that the functional and anatomical connectivity of their model coincided on long time scales. Other examples of this type of work are Brovelli et al. (2004), who used Granger causality to identify the functional relationships between recordings made from different sites in two monkeys as they pressed a hand lever during the wait discrimination task, and Friston et al. (2003) modelled the interactions between different brain areas and made predictions about the coupling between them. Information-based analyses have also been used to guide and study the evolution of artificial neural networks connected to simulated robots (Seth and Edelman, 2004; Sporns and Lungarella, 2006). An overview of this type of research can be found in Sporns et. al. (2004) and Sporns (2007).

The neural network in Section 4 was influenced by a number of other biologically-inspired neural networks and neural models of the correlates of consciousness. For example, Aleksander (2005) and Aleksander and Morton (2007) used weightless neurons to build a number of brain-inspired neural networks that included all five of Aleksander's axioms, Shanahan (2006) developed a brain-inspired cognitive architecture based on global workspace theory that directed the movements of a virtual Khepera robot, and Cotterill (2003) hoped to identify signs of consciousness in a brain-inspired neural network that controlled a virtual child. There is also the work of Krichmar et. al (2005) and Krichmar and Edelman (2006), who carried out a number of experiments with robots controlled by neural networks closely based on the brain, and a large scale model of the mammalian thalamocortical system was built by Izhikevich and Edelman (2008).

3. The Information Integration Theory of Consciousness

The information integration theory of consciousness was put forward by Tononi (2004), who claims that the capacity of a system to integrate information is correlated with its amount of consciousness, and that the conscious part of the system is the part that integrates the most

information. Information integration is measured using the value Φ and Tononi and Sporns (2003) describe an algorithm that can be used calculate Φ on any system of connected elements.

To measure the information integrated by a subset of elements, S, the subset is divided into two parts, A and B. A is then put into a state of maximum entropy ($A^{H^{MAX}}$) and the mutual information, MI, between A and B is measured to get the *effective information* (EI), as expressed in Equation 1:

$$EI(A \rightarrow B) = MI(A^{H^{MAX}}; B), \quad (1)$$

where MI(A; B) is given by Equation 2:

$$MI(A; B) = H(A) + H(B) - H(AB), \quad (2)$$

where H(*x*) is the entropy of *x*. Since A has effectively been substituted by independent noise sources, there are no causal effects of B on A, and so the mutual information between A and B is due to the causal effects of A on B. EI(A→B) also measures all possible effects of A on B and EI(A→B) and EI(B→A) are in general not symmetrical. The value of EI(A→B) will be high if the connections between A and B are strong and specialized, so that different outputs from A produce different firing patterns in B. On the other hand, EI(A→B) will be low if different outputs from A produce scarce effects or if the effect is always the same.

The next stage is the repetition of the procedure in the opposite direction by putting B into a state of maximum entropy and measuring its effect on A, giving EI(B→A). For a given bipartition of the subset S into A and B, the effective information between the two halves is given by Equation 3:

$$EI(A \leftrightarrow B) = EI(A \rightarrow B) + EI(B \rightarrow A). \quad (3)$$

The amount of information that can be integrated by a subset is limited by the bipartition in which EI(A↔B) reaches a minimum. To find this *minimum information bipartition* the analysis is run on every possible bipartition of the subset, with EI(A↔B) being normalised by maximum entropy of A or B when the effective information of each bipartition is compared. The *information integration* for subset S, or Φ(S), is the non-normalised value of EI(A↔B) for the minimum information bipartition, and this measures the amount of causally effective information that can be integrated across the informational weakest link of the subset.

A *complex* is defined by Tononi and Sporns (2003) as a part of the system that is not included in a larger part with higher Φ. To identify complexes it is necessary to consider every possible subset S of *m* elements out of the *n* elements of the system starting with $m = 2$ and finishing with $m = n$. For each subset Φ is calculated and the subsets that are included in a larger subset with higher Φ are discarded, leaving a list of complexes. The *main complex* is the complex that has the maximum value of Φ, and Tononi (2004) claims that this main complex is the conscious part of the system. To substantiate his link between Φ and consciousness, Tononi (2004) compares different network architectures with structures in the brain and shows how the architectures with high Φ map onto circuits in the brain that are associated with consciousness. Full details about the calculation of information integration can be found in Tononi and Sporns (2003).

Since Tononi's theory of consciousness was already algorithmic, relatively little work had to be done to convert it into a formal definition that could be used to automatically analyze an arbitrary system. In this analysis the main adjustment to Tononi's theory was to take account of situations in which two or more complexes share little information and have approximately the same value of Φ - a situation similar to a split brain patient (Gazzaniga, 1970). In these cases it seems rather arbitrary to say that only one of the complexes is conscious just because its Φ value is, for example, 0.01% greater than the other. To

accommodate this possibility, a definition was developed that specifies the notion of an *independent* complex:

> *None of the neurons in an independent complex, A, are part of another complex, B, that has higher Φ than A.* (1)

This definition of an independent complex was incorporated into the formal statement of Tononi's theory given in Definition 2, which introduces a 50% threshold to eliminate independent complexes whose Φ value is substantially less than the main complex of the system.

> *A state of the system will be judged to be phenomenally conscious according to Tononi (2004) if it is part of the main complex or if it is part of an independent complex whose Φ is 50% or more of the Φ of the main complex. The* ***amount*** *of consciousness will be indicated by the Φ of the complex.* (2)

More recently Balduzzi and Tononi (2008) put forward a new algorithm for analyzing the information integration of a system, which identifies complexes on the basis of the network's states. This new algorithm allows for the possibility that there could be several main complexes in a network and avoids the assumption of a steady state solution, but it retains the factorial dependencies of the algorithm in Tononi and Sporns (2003). I am currently developing software to analyze networks according to the new algorithm, which could be used to analyze the network described in this paper in future work.

4. Neural Network

4.1 Overview

This section describes a spiking neural network with 17,544 neurons and 698,625 connections that was developed to test the approach to prediction outlined in this paper. The network was modelled using the SpikeStream simulator and it directed the eye of the SIMNOS virtual robot (see Figure 1) towards 'positive' red features of its environment and away from 'negative' blue objects. To carry out this task it included an 'emotion' layer that responded differently to red and blue stimuli, and neurons that learnt the association between motor actions and visual input. These neurons were used to 'imagine' different eye movements and select the ones that were predicted to result in a positive visual stimulus. This network is a biologically inspired model of aspects of the brain's processing, not a biologically accurate copy, and so the names given to individual layers, such as "Emotion", are only intended as guides indicating that layers' functions were inspired by particular brain areas.

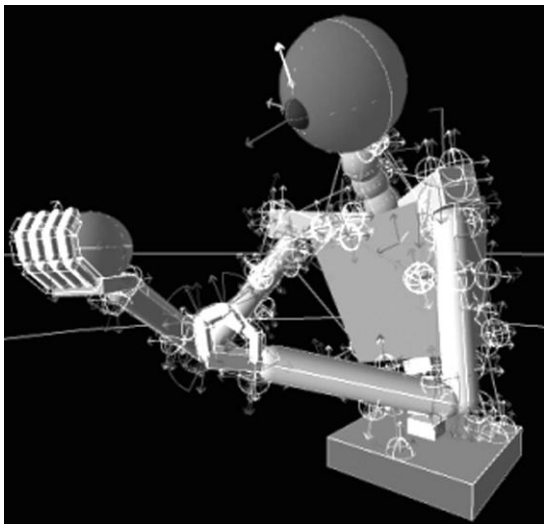

Figure 1. SIMNOS virtual robot. The thin lines are the virtual muscles; the outlines of spheres with arrows are the joints. The length of the virtual muscles and the angles of the joints were encoded into spikes and sent to the SpikeStream neural simulator.

Whilst a smaller network could have been used to demonstrate the analysis work in this paper, one of the key problems with Tononi's (2004) theory is that the required amount of computer processing power increases factorially with the number of elements in the system, and so an analysis of a system with a few hundred elements would not have addressed the scalability issues raised by Tononi's theory. A second motivation for the network's size was that the long term aim of this work is the development of analysis techniques that can make predictions about the consciousness of artificial systems that interact with the real world. One of the major constraints on this type of system is that the network has to be large enough to process visual data at an adequate resolution. In early experiments a larger network was tried with a visual resolution of 128x128 pixels, but this took too long to simulate, and so the current network was based on a visual resolution of 64x64 pixels. Whilst this is tiny compared to biological systems, it is of the same order of magnitude as many robotic systems controlled by neural networks. The application of this type of analysis technique to real systems can also help us to understand what is going on inside robots that learn from their experiences – a point that is discussed in detail in a recent paper (Gamez and Aleksander, 2009).

4.2 Architecture

An illustration of the network architecture is given in Figure 2. The parameters for the layers are provided in Table 1 and details about the connections between layers can be found in Table 2. The neuron model for these experiments was based on the Spike Response Model (Gerstner and Kistler, 2002; Marian, 2003), and learning in the network was carried out using Brader et al.'s (2006) spike time dependent learning algorithm. Full details about the neuron model and training can be found in the Supplementary Material.

