7. ANALYSIS

7.1 Introduction

This chapter describes how the neural network in Chapter 5 was analyzed for consciousness using the approach to synthetic phenomenology set out in Chapter 4. The first section in this chapter covers the calculation of the OMC rating of the network, Section 7.3 explains the method that was used to identify the representational mental states, and then Section 7.4 describes the analysis of the system for information integration using Tononi and Sporns' (2003) approach. Sections 7.5 - 7.7 look at whether the network is capable of consciousness according to Tononi's, Aleksander's and Metzinger's theories and definitions are formulated that enable the network to be automatically analyzed for phenomenal states. The final part of this chapter describes how the network's activity was recorded and combined with the analysis data to produce a sequence of XML files that predict the phenomenology of the system according to the three theories of consciousness.

All of this analysis was carried out on two 3.2 GHz Pentium IV computers with 2 GB RAM. The code for this analysis is all part of the Network Analyzer software, which was written as part of this PhD and is briefly covered in Appendix 2. No official release of Network Analyzer is planned, but the source code for the current version is included in the supporting materials.

7.2 OMC Rating

In this network all of the mental states are implemented in the same way, and so they all have the same rating on version 0.6 of the OMC scale described in Section 4.2. The system is a

biologically inspired simulated neural network running on two single processor computers at a speed that is significantly slower than the human brain, and so its factors are S1, R2, F1, FN4, TS2, AD3, giving a total weighting of 3.025×10^{-3}. This needs to be multiplied by 0.1 to compensate for the missing level of molecules, atoms and ions, leading to a final weighting of 3.0×10^{-4}, which is an OMC position of 111 out of 192 on the scale, and an OMC rating of 0.43. This OMC rating makes intuitive sense because the final arrangement of atoms and electrons in the system is substantially different from that in a human brain, but not to the extent that it is impossible to conceive that it has conscious states. This OMC rating is incorporated into the XML description of the phenomenology in Section 7.9.

7.3 Identification of Representational Mental States

7.3.1 Definition of a Mental State for this System

In this analysis a simulated neural network is being analysed for consciousness, and so the mental states are states of the simulated network.[1] Depending on how a neural network is modelled, there are many different ways of defining its states – for example, the spiking activity of a population of neurons, the voltages in the neurons, the average neuron firing rates, changes in memory addresses or activity in the processor and RAM – and in this analysis, it was decided to treat the firing of a neuron as a mental state. Although this is fairly basic, the main purpose of this analysis is to illustrate how synthetic phenomenology can be carried out, and it would have been unnecessarily complicated to use population codes or memory addresses to make predictions about the network's phenomenal states.

[1] See Section 4.3.2 for the definition of a mental state that is being used in this thesis.

7.3.2 Selection of Method

To identify the representational mental states of the network a method was needed that could identify the functional or effective connections between the input and output data and the internal states (see Definition 4.2 in Chapter 4). In this network the input and output pass through Vision Input and Motor Output, and so I decided to look at the functional and effective connections between these layers, which had known response characteristics, and the internal layers whose responses were not known. The first problem that had to be addressed was that a complete map of the representational states of the network was required for the XML description, and yet the network only activated a small selection of its possible states during normal activity. To get around this problem it was decided to inject noise into the layers that had known response characteristics, and use an algorithm or mathematical method to identify the functional or effective relationships between activity in the neurons with known response characteristics and activity in the internal neurons whose representational characteristics were being measured.

One of the first algorithms that I considered was the backtracing method developed by (Krichmar et. al. 2005), which examines the firing rate of a reference neuron at a specific time step and identifies the neurons connected to the reference neuron that were active during the previous time step. Whilst it might have been possible to trace the spikes back through the network in this way, the recurrent loops and delays in the network would have made this process extremely complicated. Another method that was considered was Granger causality (Seth and Edelman 2007), but this would have required conversion of the spiking activity into average firing rates, which I wanted to avoid if possible. Instead, it was decided to use mutual information to measure the relationships between the input/ output and internal neurons, and the next section describes how this can be calculated from the spiking activity. Although mutual

information does not directly measure causal relationships, under these experimental conditions a strong case can be made that it is a measure of effective connectivity (see Section 7.3.6).

7.3.3 Identification of Representational Mental States Using Mutual Information

The first step in the analysis for representational mental states was the selection of an input or output layer, which was given one description in natural human language and another in terms applicable to the physical world (when this could be done reasonably easily). Next, noise was injected into the input or output layer and the network activity was recorded. This data was then used to calculate how much mutual information each internal neuron shared with the input or output neurons that had been given the physical and human descriptions. This procedure was repeated separately for each input and output layer that had response characteristics that could be easily described. In theory this noise injection technique could be also used to identify mental states that represent other mental states, but the difficulty of describing internal neuron groups led me to exclude meta representational mental states from this analysis.[2]

The mutual information between each input/output neuron, X, and each internal neuron, Y, was calculated by recording the number of times that the following combinations occurred for different steps back in time ("1" indicates that the neuron was firing at that time step and "0" indicates that the neuron was quiescent):

$x = 0 \ \& \ y = 0$

$x = 1 \ \& \ y = 0$

$x = 0 \ \& \ y = 1$

$x = 1 \ \& \ y = 1$

These statistics enabled the joint probabilities to be calculated:

[2] Meta representational mental states would have been needed to analyze the network using Rosenthal's (1986) higher order thought theory. However for the reasons discussed in Section 2.3.2 this theory was not used in this analysis.

$p(x = 0,\ y = 0)$

$p(x = 1,\ y = 0)$

$p(x = 0,\ y = 1)$

$p(x = 1,\ y = 1)$

for different steps back in time as well as the marginal probabilities:

$p(x = 0)$

$p(x = 1)$

$p(y = 0)$

$p(y = 1)$.

Using these values, the mutual information between each input/output neuron X and each internal neuron Y was calculated using the standard formula for mutual information:

$$I(X;Y) = \sum_{x=0}^{1}\sum_{y=0}^{1} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{7.1}$$

Equation 7.1 was also used to work out the maximum possible mutual information under the experimental conditions. With 20% of the neurons being fired randomly at each time step:

$p(x = 0) = 0.8$

$p(x = 1) = 0.2$

$p(y = 0) = 0.8$

$p(y = 1) = 0.2$.

When the mutual information between X and Y is at a maximum, their state will always be the same, and so:

$p(x = 0,\ y = 1) = 0$

$p(x = 1,\ y = 0) = 0$,

and the remaining joint probabilities can be derived from the noise:

$p(x = 0,\ y = 0) = 0.8$

$p(x = 1, y = 1) = 0.2.$

Putting these figures into Equation 7.1 yields a maximum possible mutual information of 0.72.

During the recording of the data a time step value of 10 ms was used to avoid complications caused by the refractory period[3] and all other sources of network noise were switched off. The injection of 20% noise into each input/ output layer[4] and approximately 10,000 time steps of recorded activity were found to give mutual information values that matched expectations based on the known connectivity of the network. Since the mutual information between two neurons is rarely zero, a threshold was used to eliminate low mutual information values that would have been superfluous in the final XML description. The results for Vision Input and Motor Integration are covered in the next two sections.

7.3.4 Visual Representational Mental States

Vision Input was an obvious choice of input layer for the visual analysis because it could be easily labelled and had strong forward connections to the rest of the network. With layers of several thousand neurons the analysis for representational mental states consumes a lot of time and memory because the mutual information has to be calculated for each combination of input/ output and internal neurons. This problem can be reduced by excluding layers from the analysis that are unlikely to have any systematic link with the layer that is being used as input or output For the visual analysis Motor Cortex was excluded because it did not have any input connections from other layers, and Motor Integration, Eye Pan and Eye Tilt were also left out because they did not have any direct or indirect connections from Vision Input. The mutual information between the input and internal neurons was calculated for between zero and five

[3] The total refractory period of the neurons is approximately 10 ms.

[4] Noise injection in this part of the analysis was done by firing a random selection of 20% of the neurons at each time step.

steps back in time because activity in Vision Input took four time steps to propagate to Motor Output.

The next stage in the visual analysis was to decide on appropriate labels for the neurons in Vision Input. Since half of Vision Input carries red visual information and half carries blue, it was decided to include both red and blue in the human description and to use the corresponding wavelengths of light in the physical description.[5] The parameters for the identification of visual representational mental states are summarised in Table 7.1.

Parameter	**Value**
Input Neuron Group	Vision Input
Internal Neuron Groups	Emotion, Red Sensorimotor, Blue Sensorimotor, Inhibition, Motor Output
Human Description	"Red / blue visual input"
Physical Description	"700 nm / 450 nm electromagnetic waves"
Steps back in time	0 - 5
Mutual Information Threshold	0.1
Input Neuron Group Noise	20%

Table 7.1. Parameters for the analysis of visual representational mental states

The data structures were too large to fit in memory, and so the input/output and internal layers were split into five groups and the mutual information calculations were run on the 25 possible combinations between them, which took several days to complete.[6] A high level summary of the average mutual information shared between Vision Input and the internal layers is plotted in Figure 7.1.

[5] A more sophisticated analysis could have distinguished between light wavelengths and perceived colours when assigning the human and physical labels.

[6] This separation into separate groups did not have any effect on the final result.

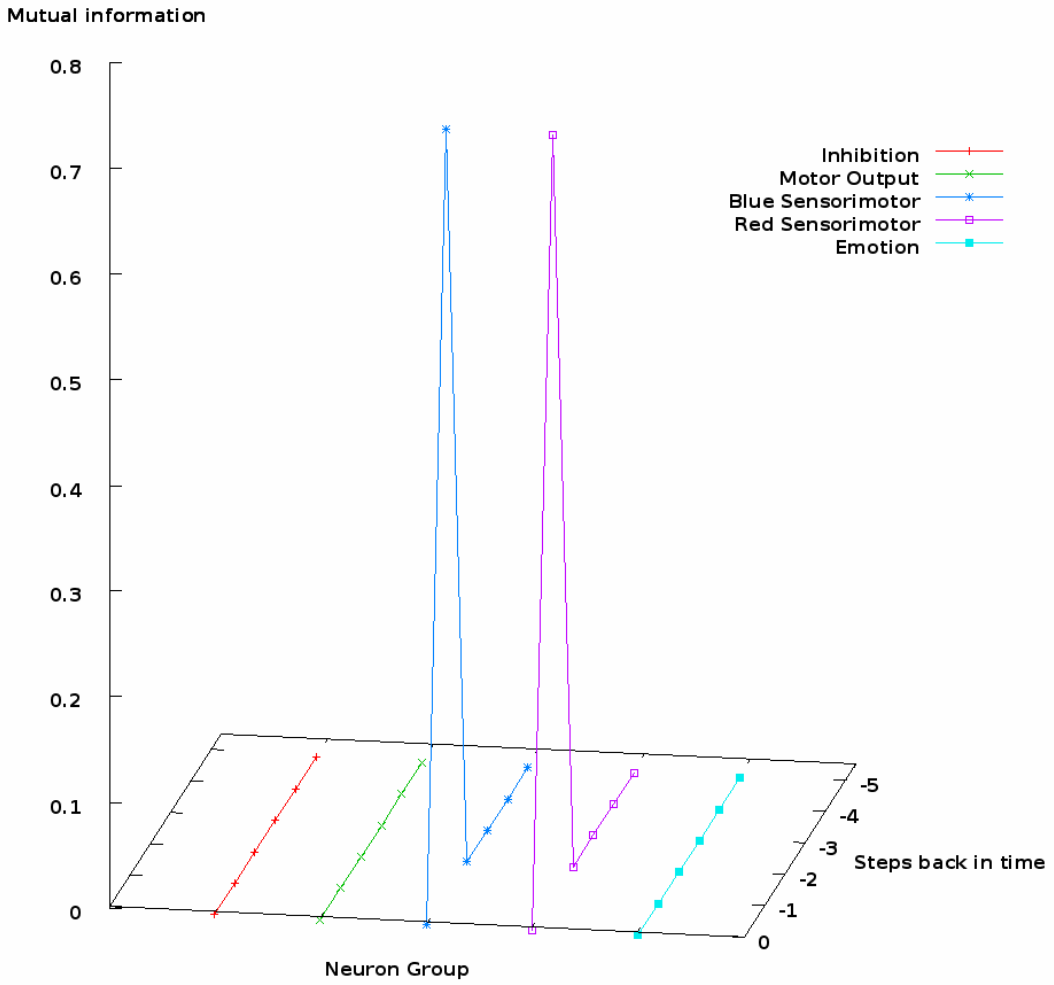

Figure 7.1. Average mutual information shared between Vision Input and the internal layers during the analysis for representational mental states