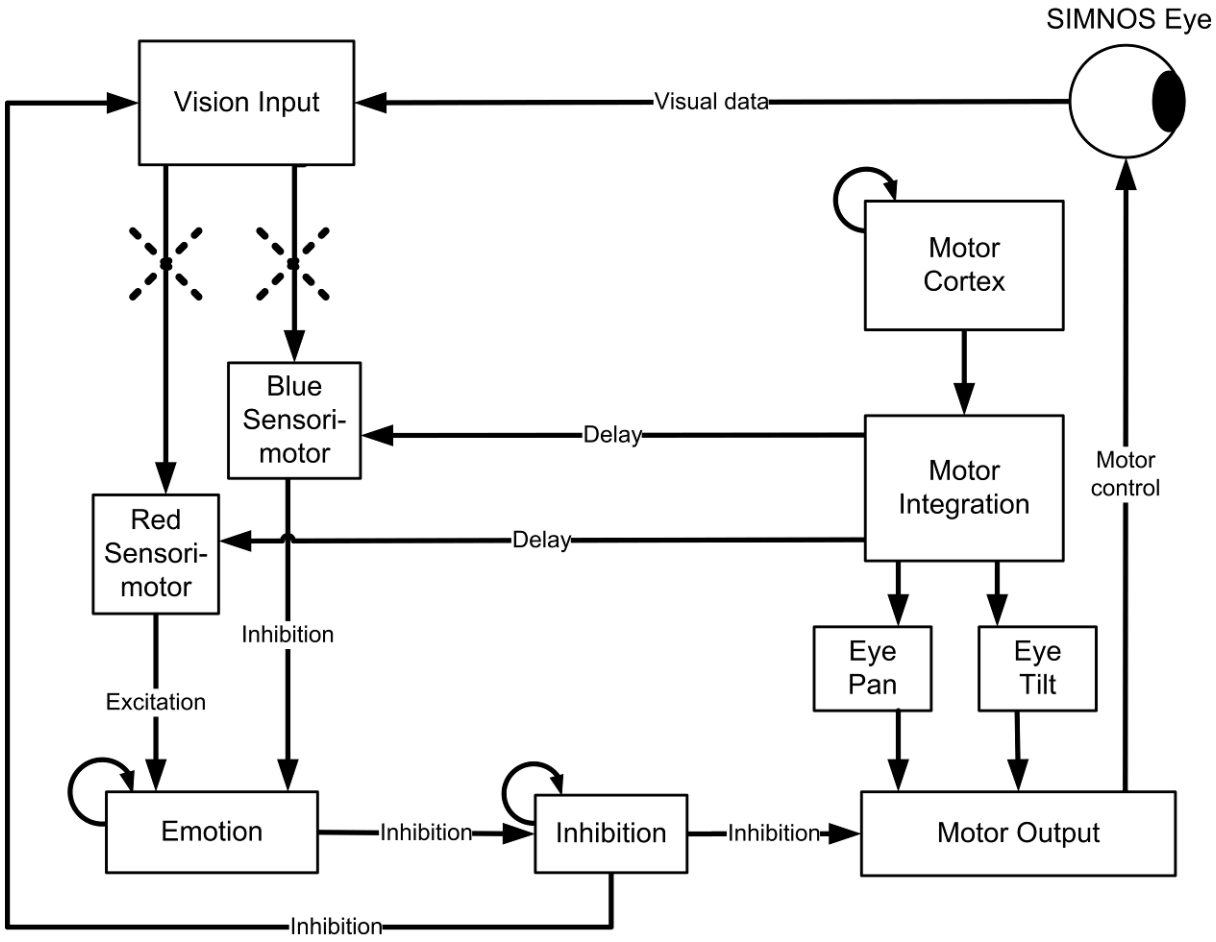

Figure 2. Neural network with SIMNOS eye. Arrows indicate connections within layers, between layers or between the neural network and SIMNOS. The connections marked with dotted crosses were disabled for the imagination test in Section 4.6.

The simulation of the network was carried out using the SpikeStream neural simulator (Gamez, 2007). SpikeStream and SIMNOS communicated using spikes, which were sent from the network to set the pan and tilt of SIMNOS's eye, and spikes containing red and blue visual information were passed back from SIMNOS to SpikeStream and mapped to neurons whose position corresponded to the location of red or blue data in the visual field (see Figure 3).

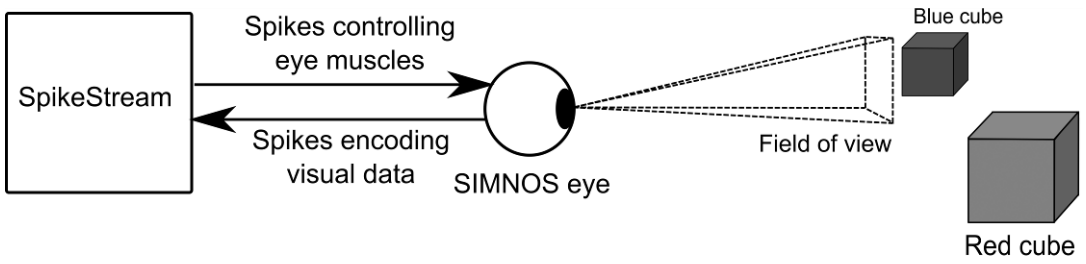

Figure 3. Experimental setup with the eye of SIMNOS in front of red and blue cubes. The eye could only view one cube at a time.

To set up the three dimensional environment of SIMNOS, red and blue cubes were created in Blender[2] and loaded into the SIMNOS environment using the Collada format.[3] The head, arms and body of SIMNOS were locked up by putting them into kinematic mode, which enabled them to be placed in an absolute position and made them unresponsive to spikes from the network, and the eye was moved in front of the red and blue cubes so that it could only view one cube at a time (see Figure 4).

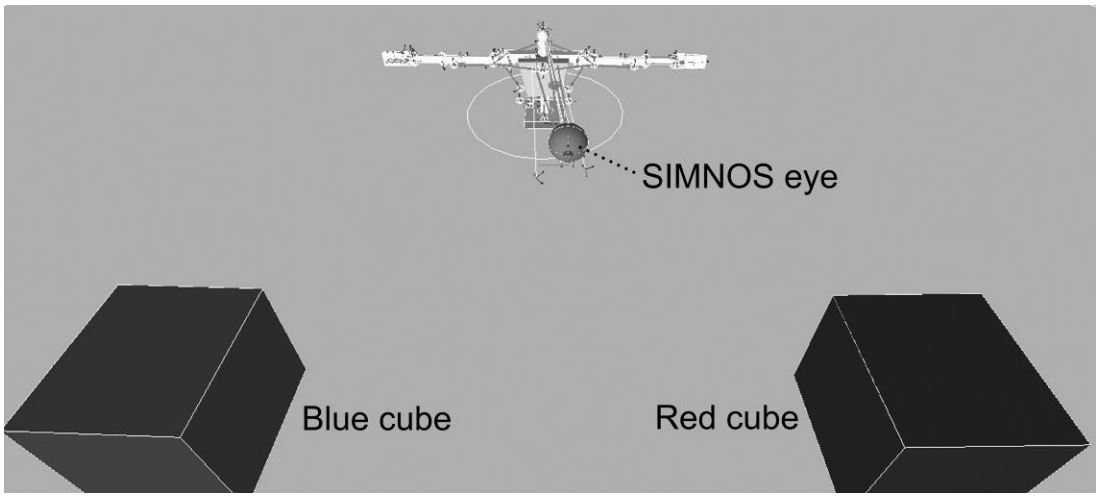

Figure 4. Screenshot of SIMNOS in front of the red and blue cubes

The next two sections highlight some of the key functions of the network and describe the design and functionality of the individual layers in more detail.

2 Blender 3D animation software: www.blender.org.

3 COLLADA format: www.collada.org.

	Area	**Size**	**Threshold**	**Noise**	**Device**
1	Vision Input	64 × 128	0.5	-	SIMNOS vision[4]
2	Red Sensorimotor	64 × 64	0.8	-	-
3	Blue Sensorimotor	64 × 64	0.8	-	-
4	Emotion	5 × 5	2	-	-
5	Inhibition	5 × 5	0.1	20% weight 1.0	-
6	Motor Cortex	20 × 20	1.5	20% weight 0.6	-
7	Motor Integration	5 × 5	0.65	-	-
8	Eye Pan	5 × 1	0.7	-	-
9	Eye Tilt	5 × 1	0.7	-	-
10	Motor Output	5 × 135	0.1	-	SIMNOS muscles

Table 1. Layer parameters

4 Spikes from SIMNOS changed the voltage of the corresponding neurons in Vision Input with a weight of 0.8.

Projection	**Arbor**	**Connection Probability**	**Weight**	**Delay**
Vision Input→Red Sensorimotor	D	1.0	1.0	0
Vision Input→Blue Sensorimotor	D	1.0	1.0	0
Red Sensorimotor →Emotion	U	0.5	0.5	0
Blue Sensorimotor →Emotion	U	0.5	-0.5	0-5
Emotion→Emotion	ECIS 5/ 10	0.5 / 0.5	0.8 ± 0.2 / -0.8 ± 0.2	0-5
Emotion→Inhibition	U	1.0	-1.0	0-5
Inhibition→Inhibition	ECIS 5/10	0.5/ 0.5	0.8 ± 0.2 / -0.8 ± 0.2	0-5
Inhibition→Vision Input	U	1.0	-1.0	0
Inhibition→Motor Output	U	1.0`	-1.0	0
Motor Cortex→Motor Cortex	ECIS 1.7/ 30	0.99/ 0.99	0.8/ -0.8	2
Motor Cortex→Motor Integration	T	1.0	0.5	0
Motor Integration→Red Sensorimotor	U	1.0	0.5	11
Motor Integration→Blue Sensorimotor	U	1.0	0.5	11
Motor Integration→Eye Pan	T	1.0	1.0	0
Motor Integration→Eye Tilt	T	1.0	1.0	0
Eye Pan→Motor Output	D	1.0	1.0	0
Eye Tilt→Motor Output	D	1.0	1.0	0

Table 2. Connection parameters. Unstructured connections (U) connect at random to the neurons in the other layer with the specified connection probability. Topographic connections (T) preserve the topology and use many to one or one to many connections when the layers are larger or smaller than each other. Excitatory centre inhibitory surround (ECIS) connections have excitatory connections to the neurons within the excitatory radius and inhibitory connections between the excitatory and the inhibitory radius - for example, ECIS 5/50 has excitatory connections to neurons within 5 units of each neuron and inhibitory connections to neurons from 5 to 50 units away. A device connection (D) connects a layer to part of an input or output layer that is connected to an external device, such as a robot or camera. So, for example, Red Sensorimotor connects to the part of Vision Input that receives red visual input from SIMNOS.

4.3 Network Functions

Input and output

The spikes containing visual data from SIMNOS's eye were routed so that red and blue visual data was passed to different halves of Vision Input. The Motor Output layer is a complete map of all the 'muscles' of SIMNOS and the activity in each of the five neuron rows was sent as spikes across the network to SIMNOS, where it set the length of the virtual muscles. In these experiments only two rows were active in Motor Output, which controlled eye pan and tilt.

Self-sustaining activity

Three of the layers – Motor Cortex, Emotion and Inhibition – had recurrent positive connections, which enabled them to sustain their activity in the absence of spikes from other layers. A random selection of 20% of the neurons in Inhibition and Motor Cortex were injected with noise at each time step by adding 1.0 or 0.6 to their voltage (see Table 1), and this enabled them to develop their self sustaining activity in the absence of spikes from other layers. The neurons in Emotion could only develop their self-sustaining activity when they received spikes from Red Sensorimotor.

Selection of motor output

The position of SIMNOS's eye was selected by the activity in Motor Cortex, which had long range inhibitory connections that limited its self-sustaining activity to a single small cluster of 2-4 neurons. The activity in Motor Cortex was passed by topological connections to one or two neurons in Motor Integration, which was a complete map of all the possible combinations of eye pan and eye tilt. The activity in Motor Integration was topologically transmitted through Eye Pan and Eye Tilt to Motor Output and passed by SpikeStream over the Ethernet to SIMNOS, where it was used to set the lengths of the eye pan and eye tilt muscles.

Learning

A delay along the connection between Motor Integration and Red Sensorimotor ensured that spikes from a motor pattern that pointed the eye at a red stimulus arrived at Red Sensorimotor at the same time as spikes containing red visual data. When these spikes arrived together, the STDP learning algorithm increased the weights of the connections between Motor Integration and the active neurons in Red Sensorimotor, and decreased the weights of the connections between Motor Integration and inactive neurons in Red Sensorimotor. The same applied to the connections between Motor Integration and Blue Sensorimotor, except that the association between motor patterns and blue visual data was learnt. Prior to the learning, repeated activation of Motor Integration neurons within a short period of time fired all of the neurons in Red/ Blue Sensorimotor. Once the training was complete, spikes from Motor Integration only fired the neurons in Red/ Blue Sensorimotor that corresponded to the pattern that was predicted to occur when the eye was moved to that position.

Online and offline modes

Inhibition had a large number of negative connections to Vision Input and Motor Output, which prevented the neurons in Vision Input and Motor Output from firing when Inhibition was active. This is called the 'imagination' or *offline* mode because in this situation the network was isolated from its environment - no spikes from SIMNOS were processed by the network or sent by the network to SIMNOS - but the system was still generating motor patterns and predicting their sensory consequences. When the neurons in Inhibition were not firing, the neurons in Vision Input were stimulated by spikes from SIMNOS and the neurons in Motor Output sent spikes to SIMNOS to set the position of the eye. This is referred to as the *online* mode of the network. The switch between online and offline modes was controlled by Emotion, which was connected to Inhibition with negative weights: when Emotion was active, Inhibition was inactive, and vice versa. Emotion entered a state of self-sustaining

activity when it received spikes with positive weights from Red Sensorimotor, and its state of self-sustaining activity ceased when it received spikes with negative weights from Blue Sensorimotor.

4.4 Overview of Individual Layers

Motor Cortex

This layer was designed to select a motor pattern at random and sustain it for a period of time. These motor patterns were used to set the lengths of the eye pan and eye tilt muscles in SIMNOS. Short range excitatory and long range inhibitory connections in Motor Cortex encouraged a small patch of neurons to fire at each point in time and this active cluster of firing neurons occasionally changed because a random selection of 20% of the neurons in Motor Cortex were injected with noise at each time step by adding 0.6 to their voltage. The topological connections between Motor Cortex and Motor Integration enabled the active cluster of neurons in Motor Cortex to send spikes to just one or two neurons in Motor Integration.

Motor Integration

Each neuron in this layer represented a different combination of eye pan and eye tilt. Activity in Motor Cortex stimulated one or two neurons in Motor Integration and this activity was transformed through Eye Pan and Eye Tilt into a pattern of motor activity that was sent to SIMNOS's eye. The activity in Motor Integration was also sent along delayed connections to Red Sensorimotor and Blue Sensorimotor, where it was used to learn the relationship between motor output and red and blue visual input.

Eye Pan

This layer connected topologically to Motor Output, where it stimulated the row corresponding to eye pan in SIMNOS. Eye Pan received topological connections from Motor Integration.

Eye Tilt

This layer connected topologically to Motor Output, where it stimulated the row corresponding to eye tilt in SIMNOS. Eye Tilt received topological connections from Motor Integration.

Motor Output

This layer was a complete map of all the 'muscles' of SIMNOS and the activity in each of the five neuron rows in this layer set the length of one of SIMNOS's virtual muscles using the encoding scheme described in Gamez et al. (2006). In these experiments, only eye pan and eye tilt were used and the rest of SIMNOS's muscles were locked up by setting them into kinematic mode. The neurons controlling the eye in Motor Output were topologically connected to Eye Pan and Eye Tilt, and strong inhibitory connections between Inhibition and Motor Output ensured that there was only activity in Motor Output (and motor output from the network) when Inhibition was inactive.

Vision Input

This layer was connected to SIMNOS's visual output so that each spike from SIMNOS stimulated the appropriate neuron in this layer with a weight of 0.8, with one half responding to red visual input from SIMNOS and the other half responding to blue visual input. When Inhibition was inactive the spikes from SIMNOS fired the neurons in Vision Input; when

Inhibition was active, a large negative potential was injected into the neurons in Vision Input, which prevented this layer from responding to visual information.

Red Sensorimotor and Blue Sensorimotor

Red Sensorimotor and Blue Sensorimotor were topologically connected to the red and blue sensitive parts of Vision Input. Positive connections between Red Sensorimotor and Emotion caused Emotion to develop self-sustaining activity when Red Sensorimotor was active. Negative connections between Blue Sensorimotor and Emotion inhibited the self-sustaining activity in Emotion. Red Sensorimotor and Blue Sensorimotor received delayed copies of the motor output from Motor Integration and the synapses on these connections used Brader et al.'s (2006) STDP rule to learn the association between motor output and visual input.

Emotion

This layer played an analogous role to emotions in biological systems, although in a greatly simplified form. Recurrent positive connections within Emotion enabled it to sustain its activity once it had been stimulated. Spikes from Red Sensorimotor set Emotion into a self-sustaining state; spikes from Blue Sensorimotor inhibited it. Emotion inhibited Inhibition, so that when Emotion was active, Inhibition was inactive, and vice versa.

Whilst this layer has been called "Emotion", it had a number of significant differences from real biological emotions. To begin with it did not receive information from the robot's body, and so it was more like the 'as if' loop described by Damasio (1995). Secondly, activity in Emotion was very basic compared to biological emotions because it lacked the detail that we sense when our viscera and skeletal muscles are changed by an emotional state (Damasio 1995, p. 138). A third limitation of Emotion was that its response did not change the way in which the neurons and synapses computed. However, Emotion did respond in a hardwired way to different characteristics of the world with a high impact low information signal that is

characteristic of the neuromodulatory aspect of emotion described by Arbib and Fellous (2004). Since this functional role of Emotion was what was critical in the network, Emotion has been described as a very primitive 'emotion' in this paper – something that is functionally analogous to basic biological emotions.

Inhibition

The Inhibition layer was intended to loosely correspond to the neurons and connections that manage the transition between imagination and online perception in the human brain. When Inhibition was active it inhibited Motor Output and Vision Input and put the system into its offline 'imagination' mode. When Inhibition was inactive Vision Input received visual data from SIMNOS and Motor Output controlled the position of SIMNOS's eye.