The results in Figure 7.1 show that the mutual information shared between Red Sensorimotor and Blue Sensorimotor and Vision Input was close to the theoretical maximum of 0.72, which matched expectations because of the strong topological connections between Vision Input and Red/ Blue Sensorimotor. Although Emotion is indirectly connected to Vision Input, it shared no mutual information above the threshold, which was probably due to the large number of internal connections within Emotion that made its self-sustaining activity largely independent of Red Sensorimotor. The other neuron groups downstream of Emotion, such as Inhibition and Motor Output, also shared no mutual information with Vision Input.

7.3.5 Proprioception/ Motor Output Representational Mental States

Motor Output would not have been a good choice of output layer for the motor analysis because it does not have any forward connections to the other layers, and no representational mental states would have been found by injecting noise into it. A better choice was Motor Integration, which has forward connections to other layers, contains a complete map of all possible motor combinations and plays a key role in action selection through its connections to Red Sensorimotor and Blue Sensorimotor. Motor Integration can also be given a clear human description because it maps directly down to Motor Output through Eye Pan and Eye Tilt. Motor Cortex and Vision Input were excluded from this part of the analysis because they did not have any incoming connections from other layers.

Although the neural network does not receive sensory data from SIMNOS's joints or muscles, the motor control signals sent from Motor Integration accurately predict the position of the eye after a delay of a few time steps, and so activity in this layer encodes both proprioceptive and motor information. To reflect this dual role, "Proprioception / motor output" was chosen as the human description of the neurons in Motor Integration and the physical description was set to "N/A" because it would have been too complicated to describe the physical movements of the eye in response to activity in this layer. There is an 11 time step delay from Motor Integration to Red Sensorimotor and Blue Sensorimotor, and so the mutual information between the input and internal neurons was calculated for between zero and twelve steps back in time. Although this had the effect of excluding potential representational links between Motor Integration and Emotion via Red/ Blue Sensorimotor, the visual analysis strongly suggested that there was no representational relationship between Emotion and Red/ Blue Sensorimotor. The parameters for the identification of proprioception/ motor output representational mental states are summarized in Table 7.2.

Parameter	Value
Input Neuron Group	Motor Integration
Internal Neuron Groups	Emotion, Red Sensorimotor, Blue Sensorimotor, Inhibition, Eye Tilt, Eye Pan, Motor Output.
Human Description	"Proprioception / motor output"
Physical Description	"N/A"
Steps back in time	0 - 12
Mutual Information Threshold	0.1
Input Neuron Group Noise	20%

Table 7.2. Parameters for the analysis of proprioception/ motor output representational mental states

The data structures for the proprioception/ motor output analysis fitted comfortably in memory and the calculations took less than an hour to complete. A high level summary of the average mutual information shared between Motor Integration and the internal layers is plotted in Figure 7.2.

In the results shown in Figure 7.2 Motor Output shows a small response with a peak of 0.01 at minus two time steps, which might have been expected to be higher since there is an indirect link between Motor Integration and Motor Output. However, the value of 0.01 represents the *average* mutual information between Motor Integration and Motor Output, and only 10 out of 675 neurons in Motor Output are indirectly connected to Motor Integration. The highest average mutual information is shared between Motor Integration and Eye Pan and Eye Tilt at -1 time steps. This is close to the theoretical maximum and it is due to the topographic connections between Motor Integration and Eye Pan and Eye Tilt. There are also average mutual information peaks for Red Sensorimotor and Blue Sensorimotor at -12 time steps, which matched expectations since there is a connection with a delay of 11 time steps from Motor Integration to Red/ Blue Sensorimotor and it takes one time step for a spike to be emitted from one group and processed in another.

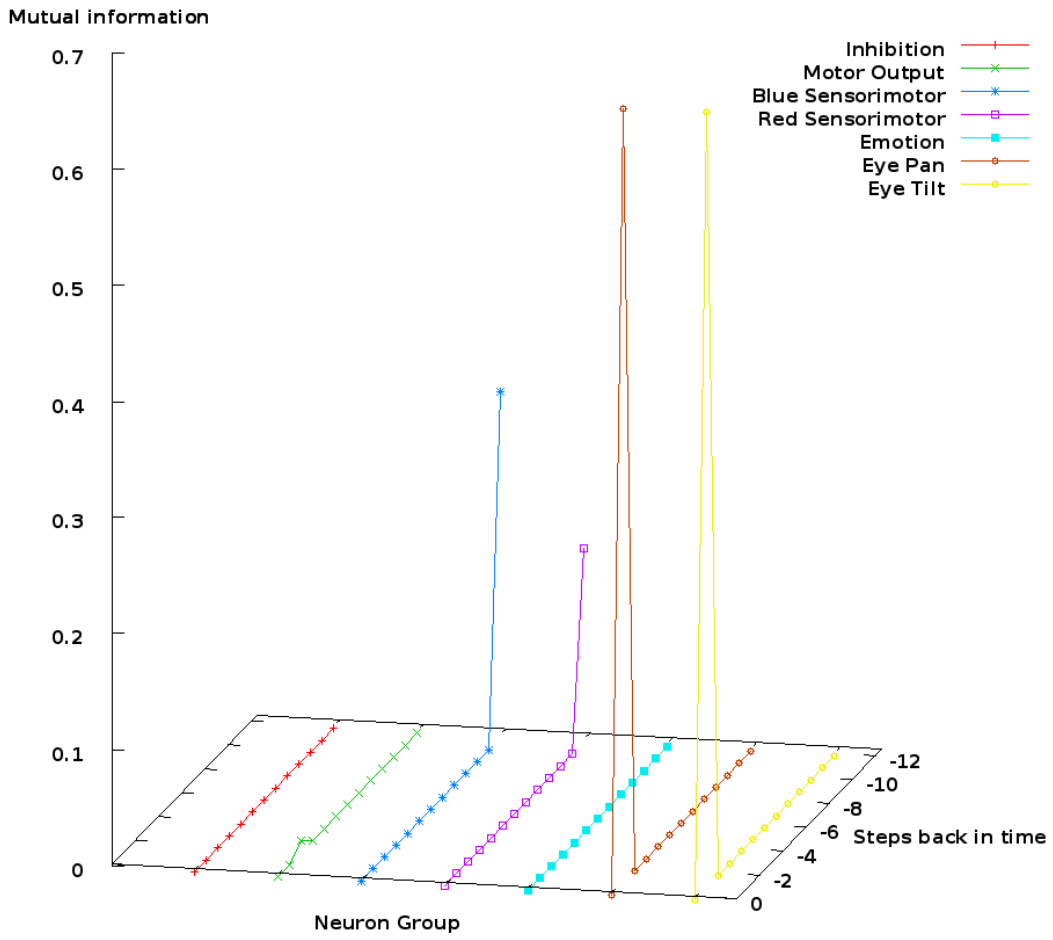

Figure 7.2. Average mutual information shared between Motor Integration and internal layers during the analysis for representational mental states

The graphs in Figure 7.1 and Figure 7.2 only give the average mutual information that is shared between the input/output and internal layers. The detailed results can be found in the VisualRepresentationalMentalStates.xml and MotorRepresentationalMentalStates.xml files, which are included in the supporting materials.

7.3.6 Representational Mental States: Discussion and Future Work

One limitation of mutual information is that by itself it is not a measure of the causal relationships between neurons. If two neurons, *A* and *B*, share mutual information, then it could be because *A* is causally influencing *B*, *B* is causally influencing *A* or because *A* and *B* are under the causal influence of a third neuron, *C*. However, in the method described Section 7.3.3 there is

good reason to believe that the activity of the internal neurons is due to their causal dependencies on the input/ output layers because the input/ output layers are individually put into a high (but not maximum) state of entropy and there is no other source of spontaneous activity within the network. A second reason why this method is likely to measure causal dependencies is because the mutual information is calculated for different numbers of steps back in time. If the mutual information between *A* and *B* peaked at time step zero, for example, then this would suggest that *A* and *B* were under the causal influence of a third neuron, *C*, but if *B* shares maximum mutual information with *A* at -2 time steps, then it is more likely that there is a causal relationship between *A* and *B* - although *A* and *B* could still be subject to a common cause *C* that is connected to A and B with different delays. Finally, the close match between the mutual information relationships and the structure and delays of the network makes it reasonable to assume that the internal neurons sharing high mutual information with input/output neurons are causally dependent on these input/output neurons.

This analysis did not attempt to identify mental states that represent other mental states because the descriptions would have been too complicated to define at both the human and physical levels. Future work in this area might be able to track the processing of data through the network by repeating the analysis a number of times at different levels. For example, mental states that responded to a combination of motor output and blue visual information could be injected with noise to discover representational mental states that respond to more abstract higher level information. In this way meta representational mental states could be described as combinations of more basic mental states that are linked to states of the world. Mental states representing more complex features of the world could also be identified using more specific test data.

The visual and motor systems of this network were extremely basic and on such a simple system the injection of noise into Vision Input and Motor Integration was a good way of

identifying representational mental states. However, a more subtle approach would be needed to identify representational mental states in more complex systems. If the system is *designed* with layers that respond to different aspects of the input signal - for example the visual input layers in Krichmar et. al. (2005) – then the layers could be individually labelled and injected with noise to identify the representational mental states. However, when the system's responses to complex aspects of the world are not known – for example, in self-organizing networks, such as the hippocampus in Krichmar et al. (2005) – then it might be possible to use the statistical methods developed by Lungarella et. al. (2005) and Lungarella and Sporns (2006) to identify regularities in the input and output signals, which could be used to label the representational mental states.

In the future it would be worth exploring whether other techniques, such as transfer entropy (Schreiber 2000, Sporns et al. 2006), backtracing (Krichmar et. al. 2005) and Granger causality (Seth and Edelman 2007), make different predictions about the representational mental states of the network. It would also be worth investigating how the definition of a system's mental states affects its representations. For example, if mental states were defined in terms of populations of neurons, then Kohonen (2001) or one of Grossberg's (1976) neural networks could be used to identify patterns in the neuron populations, and the mutual information shared between these patterns and the input/ output data could then be measured using the noise injection method.

7.4 Information Integration Analysis

7.4.1 Introduction

This section describes how the neural network was analyzed for information integration using Tononi and Sporns' (2003) method. The main motivation for this analysis was to make predictions about the consciousness of the network using Tononi's (2004) information integration theory of consciousness. The phenomenology of a system also depends on the

integration between the different pieces of conscious information (see Section 4.3.6), and since information integration is a measure of effective connectivity (Sporns 2007), it made sense to use Tononi and Sporns' method to identify the integration between the mental states in the network. Information integration is also used in the predictions about the consciousness of the network based on Metzinger's (2003) theory (see Section 7.7.3).

The central difficulty with Tononi and Sporns' (2003) method is that the analysis time increases factorially with the number of subsets and bipartitions, which makes it impossible to exhaustively analyse systems with more than fifty elements. To find out the scale of this problem, Section 7.4.3 gives an estimate of how long the full analysis would take on a network with 17,544 neurons. Since this is of the order of 10^{9000} years, optimisation strategies had to be developed for large networks, which are documented in Section 7.4.4, and Section 7.4.5 gives the result of testing these optimizations on Tononi and Sporns' (2003) sample networks. The remaining information integration sections present the results and some background and future work. Further details about the information integration results are included in Appendix 3.

7.4.2 Tononi and Sporns' Information Integration Calculation

As explained in Section 2.6.2, the complexes of a system are identified by considering 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 possible bipartition of the subset, the effective information integrated across the bipartition, $EI(A \leftrightharpoons B)$, is calculated and the minimum normalized effective information, $\min\{ EI(A \leftrightharpoons B) / H^{MAX}(A \leftrightharpoons B)\}$, is identified. The non-normalized minimum effective information is the Φ value of the subset, and a complex is a subset with $\Phi > 0$ that is not included in a larger subset with greater Φ. At the centre of this method is the calculation of $EI(A \leftrightharpoons B)$, which is repeated a large number of times during the analysis. The stages in the calculation of $EI(A \leftrightharpoons B)$ are as follows.

Normalization

The starting point for the EI(A⇋B) calculation is the connection matrix, CON(X), which is an *m* x *m* matrix representing all of the connections between the *m* elements of the subset. In this analysis all of the weights were made positive by multiplying them by -1 on the grounds that a connection is passing information regardless of whether it is excitatory or inhibitory.[7] Without this normalization of negative weights it is conceivable that the positive and negative connections between the two bipartitions of a subset would have partially cancelled each other out, leading to a value of EI(A⇋B).that did not reflect the amount of information that was exchanged between the two bipartitions.[8]

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 they set to 0.5 for their analysis Whilst this normalization method was appropriate for Tononi and Sporns' task of comparing different architectures that have 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, this analysis normalized the connection matrix 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 multiplied by this factor, which ensured that the sum of

[7] The alternative method of adding a constant to all of the weights was rejected because it would have made positive connections count for more, when in fact positive and negative connections with the same weight were transmitting the same amount of information

[8] I have not been able to find any discussion of negative weights in Tononi and Sporns (2003) or Tononi (2004) and their examples are all based on positive weights.

each neuron's afferent weights was always less than one without distorting the relationships between them.

Covariance Matrix

In each effective information calculation one part of the subset, A, is put into a state of maximum entropy and the entropy of the response of B is used to calculate EI(A→B). Since A is being substituted by independent noise sources, all of the self connections within A and the connections afferent to A are set to zero within CON(X). Under Gaussian assumptions, the elements in the system can be represented by a vector X of random variables that are subject to independent Gaussian noise R of magnitude c. When the elements settle under stationary conditions, the final state of the system is given by Equation 7.2:

$$X = X * CON(X) + cR. \tag{7.2}$$

Using standard algebra and averaging over the states produced by successive values of R, this equation can be rearranged as:

$$X = R\,(1\text{-}CON(X))^{-1}, \tag{7.3}$$

and a substitution of:

$$Q = (1\text{-}CON(X))\text{-}1 \tag{7.4}$$

into Equation 7.3 gives the formula:

$$X = RQ. \tag{7.5}$$

In Equation 7.5, the elements of R that correspond to the A bipartition of the subset are set to 1.0 to put A into a state of maximum entropy, and the elements of R that correspond to the B

bipartition are set to a value that corresponds to the background noise, which is typically 0.00001. Using the standard covariance formula:

$$COV(X) = X^T X, \tag{7.6}$$

and substituting in Equation 7.5, we obtain:

$$COV(X) = (RQ)^T RQ, \tag{7.7}$$

which is equivalent to Equation 7.8:

$$COV(X) = Q^T R^T RQ, \tag{7.8}$$

and can be calculated from CON(X) using standard matrix operations.

Entropy

EI(A⇋B) depends on the entropies H(A), H(B) and H(AB), which can be calculated by extracting the sub matrices COV(A), COV(B) and COV(AB) from the covariance matrix and putting their determinants into Equation 7.9:

$$H(X) = \frac{\ln((2\pi e)^n \mid COV(X) \mid)}{2}, \tag{7.9}$$

where | COV(X) | is the determinant of COV(X).[9]

EI(A⇋B)

The effective information from A to B, EI(A→B), is given by the mutual information between A and B when A is in a state of maximum entropy:

[9] This standard formula for calculating the entropy from the determinant of a covariance matrix can be found in Papoulis (1984, p. 541).

$$MI(A^{HMAX}:B) = H(A^{HMAX}) + H(B) - H(A^{HMAX}B), \quad (7.10)$$

which can easily be calculated from the entropy values. The process is then repeated in the opposite direction by putting B into a state of maximum entropy to calculate EI(B→A), and the final value of effective information is given by:

$$EI(A \leftrightharpoons B) = EI(A \rightarrow B) + EI(B \rightarrow A). \quad (7.11)$$

This is normalized by $H^{MAX}(A \leftrightharpoons B)$ to enable different bipartitions to be compared, and the information integration for subset S, or Φ(S), is the non-normalised value of $EI(A \leftrightharpoons B)$ for the minimum information bipartition.

The C++ code for these calculations was based on Tononi and Sporns' Matlab toolkit.[10] The most substantial change was that the Matlab code calculates Q^TR^TRQ on the whole connection matrix and then extracts the A, B and AB sub matrices to work out the entropy. Since the complete connection matrix has 17,544 rows and columns, this approach would have been impossible with the computer resources available in this project. 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 (see Section 7.4.5) 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. A brief overview of the Network Analyzer software is given in Appendix 2.

[10] The Matlab complexity toolbox is available at: http://tononi.psychiatry.wisc.edu/informationintegration/toolbox.html.

7.4.3 Time for the Full Information Integration Analysis

The full information integration analysis is computationally expensive because the EI(A⇋B) calculations are processor-heavy matrix operations that have to be run on every bipartition of every possible subset of the network. The first part of the full analysis is the extraction of all the possible subsets of the network, with the number of ways of selecting *m* elements out of the *n* elements of the system being given by:

$$\frac{n!}{m!(n-m)!}, \tag{7.12}$$

which has to be summed over all subset sizes from $m = 2$ to $m = n$.

The next part of the full analysis is the calculation of EI(A⇋B) on every possible bipartition of each subset in order to find the minimum information bipartition. A bipartition is created by selecting *k* elements out of the *m* elements in the subset, where k ranges from 1 to $m / 2$. Putting the subset selections together with the bipartition selections gives:

$$t_{analysis} = \sum_{m=2}^{n} \sum_{k=1}^{m/2} \frac{n!}{m!(n-m)!} \frac{m!}{k!(m-k)!} f(m), \tag{7.13}$$

where $t_{analysis}$ is the full analysis time and $f(m)$ is the time taken to calculate EI(A⇋B) on a single bipartition of a subset of size *m*. By cancelling out *m*! Equation 7.13 can be rearranged as follows:

$$t_{analysis} = n! \sum_{m=2}^{n} \sum_{k=1}^{m/2} \frac{1}{(n-m)!} \frac{1}{k!(m-k)!} f(m). \tag{7.14}$$

Equation 7.14 omits the fact that when the number of neurons in each half of the bipartition is exactly the same, the number of possible bipartitions has to be divided by two, because the selection of all possible combinations in one half results in the selection of all possible

combinations in the other half. This adjustment was included in the code that was used to estimate the full analysis times.

The time taken for each EI(A⇋B) bipartition calculation depends on a number of factors, including the efficiency of the code and the speed of the computer, and an estimate of this value was obtained by recording the average time that each EI(A⇋B) calculation took on subsets of different sizes (see Figure 7.3).

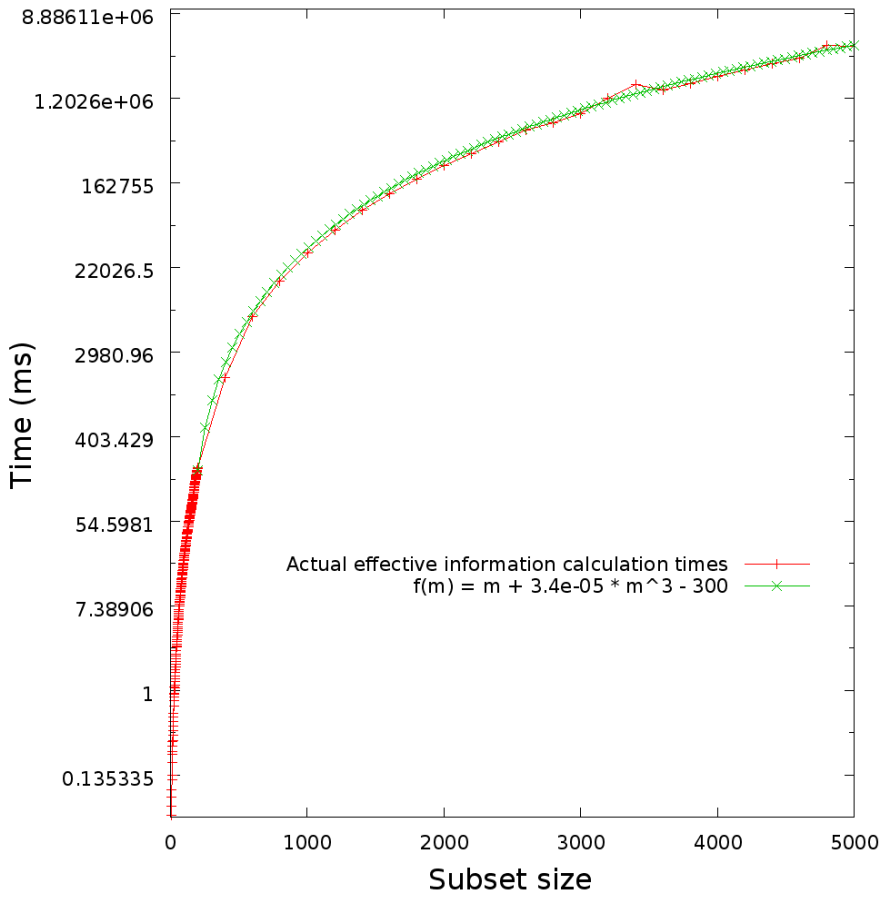

Figure 7.3. Actual and predicted times for each EI(A⇋B) bipartition calculation on subsets of different sizes