The neurons in Inhibition were injected with noise, so that in the absence of any external input this layer was active and automatically put the system into the *offline* 'imagination' mode. Inhibition received negative connections from Emotion, which inhibited its activity, so that when Emotion was active the system was put into its *online* mode.

4.5 Experimental Procedure

The first part of the experiments was a training phase in which the network learnt the association between motor output and visual input. Since the offline mode interfered with this training, it was disabled by blocking the connections from Inhibition. During the training phase spontaneous activity in Motor Cortex changed the position of SIMNOS's eye, copies of the motor signals were sent from Motor Integration to Red/ Blue Sensorimotor, and the synapse classes on these connections used Brader et al.'s (2006) rule to learn the association between motor output and red and blue visual input. By monitoring the changes in the weights over time it was empirically determined that a training period of 50,000 time steps (or 50 seconds of simulated time at 1 ms time step resolution) enabled the network to learn the

association between motor output and visual input for most combinations of eye pan and eye tilt.

Once the network had been trained, Inhibition was reconnected and the network was tested. For both training and testing a time step resolution of 1 ms was found to offer a good balance between the accuracy and speed of the simulation.

4.6 Operation of the Network

During the training phase the network spontaneously generated eye movements to different parts of its visual field and learnt the association between an eye movement and a visual stimulus. After training, the network was fully connected up and Motor Cortex moved SIMNOS's eye around at random until a blue object appeared in its visual field. This switched the network into its offline 'imagination' mode, in which it generated motor patterns and 'imagined' the red or blue visual input that was associated with these potential eye movements. This process continued until it 'imagined' a red visual stimulus that positively stimulated Emotion. This removed the inhibition, and SIMNOS's eye was moved to look at the red stimulus.[5]

A rough qualitative evaluation was carried out of the associations that the network had learnt between motor output and visual input. In this test Red Sensorimotor and Blue Sensorimotor were disconnected from Vision Input (the dotted crosses in Figure 2), so that they only received input from Motor Integration, and Vision Input continued to receive visual input from SIMNOS's eye, which remained under the control of Motor Cortex. If the system had learnt the association between motor output and visual input, then the activity in Red/ Blue Sensorimotor, caused by Motor Integration, would match the activity in Vision Input, which was driven by real visual input.

5 Videos of the network in operation are available at: http://www.davidgamez.eu/mc-thesis/pages/videos.html.

During the imagination test visual inspection of Vision Input, Red Sensorimotor and Blue Sensorimotor showed that the ‘imagined’ visual inputs were reasonably close to the real visual inputs, but often a larger area of Red Sensorimotor or Blue Sensorimotor was activated than would have been caused by visual input alone. It also happened that several different patterns were activated simultaneously in Red Sensorimotor and Blue Sensorimotor, which was probably caused by oscillation in Motor Integration between two different positions during training. Furthermore, Red/ Blue Sensorimotor sometimes contained areas of active neurons when the real stimulus was just off screen, which was again probably due to multiple neurons in Motor Integration being simultaneously active during training. Examples of the contrast between imagined and real visual input are given in Figure 5.

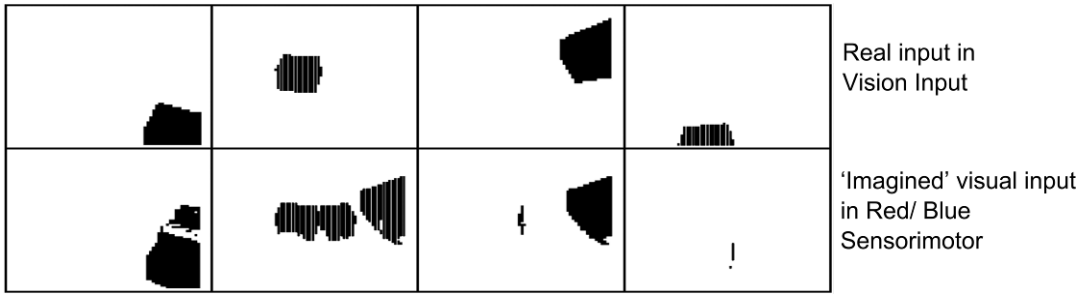

Figure 5. Examples of the contrast between real visual input (top row) and imagined visual input (bottom row)

To test whether the network could avoid exposure to ‘negative’ blue visual input the untrained network was run for 100,000 time steps (100 seconds of simulated time) with Emotion and Inhibition disabled, and the activity in the red and blue sensitive parts of Vision Input was recorded. Emotion and Inhibition were then enabled, the network was trained and the measurements were repeated. This procedure was carried out five times with SIMNOS’s environment set up from scratch on each run to reduce potential biases towards the red or blue cubes that could have been introduced by the manual positioning of the robot’s eye.

The results of the behaviour test are presented in Figure 6 which shows that the activity in the blue visual area was substantially reduced when Emotion and Inhibition were

disabled. This suggests that if the 'negative' blue stimulus was capable of damaging the system, then cognitive mechanisms, such as imagination and emotion, that have been linked by Aleksander (2005) to consciousness could play a useful role in the life of an organism.[6]

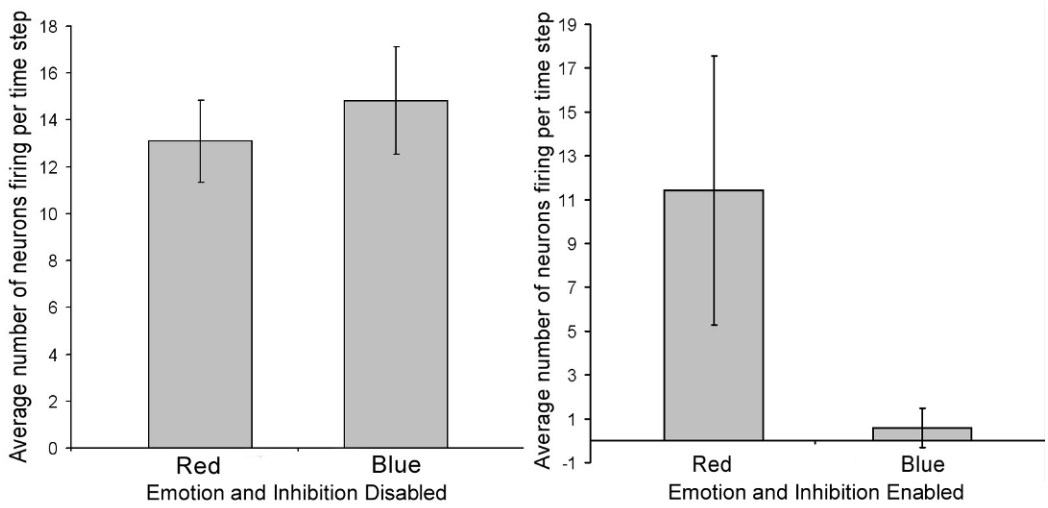

Figure 6. The average number of neurons firing per time step in the red and blue sensitive parts of Vision Input. The left graph shows the results when Emotion and Inhibition were disabled; the right graph shows the results for the fully operational network. [7]

4.7 Analysis Data

The predictions about the network's conscious states were based on two recordings of the neurons' activity. The first set of data - referred to as "Analysis Run" - was recorded as the network controlled SIMNOS's eye and used its 'imagination' to avoid looking at the blue cube. To provide a graphical representation of the activity during Analysis Run, the average number of times that each neuron fired during Analysis Run was recorded, the results were normalized to the range 0-1 and the normalized results were used to illustrate the activity of the network in Figure 7. This shows that Inhibition was the most active part of the network, followed by Emotion, and traces of motor and visual activity can also be seen.[8]