The curve fitting functions of gnuplot suggested that:

$$f(m) = m + 3.4e^{-5} \times m^3 - 300 \qquad (7.15)$$

was a good approximation to the actual values for $m > 200$ and Equation 7.15 was combined with the actual EI(A⇋B) calculation times to predict the bipartition calculation times for subsets

with 2 - 5000 neurons. A short piece of code was then written that used the bipartition calculation times, Equation 7.14 and the adjustment for equal bipartitions to calculate $t_{analysis}$ for networks of different sizes, and the results from this calculation are plotted in Figure 7.4.[11]

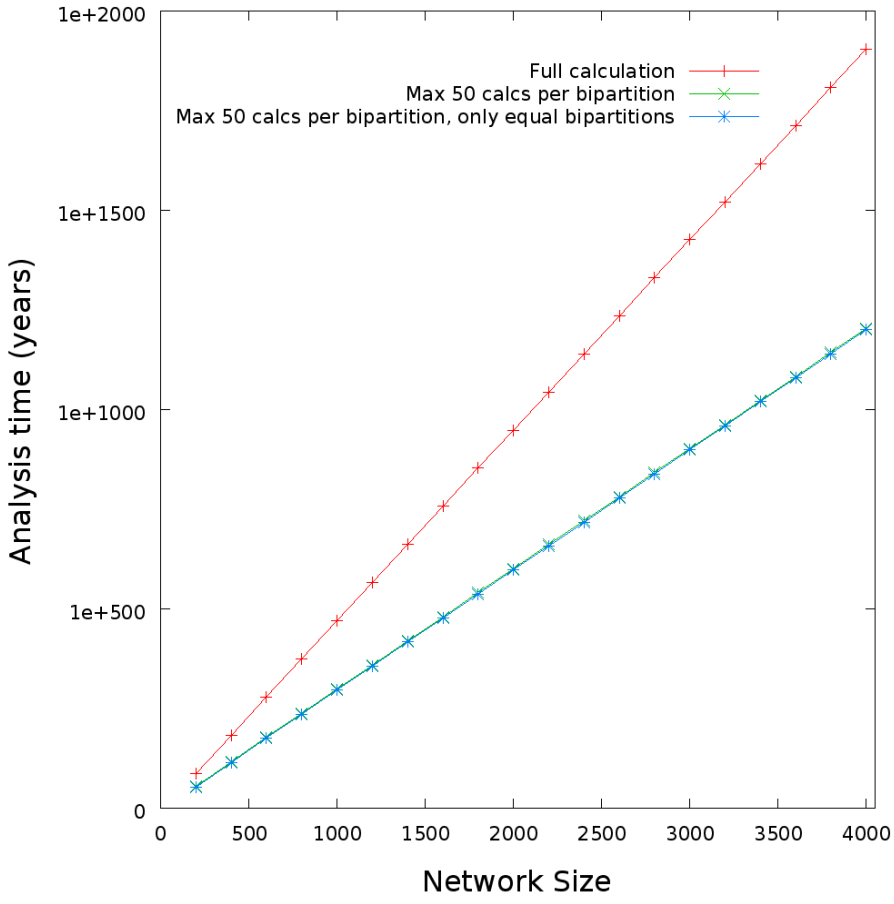

Figure 7.4. Predicted full information integration analysis times for networks of different sizes

Figure 7.4 only shows the predicted times for networks up to 4000 neurons because the factorial calculations took an increasingly long time to run as the network size increased and it was unclear whether it would reach 18,000 neurons within a reasonable time. However, the linear relationship between network size and the log of the calculation time can be extrapolated up to 18,000 neurons to give a predicted full analysis time of around 10^{9000} years. This shows that a full information integration analysis would have been completely impossible on a 17,544

[11] The data in Figure 7.4 was generated by the TimeCalculator class in Network Analyzer, which is included in the Supporting Materials.

neuron network with my current equipment. This difficulty is acknowledged by Tononi and Sporns (2003), who admit that "Practically, the procedure for finding complexes can only be applied exhaustively to systems composed of up to two dozen elements" (p. 18). The optimization and approximation strategies that were used to address this problem are discussed next.

7.4.4 Optimizations and Approximations

Given the extremely large amount of time that would have been required for the full analysis, it was necessary to develop optimizations and approximations that could identify some of the complexes in the network with the limited time and computer resources that were available to this project.

Sub-sampling

One approximation suggested by Tononi and Sporns (2003) is to evaluate EI(A⇋B) on a random selection of the possible bipartitions at each subset level. For example, to take 15 samples at each level for a 200 neuron subset, one would evaluate EI(A⇋B) for 15 samples of the 1:199 bipartition, 15 samples of the 2:198 bipartition, and so on up to 15 samples of the 100:100 bipartition. Although Tononi and Sporns suggested using 10,000 sub samples per level, the duration of each bipartition calculation suggested that orders of magnitude less sub-samples would have to be used if the calculation was going to complete in a reasonable time.

The impact of this approximation strategy is shown in Figure 7.4, where the blue line plots the predicted analysis times when the number of bipartition calculations per level is limited to 50, and the timings for the group analyses in Table A3.12 demonstrate that this approximation strategy is effective in practice. The disadvantage of this approximation is that it can dramatically reduce the proportion of bipartitions that are examined for the minimum information bipartition, which leads to a substantial loss of accuracy. In Network Analyzer, this

approximation is implemented as the *Max number of bipartitions per level* parameter. The current version of the code examines the permutations of each bipartition in an ordered sequence up to the maximum limit; in the future it might be better to select the permutations at random.[12]

Another way of reducing the number of bipartition calculations is to sub-sample the levels. For example, in a subset of 200 neurons, this could involve sampling the 20:180, 40:160 … 100:100 bipartitions instead of every possible level. Although this option was included in the Network Analyzer code as the *Percentage of bipartition levels* parameter, it was rarely used in practice.

Seed expansion method

A second strategy, developed in collaboration with Richard Newcombe at Essex, is to grow each complex incrementally from a seed. To begin with a single neuron is selected at random, either from the entire network or from one of the neuron groups in the network. Next, one or a number of neurons connected to this seed are added to the subset and the Φ of the subset is calculated. If the new Φ is greater than the old one, the neurons are left in the subset and the process is repeated again. On the other hand, if the new Φ is less than the old one, then the new neurons are removed from the subset and the process is repeated with a different set of connected neurons. When all of the connections to and from the seed have been explored, the connections to and from other neurons in the subset are tried until the subset cannot be expanded any further. The remaining subset is likely to be a complex because any larger subset with greater Φ that includes the subset would have to be connected to it, and it has been shown that the addition of further connected neurons decreases the subset's Φ. The steps in the seed method are summarised in Figure 7.5:

[12] Random selection was not done in the current analysis because of the extra processing that would have been required to calculate the full range of permutations and make a random selection from it.

```
1.  Start a new subset by choosing a neuron to act as the seed.
2   Select numNeur neurons connected to neurons in the subset that have
    not been selected before.
3.  if ( numNeur > 0 ) //Have found neurons connected to the subset
4.        Add neurons to the subset.
5.  else //No neurons connected to the subset – it is a complex
6.        Store details about the complex and return to step 1
7.  Calculate the new Φ of the subset, newPhi.
8.  if ( newPhi < oldPhi ) //Adding the neurons has reduced the Φ
9.        Delete the added neurons and return to step 2.
10. else //Adding the neurons has increased the Φ
11.       Leave the neurons in the subset, set oldPhi equal to newPhi and
          return to step 2.
```

Figure 7.5. Seed expansion algorithm

One advantage of the seed method is that it avoids all subsets with disconnected neurons and a Φ value of 0, whereas Tononi and Sporns' full analysis checks all subsets regardless of whether they contain disconnected neurons. The seed method also provides a way of identifying small complexes in large networks and it enables a limit to be set on the maximum size of the complexes, which is very useful for controlling the analysis duration.

The seed method does suffer from a number of potential and actual problems. To begin with, it can miss complexes that include subsets with higher Φ – for example the large complex in Tononi and Sporns (2003, Figure 7) was missed by this method (see Section 7.4.5). However, this was not a problem in the current analysis, which only aimed to identify the highest Φ complex that each neuron was involved in. A second disadvantage of the seed method is that the order of expansion may affect the final complex and in future work it would be worth doing some experiments to see if this is a significant effect. Finally, the seed expansion algorithm can lead to multiple calculations of Φ on the same subset, particularly when the neurons are highly connected together. Although this did occasionally happen during the analysis, it was not found to be a major issue.

A number of parameters were included in Network Analyzer to control the speed and accuracy of the seed expansion method:

- *Expansion rate per connection group*. This controls the number of neurons that are added to the subset at step 2 of the algorithm. Higher values of this parameter enable larger complexes to be identified in a shorter time, but smaller complexes may be missed when the expansion rate is greater than 1.

- *Maximum subset size*. The subset is discarded if it expands beyond this limit. This parameter is useful for searching for small complexes within a large neuron group and it was used extensively in this analysis because many seeds expanded into subsets that exceeded the available time and processing power.

- *Maximum number of consecutive expansion failures per connection group*. Some neural networks have large homogenous connections and the effect of adding one neuron from a homogenous connection group is likely to be the same as adding another neuron from the same group. When the number of failed attempts exceeds this limit, the entire connection group is discarded. For example, the network in this thesis has over 8000 connections with identical weights from each neuron in Inhibition to Vision Input. If the first twenty connections cannot be used to expand the subset, there is little reason to think that the next 8000 will, and it is more efficient to abandon the attempt to expand the connection group.[13]

- *Store Φ calculations*. When several neurons in the subset connect to the same external neuron, the same Φ calculation may be repeated several times and it might be thought that storing the results would be a good way to speed up the analysis.

[13] A variation of this approximation is to sample a random selection of neurons from a homogenous connection group. This option is available in Network Analyzer, but it was superseded by the consecutive expansion failures parameter.

However, this approach was not used in practice because the number of repeated calculations was not that large and it took a significant amount of processing to compare the neuron IDs in the current subset with the neuron IDs in each of the stored calculations.

Equal bipartitions

Another optimization strategy suggested by Tononi and Sporns (2003) is that "the bipartitions for which the normalized value of EI will be at a minimum will be most often those that cut the system in two halves, i.e. midpartitions" (p. 17). To evaluate how often mid partitions yield the minimum normalized effective information, seeds were selected from six of the layers and allowed to expand into a complex or up to a maximum subset size of 200 neurons. The percentage of times that the each bipartition had the minimum normalized effective information is plotted in Figure 7.6, which shows that mid partitions most often had the minimum normalized EI, but this was by no means always the case and during one of the seed expansions the mid partition only accounted for 40% of the minimum information bipartitions. When this approximation was applied in combination with the seed expansion method it was found that the occasional wrong expansion had a substantial effect on the final complex, and so this approximation was not used in the final analysis - although the timings presented in Section A3.3 show that the equal bipartition approximation can speed up the analysis by a factor of ten.

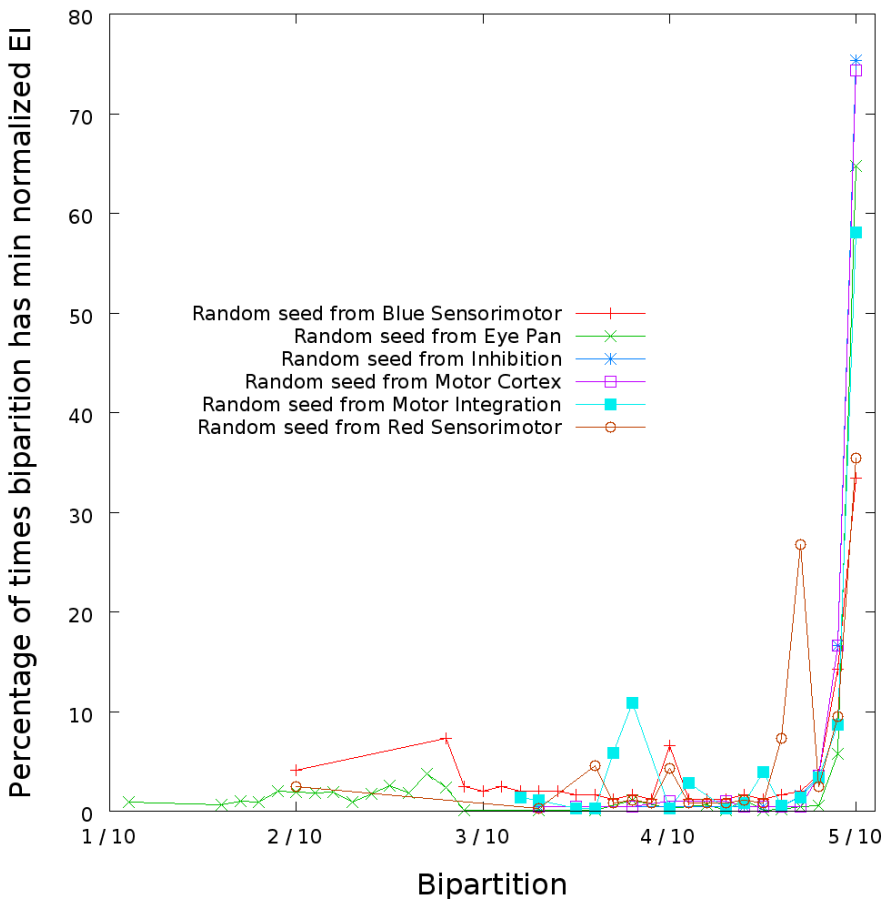

Figure 7.6. Percentage of times that different bipartitions had the minimum normalized effective information

Final strategy

The time taken to expand the seeds from each layer depends heavily on the complexes that are present in the network. For example, although Vision Input has 8,192 seed neurons, the analysis could be completed in 4.5 days because it identified a large number of complexes of approximately 30 neurons that were relatively quick to analyze. On the other hand, Inhibition has only 25 neurons, but it took 3.5 days to analyze because each seed neuron in this group had to be expanded up to the maximum subset size of 150 neurons. Since the complexes in the network were unknown at the start of the analysis, one or two test runs had to be carried out on each neuron group to identify the parameters that would enable the analysis to complete in a reasonable time. The seed expansion was then restarted on the neuron group and allowed to run to completion.

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 time available. These group analysis results are not complexes because it has not been shown that they are not included within a subset of higher Φ, and to make this distinction clear they will be referred to as *clusters*.

Although the seed and group analyses were carried out with a high level of approximation, enough information was gathered about the complexes and clusters of the network to allow predictions to be made about the network's phenomenology in Section 7.9. In the future if more accurate information about the complexes of the network could be obtained, then it would be easy to re-generate the predictions about consciousness using the improved information integration data.

7.4.5 Validation on Tononi and Sporns' Examples

The Network Analyzer code and the seed expansion method were tested on the examples supplied by Tononi and Sporns (2003) using the parameters given in Table 7.3.[14] These tests were mainly intended to establish that the seed expansion method could find the same complexes as the full analysis, and so the approximations were disabled by setting *Maximum number of consecutive expansion failures per connection group* to 1000 (greater than the number of connections in any of the examples) and *Max number of bipartitions per level* to 5000 (greater than the maximum number of possible bipartitions for this network). The results for this validation are given in Table 7.4.

[14] The connection matrices for the validation analysis were downloaded from: http://tononi.psychiatry.wisc.edu/informationintegration/toolbox.html.

Parameter	Value
Max number of bipartitions per level	5000
Percentage of bipartition levels	100
Expansion rate per connection group	1
Maximum subset size	20000
Maximum number of consecutive expansion failures per connection group	1000
Only examine equal bipartitions	false

Table 7.3. Parameters for the validation on Tononi and Sporns' examples

Example Network	Seed Expansion Algorithm		Tononi & Sporns (2003) Analysis	
	Neurons	Φ	Neurons	Φ
Figure 2	1,2,3,4	20.8	1,2,3,4	21
	5,6,7	20.1	5,6,7	20
	1,2,3,4,5,6,7,8	7.4	1,2,3,4,5,6,7,8	7
Figure 3	1,2,3,4,5,6,7,8	73.9	1,2,3,4,5,6,7,8	73
	1,4	19.1	-	-
	3,5	19.6	-	-
Figure 4	1,2,3,4,5,6,7,8	5.8	1,2,3,4,5,6,7,8	5.8
	3,6	1.8	-	-
Figure 5	1,2,3,4,5,6,7,8	60.8	1,2,3,4,5,6,7,8	60
	1,2,3,4,6,7	40.5	-	-
	5,8	20.3	-	-
Figure 6	1,2,3,4,5,6,7,8	20.5	1,2,3,4,5,6,7,8	20.5
Figure 7	1,2	20.3	1,2	20.5
	3,4	20.3	3,4	20.5
	5,6	20.3	5,6	20.5
	7,8	20.3	7,8	20.5
	-	-	1,2,3,4,5,6,7,8	19.5

Table 7.4. The complexes found in Tononi and Sporns' (2003) example networks by the full analysis and using the seed expansion algorithm. The quoted Φ values for Tononi and Sporns' (2003) analysis are approximate readings from the graphs in their figures.

The results in Table 7.4 show that the seed expansion algorithm finds most of the complexes that were identified in Tononi and Sporns (2003) and that the Network Analyzer code performed accurate Φ calculations. However the seed expansion algorithm did identify a number of false complexes in figures 3, 4 and 5, and since none of the other approximations were being used, the most likely explanation is that the order of expansion of the neurons altered the complexes.[15] The only results from the information integration analysis that are used in the predictions about consciousness are the highest Φ complexes that each neuron is involved in (see Section 7.4.6). From this perspective the identification of false complexes is not a problem as long as the larger complexes with higher Φ that incorporate the smaller complexes are also found. On these examples, all of the highest Φ complexes were correctly identified by the seed expansion algorithm and the false complexes could have been easily eliminated by post-processing the seed analysis results.[16]

The only other disparity between the results from the seed algorithm and the full analysis are that the expansion algorithm can miss complexes that include smaller complexes with higher Φ – see the last row of the results in Table 7.4. This was also not a problem in an analysis which is only looking for the highest Φ complex that is associated with each neuron.

7.4.6 The Information Integration of the Network

Since there was a great deal of overlap between the different complexes and clusters, the results from the seed and group analyses were integrated together to identify the main complex, the independent complexes and the information integration between different parts of the network. More detailed results from the seed and group analyses and illustrations of some of the

[15] For example, suppose that the subset contains two neurons, A and B, and A is connected to another two neurons, C and D. It might be the case that adding C before D reduces the Φ of the subset, whereas adding C after D causes the Φ value of the subset to increase. It is also possible that adding C or D individually to the subset reduces its Φ, whereas adding both together increases it.

[16] For example, in the figure 3 example in Table 7.4, the seed method claims that neurons 1 and 4 form a complex with a Φ value of 19.1 and that these neurons are also part of another complex with Φ = 73.9. According to the definition of a complex, it is easy to see that the complex containing only neurons 1 and 4 is a false complex.

complexes are given in Appendix 3, and the results are also available in XML format in the Supporting Materials. To present the results as clearly as possible the neuron groups in figures 7.7 - 7.15 are labelled using the IDs in Table 7.5, which correspond to the IDs that were used for these neuron groups in the database.

ID	Neuron Group
24	Vision Input
28	Red Sensorimotor
29	Blue Sensorimotor
62	Emotion
34	Inhibition
61	Motor Cortex
60	Motor Integration
54	Eye Pan
55	Eye Tilt
53	Motor Output

Table 7.5. Neuron group IDs

According to Tononi and Sporns (2003) the *main* complex of the network is the one with the highest Φ. In this network the main complex has 91 neurons, a Φ value of 103 and it includes 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 7.7). Tononi (2004) claims that the main complex is the conscious part of the network.

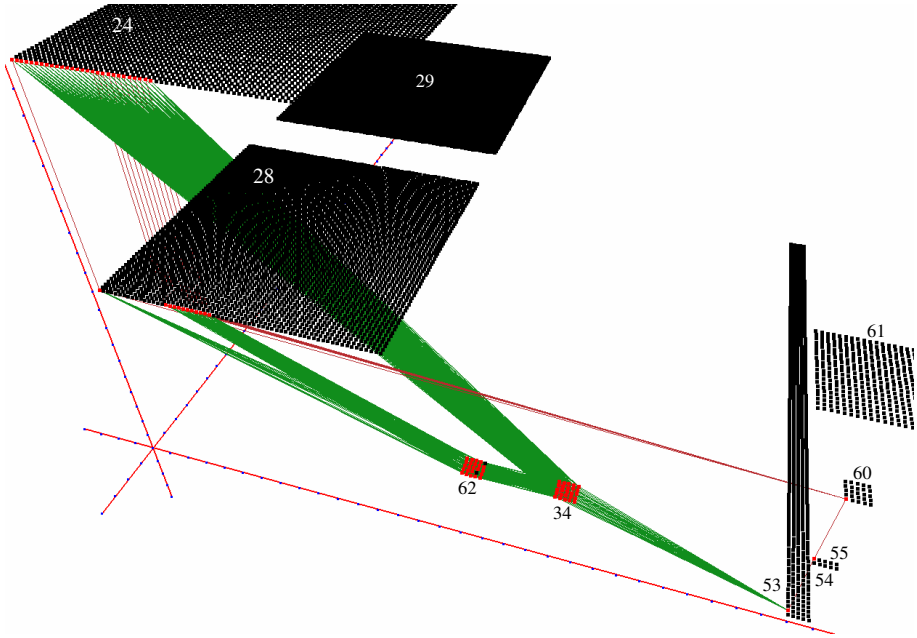

Figure 7.7. The main complex of the network.

A second aspect of information integration is whether different parts of the network integrate their information in isolation from each other (see Section 4.3.6). In this analysis, the notion of an *independent* complex is defined as follows:

> *None of the neurons in an independent complex, A, are part of another complex, B, that has higher Φ than A.* (7.1)

This definition was used to search for independent complexes in the network, and it turned out that the main complex was the only independent complex, with all of the other complexes and clusters having some overlap with the main complex and thus not being independent by this definition.