6 It is worth noting that the imagination did not have to be particularly accurate to carry out this function.

7 The error bars are +/- 2 standard deviations.

8 A video of Analysis Run is available at: http://www.davidgamez.eu/mc-thesis/pages/videos.html.

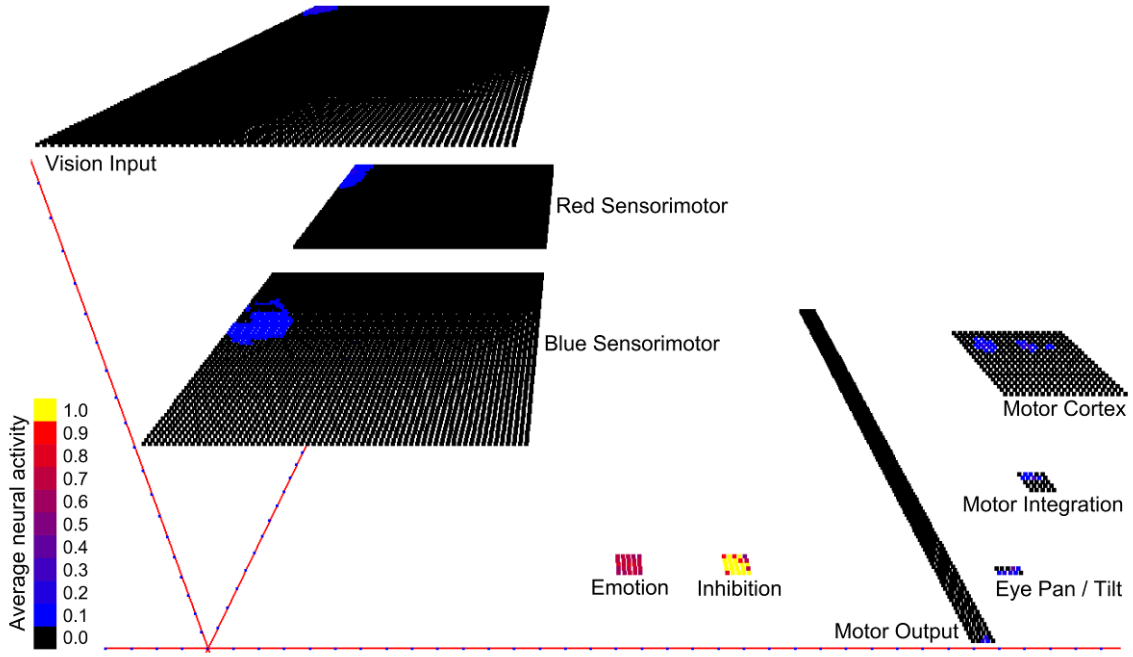

Figure 7. Normalized average firing frequency of neurons during Analysis Run

The data from Analysis Run can be used to predict the *actual* consciousness that was experienced by the network as it interacted with the world. However, in this recording only a small part of the network is active, and so it does not tell us about the consciousness that might be predicted to be associated with other parts of the network if neurons were active in these areas. To address this problem, a second recording was made in which the neuron groups were disconnected from each other and connections within each layer were cut, and 5% noise was injected into each layer at each time step for 100 time steps. The normalized average firing frequency of each neuron is illustrated in Figure 8, which shows an even spread of activity across the layers. This noise recording is referred to as “Noise Run”.

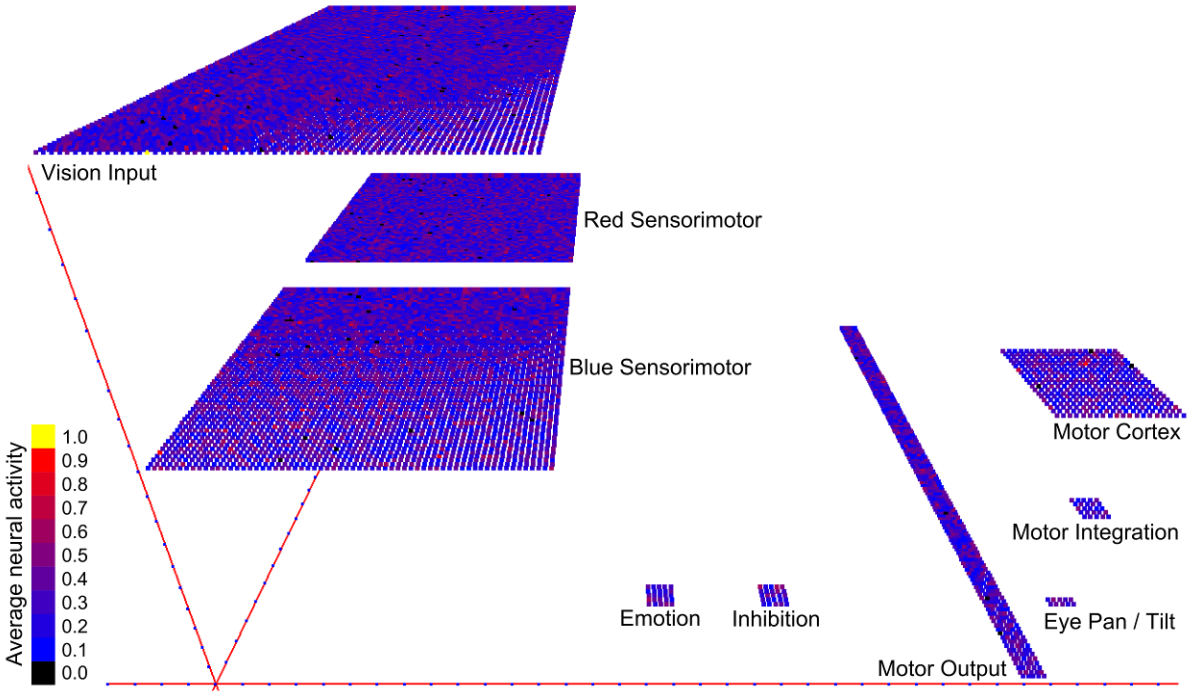

Figure 8. Normalized average firing frequency of neurons during Noise Run

Noise Run was recorded with the layers completely disconnected, and so the predictions about the consciousness of the network during Noise Run were made *as if* the noise patterns had been present when the network was fully connected. In other words, the noise data provides a useful way of understanding the *potential* for consciousness of different parts of the network: the consciousness that the network *would* have *if* it was fully connected and in that firing state.

5. Analysis of the Network for Information Integration

5.1 Introduction

As explained in Section 3, Tononi's (2004) theory links the amount and distribution of consciousness in a network to its information integration. To make predictions about the consciousness of the network according to Tononi's theory, it was necessary to calculate the network's information integration using the algorithm set out in Tononi and Sporns (2003), with a number of adjustments to take account of the network's size:

- *Entropy calculations on sub-matrices*. Tononi and Sporns' information integration Matlab code[9] calculates the entire covariance matrix and then extracts the A, B and AB sub matrices to work out the entropy. Since the complete connection matrix for the network had 17,544 rows and columns, it would have been impossible to use this approach with the available computer resources. To get around this problem, the connection matrix was generated for each bipartition and then the determinants of A, B and AB were extracted. This yielded nearly identical results to the Matlab code on the validation tests (Gamez, 2008b) and can be justified by assuming that the effect of A on B when A is in a state of maximum entropy is much greater than the effect of the rest of the system on B.

- *Normalization*. Tononi and Sporns (2003) normalized the connection matrix by multiplying the weights so that the absolute value of the sum of the afferent synaptic weights per element was a constant value, *w*, which was set to 0.5 in their analysis. Whilst this normalization method was appropriate for Tononi and Sporns' task of comparing different architectures that had been artificially evolved, it substantially distorts the relationships between the weights and does not correctly measure the information integrated by the system. For example, two neurons connected with a weight of 0.00001 have very little effective information between them, but the constant value weight normalization changes the connection weight to 0.5 and substantially alters the information exchanged between the two neurons. To avoid these problems, the connection matrix was normalized by summing each neuron's afferent weights, finding the maximum value and calculating the factor that would reduce this maximum to less than 1.0. All of the weights in the network were then

9 Tononi and Sporns' Matlab code is available at: http://tononi.psychiatry.wisc.edu/informationintegration/toolbox.html.

multiplied by this factor to ensure that the sum of each neuron's afferent weights was always less than 1.0 without distorting the relationships between them.

- *Optimization*. The key problem with Tononi and Sporns' approach is that the analysis time increases factorially with the number of subsets and bipartitions, and it was estimated that it would have taken 10^{9000} years to exhaustively analyse a network with 17,544 neurons (Gamez, 2008b). To complete the analysis in a reasonable time a number of optimization and approximation strategies were used. One of the main strategies was an algorithm that expanded the subset from a seed and progressively added neurons until the subset could not be expanded without reducing the Φ, which indicated that a complex had been found. This avoided the analysis of subsets with disconnected neurons and allowed small complexes to be identified without a large computational overhead. A second approximation strategy, suggested by Tononi and Sporns (2003), was to sub-sample the number of calculations on each subset bipartition. Even with these strategies in place the analysis took approximately a month to run on two 3.2 GHz Pentium IV computers.

- *Clusters.* To fill in the gaps left by the seed-based analysis the Φ calculations were also run on combinations of neuron groups up to a maximum size of 700 neurons – a number that was found to be a reasonable compromise between the information gained about the network and the calculation time. These group analysis results are not complexes because it has not been shown that they are not included within a subset of higher Φ. To make this distinction clear they are referred to as *clusters*.

Full details about the analysis methodology and the complete results are available in Gamez (2008b).

5.2 Information Integration Results

According to Tononi and Sporns (2003), the complex with the highest Φ is the *main* complex of the network, and Tononi (2004) claims that this is the conscious part of system. In this analysis a main complex was identified that had 91 neurons, a Φ value of 103 and included all of Inhibition, most of Emotion and small numbers of neurons from Vision Input, Red Sensorimotor, Motor Output, Eye Tilt and Motor Integration (see Figure 9). A search was carried out for independent complexes using Definition 1, and the main complex was found to be the only independent complex, with all the other complexes and clusters having some overlap with the main complex and thus not being independent by this definition.