In order to understand the information integration 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 the phenomenal predictions in sections 7.5 and 7.7 are based on the maximum information integration of each mental state, and the most significant information

relationships of each neuron will be with other neurons in its highest Φ complex. The results from this analysis were as follows:

Vision Input

All of the sampled neuron's highest Φ complexes included Inhibition and different combinations of Blue Sensorimotor, Red Sensorimotor, Emotion and Motor Output. Amongst the sampled neurons, the typical highest Φ complex contained 29-31 neurons, with Φ ranging from 75-93.

Red Sensorimotor

All of the sampled neuron's highest Φ complexes included Inhibition and Vision Input, along with different combinations of Blue Sensorimotor, Emotion, Motor Integration and Motor Output. Amongst the sampled neurons, the typical highest Φ complex contained 29-31 neurons, with Φ ranging from 75-93.

Blue Sensorimotor

All of the sampled neuron's highest Φ complexes included Inhibition and Vision Input, along with different combinations of Blue Sensorimotor, Emotion, Motor Integration and Motor Output. Amongst the sampled neurons, the typical highest Φ complex contained 29-31 neurons, with Φ ranging from 75-93.

Emotion

Although this neuron group was strongly integrated with itself, higher values of Φ were found in complexes that included Inhibition and other layers. The sampled neurons' highest Φ complex was the main complex.

Inhibition

All of the sampled neuron's highest Φ complexes were part of the main complex. The Inhibition layer is a key part of many high Φ complexes because of its recurrent connections and its large number of strong connections to Vision Input and Motor Output. On its own Inhibition has a Φ of 77.3 and this increases to more than 129 when it is combined with a small number of neurons from other layers.[17]

Motor Cortex

Despite a large number of recurrent connections, Motor Cortex only had a Φ value of 17.9 when it was measured by itself. The sampled neurons had two highest Φ clusters: one with Φ = 59 and 425 neurons from Motor Cortex and Motor Integration, and another with Φ = 59 and 435 neurons from Motor Cortex, Motor Integration, Eye Pan, and Eye Tilt.

Motor Integration

One of the sampled neurons in Motor Integration had 129 highest Φ complexes with Φ=75 and 25 neurons from other layers. Some of the other highest Φ complexes of the sampled neurons had 75-91 neurons and Φ ranging from 84-103. Motor Integration also had sampled neurons that were not included in any of the complexes identified by the seed-based analysis. These had two highest Φ clusters: one with Φ=58.7 and 425 neurons from Motor Cortex and Motor Integration, and another with Φ=58.7 and 435 neurons from Motor Cortex, Motor Integration, Eye Pan and Eye Tilt.

Eye Pan

One of the seeds in this layer expanded beyond the maximum subset size of 150 and its highest Φ value came from the group analysis, which identified two highest Φ clusters: one with Φ=58.7

[17] Some of the subsets expanded from Motor Output included Inhibition and achieved a Φ value of 129 before the maximum subset size was exceeded.

and 425 neurons from Motor Cortex and Motor Integration, and another with Φ=58.7 and 435 neurons from Motor Cortex, Motor Integration, Eye Pan and Eye Tilt. The other four neurons in this group had highest Φ complexes with 75-79 neurons from all of the other layers and Φ ranging from 84–102.

Eye Tilt

The sampled neuron's highest Φ complexes had 71-91 neurons from some or all of the other layers and Φ ranging from 80–103.

Motor Output

The sampled neuron's highest Φ complexes had Φ=57 and 22 neurons from Inhibition. Ten of the neurons in Motor Output, which were not included in the random sample, are connected through Eye Pan and Eye Tilt into complexes with Φ up to 103.

These results show that the highest Φ complexes of neurons in different layers have a consistent level of information integration that typically ranges from 58 - 103. The most important neuron group for information integration was Inhibition, which played a central role in many of the complexes with higher Φ.

7.4.7 Previous Work on Information Integration

Evidence for a link between information integration and consciousness was provided by Lee et al. (2007), who made multi-channel EEG recordings from 8 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. To begin with, Seth et al. showed that simple Hopfield-type networks can be designed to have arbitrary values of Φ, which suggests that Φ may not be an adequate sole measure of the consciousness of a system. A second problem identified by Seth et al. is that the value of Φ depends on arbitrary measurement choices made by the observer. Different descriptions of the system lead to different predictions about its information integration, and Seth et al. demonstrate that a simple continuous system consisting of two coupled oscillators can generate arbitrary and even infinite values of Φ depending on the measurement units that are used. Both of these criticisms highlight the fact that Tononi and Sporns' (2003) method is at an early stage of development and needs further refinement to increase the accuracy of its predictions about real biological networks. Seth et al. also point out that Φ is essentially a static measure of consciousness, which makes it unable to distinguish between a conscious and an unconscious brain, and they discuss the difficulties of calculating the information integration of a realistic system.

Tononi and Sporns' (2003) Φ measure is based on their earlier work on neural complexity (Tononi et al. 1994, 1998). Neural complexity is defined as the average mutual information that is shared between a subset of the network and the rest of the system, where this average is taken over all subset sizes. Whilst Tononi and Sporns' (2003) method looks for the minimum information bipartition of the subset and introduces the concept of a complex, neural complexity is calculated once for the whole network without searching for the most integrated part. The computation cost of calculating neural complexity increases factorially in a similar way to effective information, but it can be approximated by limiting the analysis to bipartitions between a single element and the rest of the network (Seth et al. 2006). Since neural complexity

depends solely on mutual information it is only a measure of functional and not effective connectivity.[18]

Another way of measuring effective connectivity is the causal density measure put forward by Seth et. al. (2006), which identifies the causally significant interactions amongst a network's elements using Granger causality, and then calculates the causal density using Equation 7.16:

$$cd = \frac{\alpha}{n(n-1)}, \tag{7.16}$$

where *cd* is the causal density, α is the total number of significant causal interactions and *n(n - 1)* is the total number of directed edges in a fully connected network with *n* nodes.[19] Causal density depends on a comprehensive set of test data because it is calculated using the actual activity of the network, and it also has scaling problems since the multivariate regression models become difficult to estimate accurately as the number of variables increases. However, these scaling problems are substantially less serious than the factorial dependencies associated with neural complexity and Φ.

There has also been a substantial amount of analysis of the anatomical, functional and effective connectivity of biological networks, either using scanning or electrode data, or based on 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

[18] See Sporns et al. (2004) for the difference between anatomical, functional and effective connectivity.

[19] Granger causality has also been used by Seth and Edelman (2007) to identify causal cores within a large network.

task, and Friston et al. (2003), who modelled the interactions between different brain areas and made predictions about the coupling between them. There is also the work by Massimini et. al. (2005) who measured the cortical effective connectivity during non-REM sleep and waking. An overview of this type of research can be found in Sporns et. al. (2004) and Sporns (2007).

7.4.8 Information Integration: Discussion and Future Work

The seed expansion method was found to be an effective way of speeding up the calculations and offered a valuable way of controlling the analysis time by limiting the maximum subset size. However, this method did have the problem that errors introduced by other approximations could lead to erroneous expansions of the subset, and it is also probable that the order of expansion of the connected neurons significantly altered the final complex. Future work in this area could evaluate the effect of different expansion orders on the complexes found in the network.

One possible improvement to this analysis would be to use a shuffling algorithm to randomly select different neurons from homogenous connections, in order to identify complexes with similar Φ and connection patterns. For example, the high information integration of the main complex partly depends on connections to Vision Input that are selected from a large uniform set, and a different selection of these connections could be used to identify a different complex with similar Φ.

In this analysis, the main compromise between speed and accuracy was the limitation on the number of calculations per bipartition, which had a big effect on the calculation time (see Figure 7.4 and Table A3.12 in Appendix 3), and a proportionally greater impact on larger networks. On most calculations this approximation would have made the final Φ higher than it actually was by reducing the number of bipartitions that were examined for the minimum normalized effective information. However, in some circumstances this approximation might have artificially reduced the Φ by changing the way in which the subset expanded.

Although the equal bipartition approximation speeded up the analysis considerably, the results in Figure 7.6 show that a significant number of other bipartitions had the minimum normalised effective information. When this approximation was combined with the seed method it led to substantially different complexes, and so it was not used in the final analysis. In future work it would be worth investigating the strengths and limitations of this approximation in more detail and it might be possible to use the structure of the network to decide when the equal bipartition approximation is most likely to be accurate.

The main limitation of this analysis was the extremely long time that was required to calculate Φ. One way of addressing this problem would be to use graphics cards for the matrix calculations - for example, using the NVIDIA CUDA system.[20] Although this analysis did run partly in parallel by expanding the seeds from different neuron groups on different computers, the code could be rewritten to automatically distribute itself across an arbitrary number of processors. This would enable it to run on supercomputers and address some of the memory limitations that were encountered with large neuron groups.[21]

Work is already in progress on the simulation of networks with a billion spiking neurons (see Section 5.6) and on networks of this size even supercomputing power will not be enough to identify the complexes of the network. Future work should investigate other methods of estimating the effective connectivity of neural networks, such as Seth et. al.'s (2006) causal density measure, and it would also be worth investigating whether Φ can be estimated on the basis of sub-samples of each bipartition.

A further limitation of Tononi and Sporns' (2003) method is that it is essentially static and ignores the fact that complexes in a real network might change over time. In future work, it would be much better to record the network as it interacts with the world and use transfer

[20] NVIDIA CUDA: http://www.nvidia.com/object/cuda_home.html.

[21] For example, the Φ of Vision Input could not be calculated because it used more than 2GB of RAM, which was the maximum that could be installed on the computers used for this analysis.

entropy (Schreiber 2000) or a similar method to identify the effective information that is integrated across different bipartitions of each subset. It would also be worth analyzing the system at a number of different levels - for example, using populations of neurons, ion channels and memory addresses as well as neurons – to increase our understanding of the difference between simulated and physical systems.

In the next three sections definitions based on Tononi's, Aleksander's and Metzinger's theories of consciousness are developed, which are used make predictions about the phenomenology of the network in Section 7.9.

7.5 Phenomenal Predictions based on Tononi's Information Integration Theory of Consciousness

Tononi (2004) makes an explicit connection between the consciousness of a system and its capacity to integrate information: "consciousness corresponds to the capacity to integrate information. This capacity, corresponding to the quantity of consciousness, is given by the Φ value of a complex." This link between Φ and consciousness is independent of the material that the system is made from, but there is not a simple proportional relationship between Φ and consciousness because only the main complex is capable of consciousness according to Tononi's theory – parts of the system that are outside the main complex are completely unconscious.

When complexes overlap it seems reasonable to follow Tononi (2004) and only allocate consciousness to the one with the highest Φ.[22] However, when complexes do not overlap and exchange relatively little information, it seems more sensible to attribute two consciousnesses to the system, rather than saying rather arbitrarily that the one with slightly higher Φ is conscious and the other not conscious at all. To accommodate this type of case without including all of the independent complexes of the system, this analysis will consider a firing neuron to be conscious

[22] The problems with this are discussed in Section 7.9.4.

according to Tononi's theory if it is part of the main complex or if it is part of an independent complex whose Φ value is at least 50% that of the main complex. The explicit definition is as follows:

> *A mental state will be judged to be included in the phenomenally conscious part of the system according to Tononi 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.* (7.2)

The results from the information integration analysis showed that the main complex was the only independent complex, and so Tononi's theory predicts that the 91 neurons in the main complex will be the only parts of the network that are associated with conscious states. Tononi (2004) claims that the amount or quantity of consciousness in the conscious part of the network is given by the Φ value of the main complex, which is 103.

7.6 Phenomenal Predictions based on Aleksander's Axioms

7.6.1 Is the System Synthetically Phenomenological?

In earlier work, Aleksander and Dunmall (2003) set out five axiomatic mechanisms and claimed that these are minimally necessary for consciousness (see Section 2.6.3). Objects that did not possess these mechanisms were not considered to be conscious according to this theory. Over the last few years Aleksander's thinking has evolved and he now emphasises the importance of depiction over the other axioms, as illustrated in the following quotation:

> **Def 1:** To be **synthetically phenomenological**, a system S must contain machinery that represents what the world and the system S within it *seem* like, from the point of view of S. …
>
> **Def 2:** A **depiction** is a state in system S that represents, as accurately as required by the purposes of S the world, from a virtual point of view within S.

> **Assertion 1:** A depiction of Def. 2 is the mechanism that is necessary to satisfy that a system be synthetically phenomenological according to Def. 1.
>
> Aleksander and Morton (2007a, p. 72)

This section will take a brief look at whether the neural network developed in this thesis conforms to all five of Aleksander's axioms, but it will only consider the network to be capable of consciousness (or synthetically phenomenological) if it includes depiction.

1. Depiction

Although the network described in this paper does not have gaze locked cells, the neurons in Red Sensorimotor and Blue Sensorimotor are connected to both Vision Input and Motor Integration, and respond to both visual data and the motor signals sent to control the eye, which contain proprioceptive information. These observations are confirmed by the measurements of representational mental states in Section 7.3, which showed that neurons in Red Sensorimotor and Blue Sensorimotor share mutual information with Vision Input and Motor Integration. It also appears to be consistent with the interpretation of depiction in this thesis that it could be implemented as a population code in which the *combined* activity of the motor and visual layers represents the presence of an out there world. In this case some kind of binding or integration between the motor and visual layers would be all that was needed for depiction.

2. Imagination

The network has an offline mode in which it can 'imagine' the consequences of different motor actions without carrying them out.

3. Attention

This network's 'imagination' is used to select the part of the world that is looked at by the system.

4. Volition

When Vision Input and Motor Output are inhibited, the 'imagination' circuit decides which part of the world to look at and then executes the selected motor action based on the response of its 'emotion' layer.

5. Emotion

The neural network has an 'emotion' layer, which responds 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 (Arbib and Fellous 2004). However, it could be argued that this 'emotion' layer does not directly represent the state of SIMNOS's body, and so it is at best something like the 'as if' circuit discussed by Damasio (1995). Other limitations of the 'emotional' response are that it does not modulate the way in which neurons and synapses compute and it lacks the detail that we sense when our viscera and skeletal muscles are changed by an emotional state, such as fear or love (Damasio 1995, p. 138). These limitations do not completely exclude the possibility that the 'emotional' response of the network can be counted as an emotion, and so it will be provisionally accepted as a very primitive emotion that is much simpler than our basic human emotions.

This discussion suggests that the neural network in this thesis is capable of depiction and minimally conforms to Aleksander's other axioms, and so it is likely to possess a very simple form of consciousness according to this theory. Since the network is simulated and operates very differently from a real biological network on a much smaller scale, the contents and qualitative character of this consciousness will be very different from the consciousness of biological creatures that have the axiomatic mechanisms.[23]

[23] These differences are likely to be much greater than those identified by Nagel (1974) between human and bat consciousness.

7.6.2 What are Aleksander's Predictions about Phenomenal States at Time *t* ?

In this analysis, predictions about phenomenal states according to Aleksander's theory are based on his link between depiction and consciousness. The depictive neurons are identified by using the method set out in Section 7.3 to look for representational relationships between input/ output neurons and internal states of the system. Under these experimental conditions, high mutual information between an input/ output and internal neuron indicates a strong representational relationship, and so an internal neuron that shares a high level of mutual information with both visual and proprioceptive data is likely to be depictive. Since depictive neurons are defined by the fact that they respond to both sensory and proprioceptive data, the amount of depiction will be limited by whichever of these is smallest. This leads to the following definition:[24]

> *A mental state will be judged to be within the phenomenally conscious part of the system according to Aleksander if it shares mutual information with both sensory and proprioceptive layers. The* ***amount*** *of consciousness will be measured by the minimum mutual information that is shared with sensory and proprioceptive layers. So, for example, if the neuron has 0.4 mutual information with an auditory input layer, 0.2 mutual information with a visual input layer and 1.0 mutual information with a proprioception layer, then its amount of consciousness would be judged to be min{0.4, 0.2, 1.0} = 0.2, according to Aleksander's theory.* (7.3)

Based on this definition, the only parts of the network that share mutual information with both visual input and proprioception/ motor output are Red Sensorimotor and Blue

[24] It might be thought that the sensory and motor mutual information values could be added or multiplied together to get the amount of depiction. However, consider two neurons: neuron A that has 1000 mutual information with visual input and 0.1 mutual information with motor output, and neuron B that has 10 mutual information with visual input and 10 mutual information with motor output. A's strong response to visual information makes it much more like a photographic representation, whereas neuron B is much closer to the gaze-locked neurons discovered by Galletti and Battaglini (1989) that respond to a particular combination of sensory and muscle information, and are cited by Aleksander (2005) as a key example of depictive neurons. In this example, addition of the mutual information values gives 1000.1, for neuron A and 20 for neuron B, which erroneously suggests that neuron A is more depictive than neuron B. The product of the mutual information values gives 100 for neuron A and 100 for neuron B, which is also an incorrect measure of their relative levels of depiction. In this example, the minimum of the two values, which is 0.1 for neuron A and 10 for neuron B most accurately predicts which neuron is most depictive.

Sensorimotor, and so activity in these parts of the network will be conscious according to Aleksander's theory. For the phenomenal predictions in Section 7.9 the mutual information values were normalized to the range 0-1, and so the maximum amount of consciousness is 1. In the future, methods such as transfer entropy (Schreiber 2000, Sporns and Lungarella, 2006), backtracing (Krichmar et. al. 2005) and Granger causality (Seth and Edelman 2007) could be used to identify the depictive parts of the network.

7.7 Phenomenal Predictions based on Metzinger's Constraints

7.7.1 Is Artificial Subjectivity Possible?

Although Metzinger (2003) believes that machines are capable of consciousness, he points out that our current simulations and robotic models are too coarse to replicate the extremely fine levels of detail of biological systems:

> The subtlety of bodily and emotional selfhood, the qualitative wealth and dynamic elegance of the *human* variety of having a conscious self, will not be available to any machine for a long time. The reason is that the microfunctional structure of our emotional self model simply is much too fine-grained, and possibly even mathematically intractable. … Self-models emerge from elementary forms of bioregulation, from complex chemical and immunological loops—and this is something machines don't possess.
>
> Metzinger (2003, p. 619)

One way of developing machines with a fine-grained biological structure is to use biological neurons to control a real or virtual robotic body, as was done in the work of DeMarse et al. (2001). Metzinger also points out that consciousness is a *graded* phenomena and that there are degrees of constraint satisfaction and phenomenality: "just as with animals and many primitive organisms surrounding us on this planet, it is rather likely that there will soon be artificial or postbiotic systems possessing simple self-models and weaker forms of conscious experience in

our environment." (Metzinger 2000, p. 620). If consciousness is graded, then systems as simple as the neural network developed by this project may be capable of an extremely limited form of consciousness if they can satisfy Metzinger's minimal set of constraints.

7.7.2 Does the Network Conform to Metzinger's Constraints?

Although Metzinger describes his constraints on conscious experience at a number of different levels, these descriptions remain at a fairly high level of abstraction and in some cases it is quite difficult to say whether the network developed by this thesis matches them or not. This section is a general discussion about the degree to which the network conforms to the constraints; a more precise definition of what it would mean for the network to conform to Metzinger's minimal definition of consciousness is given in the next section.

1. Global availability

The network can access information in different parts of its real and imaginary environment and this information is available for the control of action, and so the network does possess a limited form of global availability. Metzinger links global availability with Tononi's earlier work on information integration, and so it might be possible to use Φ to measure this constraint.

2. Window of presence

Activity within the network does exist in a single now and there is a certain amount of temporal integration along the connections with different delays. The reverberatory activity within Emotion, Inhibition and Motor Cortex also stores a limited amount of information about earlier states of the system. Taken together these observations suggest that the window of presence of the network is very thin, but not completely non-existent.

3. Integration into a coherent global state

Global availability (constraint 1) is a functionalist-level description of global integration, which Metzinger links to Tononi's earlier work on information integration. This suggests that Φ could be used to measure the degree to which the network integrates information into a coherent global state.[25]

4. Convolved holism

The visual processing of the neural network is too basic to identify wholes at different levels of scale, and so it does not even minimally conform to this constraint. In the future, more complex processing could be added to the network to enable it to identify part-whole relationships.

5. Dynamicity

The network can sustain the activation of a neuron group over time, but it has a very limited ability to integrate information between points in time and it is not sensitive to the part-whole structure of temporal information.

6. Perspectivalness

This constraint has a certain amount of overlap with Aleksander's depiction axiom and the network's integration between sensory and proprioceptive information should give it some kind of rudimentary sense of seeing the world from somewhere. Since the size of objects changes with distance and the network only perceives part of the world at any one time, there is also some sense to the idea that it has a perspective.

[25] Metzinger (2003) was published in the same year as Tononi and Sporns (2003), and so it is unlikely that Metzinger (2003) knew about Tononi's work on Φ. Metzinger's more recent work, such as Metzinger and Windt (2007) and Metzinger (2008), has focused on the phenomenal self model and the phenomenal model of the intentional relation.

7. Transparency

Since the network lacks internal sensors there is some basis to the claim that it is as transparent as a biological neural network, such as the brain. However, Metzinger distinguishes between conscious and unconscious transparency and claims that almost nothing is known about the neural basis of phenomenal transparency. This suggests that we have no reason to believe that the neural network is *less* transparent than the human brain, but much more research needs to be done on transparency.

8. Offline activation

This constraint is similar to Aleksander's second axiom of imagination and the system is capable of inhibiting its sensory input and motor output whilst it 'imagines' an eye movement that would look at a red or blue object.

9. Representation of intensities

Information in the network is held as neurons that spike at different rates, and so this constraint is implemented by the system.

10. "Ultrasmoothness": the homogeneity of simple content.

Although individual neurons represent individual areas of colour, there is no representation within the system of the gaps between neurons, and so the network cannot access the graininess of the neurons' spatial firing patterns that is visible to us as outside observers. The network is also unable to represent the graininess of its temporal representations, and so it is probably reasonable to claim that its mental states are ultrasmooth.

11. Adaptivity

This network did not come about through natural selection, and so it does not conform to this constraint.