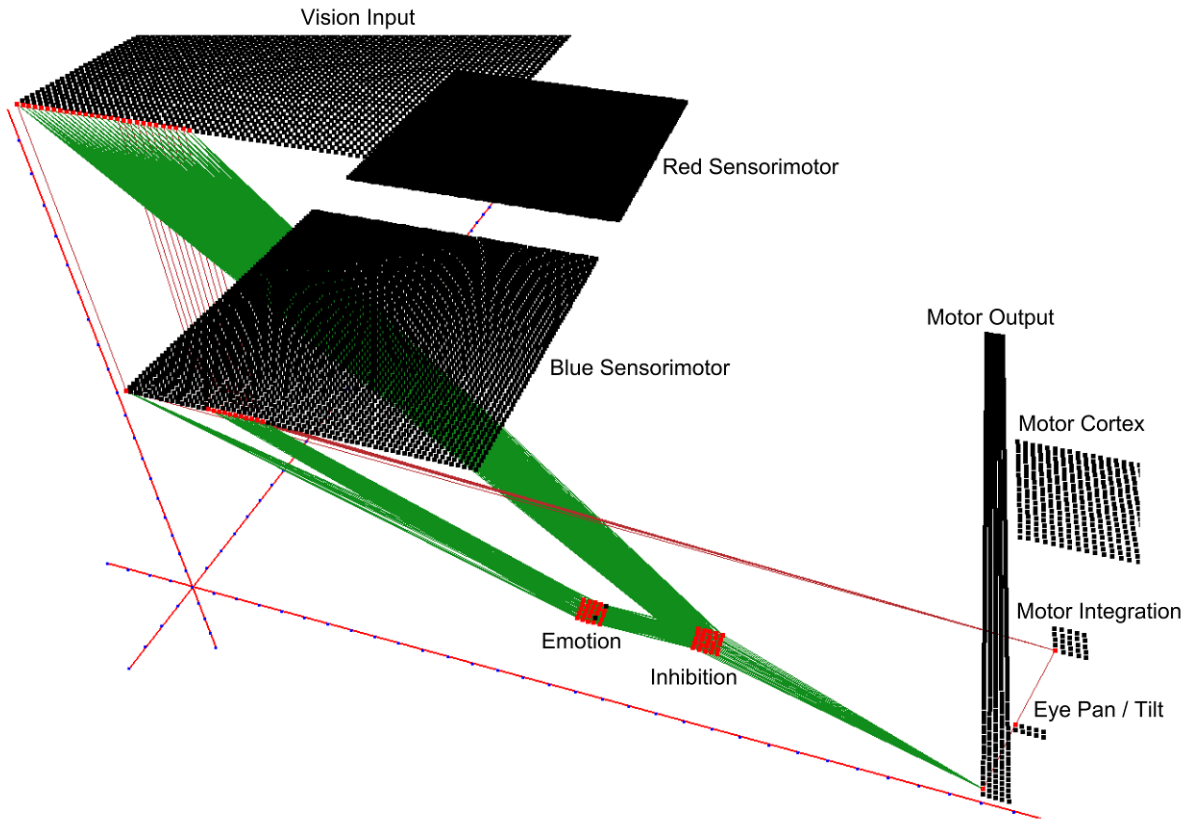

Figure 9. The main complex of the network

To understand the information integrated between different parts of the network, ten neurons were selected at random from each neuron group and the complex(es) with the highest Φ that each neuron was involved in were identified. Only the highest Φ complexes

were considered because a neuron's most significant information relationships are with the other neurons in its highest Φ complex. The analysis showed that most neurons were part of a complex with 22-435 neurons and a Φ value ranging from 57-103. The most important neuron group for information integration was Inhibition, which played a central role in many of the complexes with higher Φ.[10]

6. Predictions about Phenomenal States

6.1 Methodology

To generate predictions about the phenomenal states associated with the network, the definition in Section 3 was combined with the results from the information integration analysis and applied to the Analysis Run and Noise Run data. To visualize the predictions about the distribution of consciousness, the amount of predicted consciousness per neuron was averaged over each run, normalized to the range 0-1 and used to highlight the network in Figure 10 and Figure 11. Detailed predictions about the phenomenal states at each time step were also output as XML files, which are included in the supporting data.

6.2 Predictions about the Distribution of Consciousness in the Network According to Tononi's Theory

Tononi's (2004) link between information integration and consciousness is independent of the material that the system is made from, and so the simulated neural network described in this paper has the same potential for consciousness as a biological system. According to the interpretation of Tononi's theory in Definition 2, the predicted consciousness of the network at each point in time is the intersection of the neuron activity with the main complex. In this network the main complex had a Φ of 103 and included all of the neurons highlighted in Figure 9. In Noise Run there was fairly uniform activity across the network, and so the

10 The full information integration results are included in the supporting data.

predicted distribution of consciousness for Noise Run was an extract from the average activity shown in Figure 8 that was shaped like the main complex (see Figure 10).

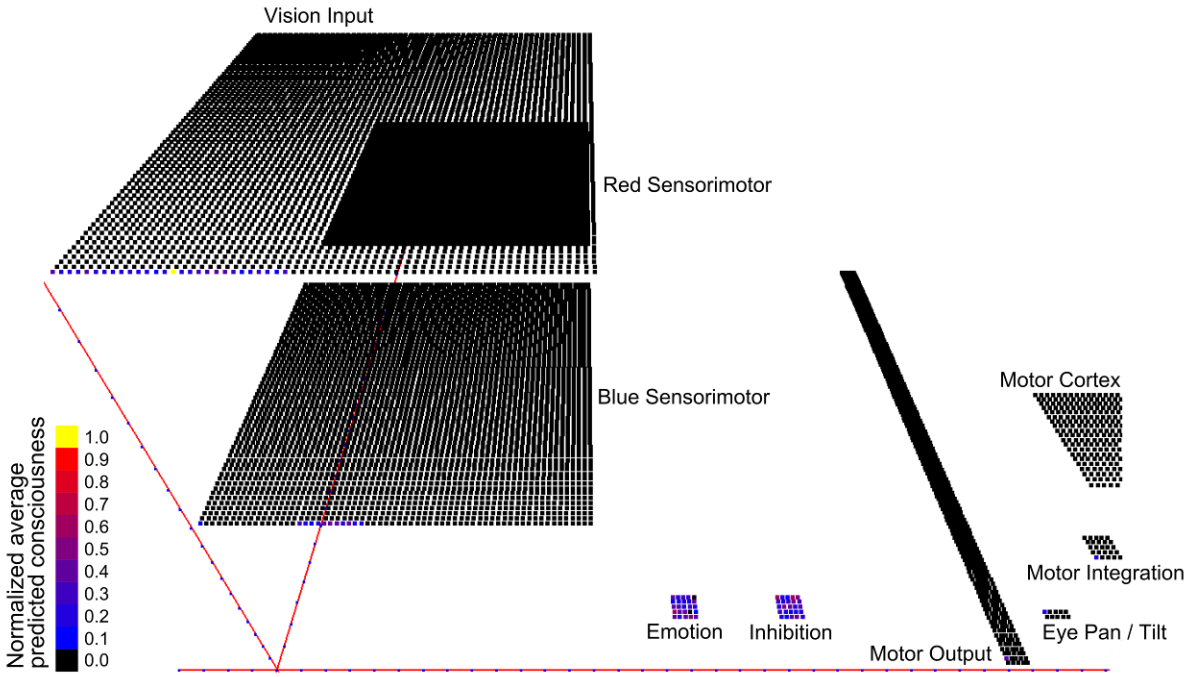

Figure 10. Predicted distribution of consciousness during Noise Run according to Tononi's theory

The more specific neuron activity during Analysis Run did not include any of the main complex neurons outside of Emotion and Inhibition, and so the predicted distribution of consciousness in Figure 11 only includes neurons in Emotion and Inhibition, with the pattern closely matching the average firing frequencies shown in Figure 7.

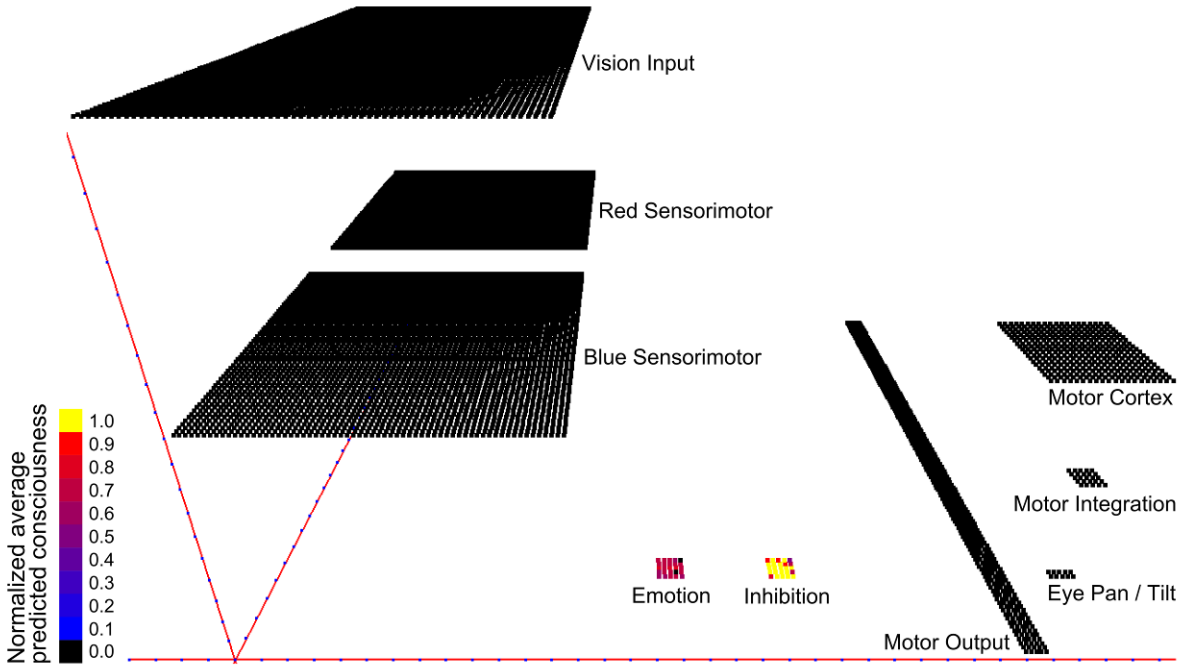

Figure 11. Predicted distribution of consciousness during Analysis Run according to Tononi's theory

7. Discussion and Future Work

One outcome of these results is that they highlight a problem with Tononi's (2004) simplistic link between the main complex and consciousness. In this network there were a number of overlapping complexes with approximately the same value of Φ as the main complex, and it seemed somewhat arbitrary to interpret just one of these as the main complex, when it was also conceivable that several overlapping complexes could be part of the same consciousness. In such a consciousness, there would be strong integration between the neurons in Inhibition and Vision Input, but low integration between different neurons in Vision Input. This appears to reflect our own phenomenology since we appear to be most conscious of our intentional relationship with the world and much less conscious of the relationships that different parts of the world have to each other. One way in which overlapping complexes could be combined would be to look at the rate of change of Φ between adjacent overlapping complexes, and use a high rate of change of Φ to indicate a boundary between the conscious and unconscious

parts of the system. This method of combining overlapping complexes with high Φ could also be applied to Tononi's more recent work (Balduzzi and Tononi, 2008), which does allow for the possibility of multiple main complexes.

One of the main limitations of this analysis was that it did not address the question about how *much* consciousness was present. Ideally the results would have stated that the network exhibited 5% of the consciousness of the average waking human brain, for example, but without calibration of the measurement scales it was impossible to say how much consciousness was associated with the network. Although Tononi (2004) claims that Φ is an absolute measure of the amount of consciousness, he has made no attempt, as far as I am aware, to estimate or measure the Φ of the main complex in an average waking human brain, and without this reference point, the Φ values are without absolute meaning.

To address this problem, more work is needed to measure or estimate the Φ of a waking human brain, so that predictions about consciousness can be compared with a system that can (at least to begin with) be taken as a reference standard. Without such a 'platinum bar' it is impossible to measure the amount of consciousness in a system using numerical methods. A first step towards obtaining these figures would be to measure the Φ of more realistic simulations, such as the networks created by the Blue Brain project (Markram, 2006). This would give some idea about the Φ values that might be found in a real biological system and help us to understand what level of consciousness might be associated with the Φ value of 103 that was found in the network described in this paper. Better ways of quantifying the amount of consciousness in a system will also go some way towards addressing the "small networks" argument put forward by Herzog et al. (2007), which suggests that many influential theories of consciousness can be implemented by very small networks of less than ten neurons that we would unwilling to attribute much consciousness to.

Once predictions have been made about a system's consciousness it is possible to suggest ways in which this consciousness can be extended or enhanced. Before any thought can be given to extending the consciousness predicted by Tononi's theory, it is essential to improve the information integration analysis to take account of overlapping complexes in a more flexible way. When this has been done, it might be possible to increase the amount and distribution of consciousness by evolving connection patterns that extend the main complex and give it a higher value of Φ.

The question about what a system is conscious *of* is a useful and important aspect of any prediction about phenomenal states. Relatively little work has been done on this in artificial systems and future work is likely to be based on the method pioneered by Hubel and Wiesel (1959), in which the responses of the internal parts of the system are recorded whilst it is exposed to different stimuli. In traditional phenomenology – for example, Husserl (1960) and Merleau-Ponty (1989) - the contents of the conscious states are described using human language, but Nagel (1974) and Chrisley (1995) have identified significant problems that arise when human language is used to describe the phenomenal states of artificial systems and non-human animals. One response to this issue has been to use a markup language to describe the states (Gamez, 2006; Gamez, 2008b) and other promising solutions include semantic maps (Ascoli and Samsonovich, 2008) and the use of concrete systems to specify the conscious states (Chrisley and Parthemore, 2007).

To compare the 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 – possibly using the work of Steels (2001, 2003). Since many people would not consider the report of an artificial system to be evidence that it is conscious, formalized theories of consciousness could be used to make predictions about the consciousness of biological systems that can report their conscious states, and these

predictions 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.

Other theories of consciousness could also be used to make predictions about the consciousness of the network described in this paper. In previous work I generated predictions about the phenomenal states of the network using Aleksander's (2005) and Metzinger's (2003) theories, which showed very different distributions from those in Figure 10 and Figure 11. A key precondition for this type of work is a formal definition of each theory, which could be a mathematical equation, an algorithm or a piece of code – the only requirement is that it takes the states of an arbitrary system as input and generates predictions about its consciousness.

8. Conclusions

This paper has argued that a greater emphasis on prediction could help some of the current research on consciousness to become more scientific. An approach to making predictions about conscious states was put forward and used to make detailed predictions about the distribution of phenomenal states in a spiking neural network according to Tononi's (2004) information integration theory. To establish whether a theory of consciousness is actually *correct*, its predictions need be compared with first person reports from systems that are known to be conscious, and this will only become possible when low level access to the human brain has improved. For the moment, artificial systems provide a good platform for the development of prediction techniques, and the process of making predictions about artificial systems can help us to refine and improve our theories about consciousness.

Acknowledgements

Many thanks to Owen Holland for feedback, support and advice about this work. The interface between SIMNOS and SpikeStream was developed in collaboration with Richard Newcombe, who designed the spike conversion methods, and I would also like to thank Igor Aleksander, Renzo De Nardi and Hugo Gravato Marques for many useful suggestions and discussions. This research was funded by the Engineering and Physical Science Research Council Adventure Fund (GR/S47946/01).

References

Aleksander, I. (2005). *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines*. Exeter: Imprint Academic.

Aleksander, I. and Morton, H. (2007). Why Axiomatic Models of Being Conscious? *Journal of Consciousness Studies* 14(7): 15-27.

Arbib, M.A. and Fellous, J.-M. (2004). Emotions: from brain to robot. *TRENDS in Cognitive Sciences* 8(12): 554-61.

Ascoli, G. A. and Samsonovich, A. V (2008). Science of the Conscious Mind. *The Biological Bulletin* 215: 204-215.

Baars, B. J. (2000). There are no known Differences in Brain Mechanisms of Consciousness Between Humans and other Mammals. *Animal Welfare* 10(1): 31-40.

Balduzzi, D. and Tononi, G. (2008). Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLOS Computational Biology* 4(6): e1000091.

Brader, J.M., Senn, W. and Fusi, S. (2006). Learning real world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Computation* 19 (11): 2881-912.

Brovelli, A, Ding, M., Ledberg, A., Chen, Y., Nakamura, R. and Bressler, S.L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proc Natl Acad Sci USA* 101: 9849-54.

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Chrisley, R. J. (1995). Taking Embodiment Seriously: Nonconceptual Content and Robotics. In K.M. Ford, C. Glymour and P.J. Hayes (eds.), *Android Epistemology*. Menlo Park, Cambridge and London: AAAI Press/ The MIT Press.

Chrisley, R. J. and Parthemore, P. (2007). Synthetic Phenomenology: Exploiting Embodiment to Specify the Non-Conceptual Content of Visual Experience. *Journal of Consciousness Studies* 14 (7): 44-58.

Cotterill, R. (2003). CyberChild: A Simulation Test-Bed for Consciousness Studies. In O. Holland (ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Crick, F. and Koch, C. (2000). The Unconscious Homunculus. In T. Metzinger (ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge, Massachusetts: The MIT Press, pp. 103-10.

Crook, J.H. (1983). On attributing consciousness to animals. *Nature* 303: 11-14.

Damasio, A. R. (1995). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Picador.

Friston, K.J., Harrison, L. and Penny, W. (2003). Dynamic causal modelling. *NeuroImage* 19: 1273–302.

Gamez, D. (2006). The XML Approach to Synthetic Phenomenology. In R. Chrisley, R. Clowes and S. Torrance (eds.), *Proceedings of the AISB06 Symposium on Integrative Approaches to Machine Consciousness*, Bristol, UK, pp. 128-35.

Gamez, D. (2007). SpikeStream: A Fast and Flexible Simulator of Spiking Neural Networks. In J. Marques de Sá, L.A. Alexandre, W. Duch and D.P. Mandic (eds.), *Proceedings of ICANN 2007*, Lecture Notes in Computer Science Volume 4668, Springer Verlag, pp. 370-9.

Gamez, D. (2008a). Progress in Machine Consciousness. *Consciousness and Cognition* 17(3): 887-910.

Gamez, D. (2008b). *The Development and Analysis of Conscious Machines*. Unpublished PhD thesis, University of Essex, UK. Available at: http://www.davidgamez.eu/mc-thesis/.

Gamez, D. and Aleksander, I. (2009). Taking a Mental Stance Towards Artificial Systems. *Biologically Inspired Cognitive Architectures. Papers from the AAAI Fall Symposium*. AAAI Technical Report FS-09-01, forthcoming.

Gamez, D., Newcombe, R., Holland, O. and Knight, R. (2006). Two Simulation Tools for Biologically Inspired Virtual Robotics. *Proceedings of the IEEE 5th Chapter Conference on Advances in Cybernetic Systems*, Sheffield, pp. 85-90.

Gazzaniga, M.S. (1970). *The Bisected Brain*. New York: Appleton-Century-Crofts.

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models*. Cambridge University Press, Cambridge.

Haynes, J. -D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology* 17(4): 323-328.

Herzog, M.H., Esfeld, M. and Gerstner, W. (2007). Consciousness & the small network argument. *Neural Networks* 20: 1054–6.

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

Honey, C.J., Kötter, R., Breakspear, M. and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *PNAS* 104(24): 10240–5.

Hubel, D.H. and Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology* 148(3): 574–91.

Husserl, E. (1960). *Cartesian Meditations*. Translated by Dorion Cairns. The Hague: Nijhoff.

Izhikevich, E. M. & Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proc Natl Acad Sci U S A* 105: 3593-8.

Kay, K. N., Naselaris, T., Prenger, R. J. and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* 452: 352–355.

Krichmar, J. L. and Edelman, G. M. (2006). Principles Underlying the Construction of Brain-Based Devices. In T. Kovacs, and J.A.R. Marshall (eds.), *Proceedings of AISB'06: Adaptation in Artificial and Biological Systems*, Bristol, UK, pp. 37-42.

Krichmar, J. L., Nitz, D. A., Gally, J. A. and Edelman, G. M. (2005). Characterizing functional hippocampal pathways in a brain-based device as it solves a spatial memory task. *PNAS* 102(6): 2111-6.

Lee, U., Mashour, G. A., Kim, S., Noh, G.-J and Choi, B.-M. (2009). Propofol induction reduces the capacity for neural information integration: Implications for the mechanism of consciousness and general anesthesia. *Consciousness and Cognition* 18(1): 56-64.

Linåker, F. and Niklasson, L. (2000). Time series segmentation using an adaptive resource allocating vector quantization network based on change detection. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000)*, pp. 323-8.

Marian, I. (2003). *A biologically inspired computational model of motor control development*. Unpublished MSc Thesis, University College Dublin, Ireland.

Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience* 7: 153-60.

Merleau-Ponty, M. (1989). *Phenomenology of Perception*. Translated by C. Smith. London: Routledge.

Metzinger, T. (2003). *Being No One*. Cambridge, Massachusetts: The MIT Press.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A. and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* 320: 1191–1195.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review* 83: 435-56.

O'Regan, J.K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24: 939-1031.

Popper, K. (2002). *The Logic of Scientific Discovery*. London and New York: Routledge.

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

Schreiber, T. (2000). Measuring Information Transfer. *Physical Review Letters* 85(2): 461-4.

Seth, A.K. (2009). The Strength of Weak Artificial Consciousness. *International Journal of Machine Consciousness* 1(1): 71-82.

Seth, A. K. & Edelman, G. M. (2004). Environment and behavior influence the complexity of evolved neural networks. *Adaptive Behavior* 12: 5-20.

Seth, A. K., Izhikevich, E., Reeke, G. N. and Edelman, G. M. (2006). Theories and measures of consciousness: An extended framework. *PNAS* 103(28): 10799–804.

Shanahan, M.P. (2006). A Cognitive Architecture that Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition* 15: 433-49.

Sporns, O. (2007) Brain Connectivity. *Scholarpedia* 2(10):4695.

Sporns, O., Chialvo, D.R., Kaiser, M. and Hilgetag, C.C. (2004). Organization, development and function of complex brain networks. *TRENDS in Cognitive Sciences* 8(9): 418-25.

Sporns, O. & Lungarella, M. (2006). Evolving coordinated behavior by maximizing information structure. In L. Rocha, L. Yaeger, M. Bedau, D. Floreano, R. L. Goldstone and A. Vespigniani (eds.), *Artificial Life X: Proceedings of the 10th International Conference on the Simulation and Synthesis of Living Systems*, Cambridge, MA: MIT Press, pp. 322-329.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems and Their Applications* 16(5): 16-22.

Steels, L. (2003). Language Re-Entrance and the 'Inner Voice'. In O. Holland (ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Stening, J., Jacobsson, H. and Ziemke, T. (2005). Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model. In R. Chrisley, R. Clowes and S. Torrance, (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness*, Hatfield, UK.

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:42.

Tononi, G., Edelman, G.M. and Sporns, O. (1998). Complexity and coherency: integrating information in the brain. *Trends in Cognitive Sciences* 2(12): 474-84.

Tononi, G. and Sporns, O. (2003). Measuring information integration. *BMC Neuroscience* 4:31.

Tononi, G., Sporns, O. and Edelman, G.M. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* 91: 5033-7.

Witelson, S.F., Glezer, I.I. and Kigar, D.L. (1995). Women Have Greater Density of Neurons in Posterior Temporal Cortex. *Journal of Neuroscience* 15(5): 3418-28.

Information Integration Based Predictions about the Conscious States of a Spiking Neural Network – Supplementary Material

David Gamez

1. Neuron and Synapse Model

The neuron model for these experiments was based on the Spike Response Model (Gerstner and Kistler, 2002; Marian, 2003), which has three components: a leaky integrate and fire of the weights of incoming spikes, an absolute refractory period in which the neuron ignores incoming spikes, and a relative refractory period in which it is harder for incoming spikes to push the neuron beyond its threshold potential. The resting potential of the neuron is zero and when it exceeds the threshold the neuron is fired and the contributions from previous spikes are reset to zero. There is no external driving current and the voltage V_i at time t for a neuron i that last fired at $\hat{t}$ is given by Equation 1:

$$V_i(t) = \sum_j \sum_f w_{ij} e^{-\frac{(t-t_j^{(f)})}{\tau_m}} - e^{n-(t-\hat{t}_i)^m} \mathrm{H}'(t-\hat{t}_i), \quad (1)$$

where w_{ij} is the synaptic weight between i and j, τ_m is the membrane time constant, f is the last firing time of neuron j, m and n are parameters controlling the relative refractory period, and H' is given by Equation 2:

$$\mathrm{H}'(t-\hat{t}_i) = \begin{cases} \infty, & if\ 0 \le (t-\hat{t}_i) < \rho \\ 1, & otherwise \end{cases}, \quad (2)$$

for an absolute refractory period ρ. To facilitate the learning algorithm, the neuron model also contained a variable c that represented the calcium concentration at time t. Each time the

neuron fired, this calcium concentration was increased by C_S and it decayed over time according to Equation 3, where C_D is the calcium decay constant.

$$c(t) = \sum_i C_S e^{-\frac{t-\hat{t}_i}{C_D}} \quad (3)$$

The thresholds were adjusted in each neuron group until the network produced the desired behaviour. The values for the other neuron parameters were based on Marian (2003) and Brader et al. (2006), and are given in Table 1. The synapse model was very basic, with each synapse class passing its weight to the neuron when it received a spike.

Parameter	**Value**
C_S	1
C_D	60
P	1 ms
τ_m	1
M	0.8
N	3
Minimum postsynaptic potential	-5

Table 1. Parameters common to all neurons

2. Learning

Learning in the network was carried out using Brader et al.'s (2006) spike time dependent learning algorithm. In Brader et al.'s model the internal state of the synapse is represented by *X(t)* and the efficacy of the synapse is determined by whether *X(t)* is above a threshold. In the model used in this paper, the state of the synapse is represented by a weight variable, *w*, which is the amount by which the post-synaptic membrane potential is increased when the

neuron fires. When a spike is received at time t_{pre}, this variable w is changed according to equations 4 and 5:

$$w \rightarrow w + a \quad if \quad V(t_{pre}) > \theta_V \quad and \quad \theta_{up}^{l} < c(t_{pre}) < \theta_{up}^{h} \tag{4}$$

$$w \rightarrow w - b \quad if \quad V(t_{pre}) \leq \theta_V \quad and \quad \theta_{down}^{l} < c(t_{pre}) < \theta_{down}^{h} \ , \tag{5}$$

where a and b are jump sizes, θ_V is a voltage threshold, $c(t)$ is the calcium concentration at time t, and θ_{up}^{l}, θ_{up}^{h}, $\theta_{down}^{l}$and θ_{down}^{h} are thresholds on the calcium variable. In the absence of a pre-synaptic spike or if the two conditions in equations 4 and 5 are not satisfied, the weight changes at the rate given by equations 6 and 7:

$$\frac{dw}{dt} = \alpha \quad if \quad w > \theta_w \tag{6}$$

$$\frac{dw}{dt} = -\beta \quad if \quad w \leq \theta_w, \tag{7}$$

where α and β are positive constants and θ_w is a threshold. If w drops below 0 or exceeds 1, then it is held at these boundary values. The parameters that were used for training the network are given in Table 2. These parameters were initially set using Brader et al. ‘s (2006) values and fine tuned until the network successfully learnt the association between motor output and visual input.

Parameter	Value
θ_{up}^{l}	4
θ_{up}^{h}	120
θ_{down}^{l}	0
θ_{down}^{h}	4
θ_{V}	0.4
a	0.01
b	0.01
θ_{w}	0.7
α	0.00001
β	0.00001

Table 2. Synapse parameters used during training

References

Brader, J.M., Senn, W. and Fusi, S. (2006). Learning real world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Computation* 19 (11): 2881-912.

Gerstner, W. and Kistler, W. (2002). *Spiking Neuron Models*. Cambridge University Press, Cambridge.

Marian, I. (2003). *A biologically inspired computational model of motor control development*. Unpublished MSc Thesis, University College Dublin, Ireland.