7.7.3 What are Metzinger's Predictions about Phenomenal States at Time *t* ?

The discussion in the previous section demonstrated that the network is likely to conform to a number of Metzinger's constraints, including a coherent global model of reality (*constraint 3*), a window of presence (*constraint 2*) and transparency (*constraint 7*), which are sufficient for Metzinger's minimal notion of consciousness. In this analysis, the degree to which a mental state is involved in a coherent global model of reality will be indicated by the Φ value of the highest Φ complex that it is involved in. Since recurrency is a key way in which information can be integrated over time, a window of presence will be attributed to neurons whose highest Φ complex includes a recurrent part of the system. Transparency will be left out of this analysis because it cannot be directly identified, and it has been argued that we do not have any reason for believing that the network is less transparent than the human brain. The final definition is as follows:

> *A mental state will be judged to be minimally conscious according to Metzinger if the highest Φ complex that it is involved in includes one or more recurrent layers. The* ***amount*** *of consciousness will be indicated by the Φ of this complex.* (7.4)

According to this definition, the conscious parts of the network will be the complexes that include Motor Cortex, Emotion and Inhibition. The amount of consciousness will be the Φ of these complexes.

7.8 Other Phenomenal Predictions

For the reasons discussed in Section 2.6.1, only three theories of consciousness are being used to make predictions about the consciousness of the network in this thesis. However, to provide more context for this work I will make brief remarks about some other theories that make fairly

explicit predictions about the consciousness of the network. None of these predictions were used to generate the final XML description in Section 7.9.

Pantheism

Pantheists, such as Spinoza (1992), believe that all matter is conscious to some degree, and so the physical computer running the simulation is conscious even when it is switched off. From this perspective, the task of synthetic phenomenology is to determine the amount of consciousness in the system and the qualitative character of this consciousness at different points in time. Pantheism is a type I theory because the behaviour of the system does not affect the attribution of consciousness to it.

Information states

Chalmers (1996, p. 292) claims that conscious experiences are realizations of information states, and so systems as simple as thermostats are conscious because they contain information. Since the neural network contains a large number of information states, it is conscious according to this hypothesis. This link between consciousness and information states is a type I theory because every object in the universe interacts to some degree and stores 'information' about the particles and forces affecting it.

Non-biological systems cannot be conscious

A number of people would argue that the neural network developed by this project can never become conscious because it is a simulated artificial system (Searle 2002) or because the calculations that are used to simulate it are all algorithmic (Penrose 1990, 1995). These theories are discussed in detail in Section 3.4.

Internal models

Holland (2007) claims that internal models play an important role in our conscious cognitive states and may be a cause or correlate of consciousness in humans (see Section 3.5.2). In this network, the activity in Motor Cortex, Motor Integration, Eye Pan, Eye Tilt and Motor Output accurately reflects the position of SIMNOS's eye 'muscles', and this could be interpreted as an internal model in an extremely limited sense. This internal modelling could be made more realistic by making the activity in Emotion reflect the internal states of SIMNOS's body, which would also link the network more closely to Damasio's (1995) work.

7.9 XML Description of the Phenomenology of the Network

7.9.1 Introduction

This section explains how the data about representational mental states and complexes was integrated with definitions 7.2 - 7.4 to generate a sequence of XML files that predicts the phenomenology of the network at each time step. The first parts of this procedure were two recordings of the network, which are documented in Section 7.9.2. The next section explains how the XML files were generated, and then sections 7.9.4 – 7.9.6 examine the predictions that were made about the consciousness of the network using Tononi's, Aleksander's and Metzinger's theories of consciousness. After discussing what these results show about the relationship between consciousness and action, some extensions and enhancements of the consciousness of the network are suggested in Section 7.9.8, and the analysis concludes with a discussion and suggestions for future work.

7.9.2 Analysis Data

The main data for this analysis was recorded as the neural network moved SIMNOS's eye and used its 'imagination' to avoid looking at the blue cube, as described in Section 5.5.1. The

recording starts at time step 13100 with the network in its online perception mode and an empty visual field. At time step 13138 a red object starts to appear in the top left corner of the visual field and this moves in and out of view until time step 13503, when a blue object appears in the bottom left corner of the visual field. This leads to the activation of Inhibition after time step 13520 and the system switches into its offline 'imagination' mode. At time step 13745 the system 'imagines' a blue blob in the left half of its visual field and eventually it 'imagines' a red object at time step 13945, which activates Emotion and returns the system to online perception. Finally at time step 13966 the network starts to perceive a red object in the top left corner of its visual field. This recording of data from time steps 13100 to 14004 will be referred to as "Analysis Run 1", and a video of Analysis Run 1 is included in the supporting materials.

The average number of times that each neuron fired during Analysis Run 1 was recorded and the results were normalized to the range 0-1 and used to illustrate the activity of the network in Figure 7.8. This shows that Inhibition was the most active part of the network, followed by Emotion. Traces of motor and visual activity can also be seen in Figure 7.8.

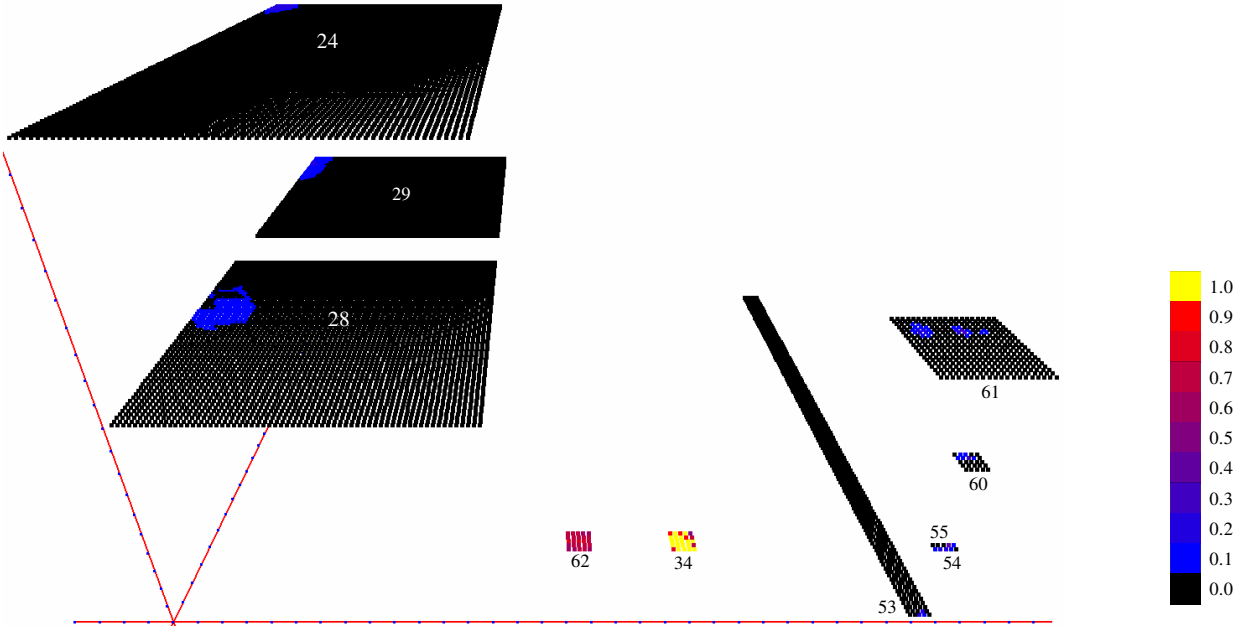

Figure 7.8. Normalized average firing frequency of neurons during Analysis Run 1

A recording was also made in which the neuron groups were disconnected from each other and themselves and 5% noise was injected into each layer at each time step for 100 time steps. The normalized average firing frequency of each neuron was used to illustrate the activity of the network in Figure 7.9, which shows that there was a reasonably even spread of activity across the layers. This noise recording will be referred to as "Noise Run 1".

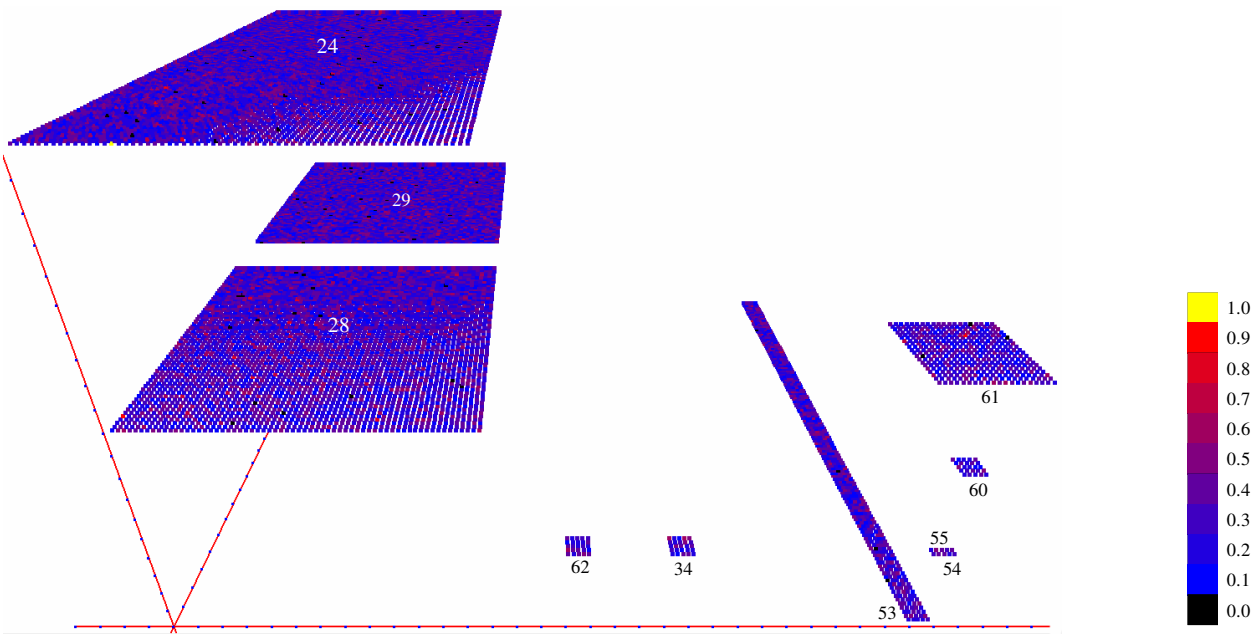

Figure 7.9. Normalized average firing frequency of neurons during Noise Run 1

The data from Analysis Run 1 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 was active, and so it does not tell us about the consciousness that might be predicted to be associated with the other parts of the network. On the other hand, the noise data has an even spread of activity that includes all of the neurons, but it was recorded with the layers disconnected from themselves and each other, and so the predictions about the consciousness of the network during Noise Run 1 are 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 the different parts of the network.

7.9.3 Generation of the XML Description

To generate the XML files, the recordings of the network's activity were combined with the OMC rating, representational mental states, complexes, clusters and definitions to produce a sequence of XML files describing the phenomenology of the system at each time step. As discussed in Section 7.3.1, firing neurons are being treated as mental states for this analysis and the predictions about the consciousness associated with each mental state are given by definitions 7.2 – 7.4. It was decided not to normalize the predictions based on Tononi's and Metzinger's theories of consciousness, both because Φ does not have a maximum value and because Tononi interprets Φ as an absolute measure of a system's consciousness. The predictions based on Aleksander's theory were normalized to the range 0-1 by dividing the mutual information by the maximum possible mutual information of 0.72.[26]

In addition to the representational mental states identified in sections 7.3.4 and 7.3.5, the neurons in the input and output layers were also treated as representational mental states in the final XML description and assigned a mutual information value of 1 to reflect the fact that they shared the maximum amount of mutual information with themselves. The other mutual information values for the representational mental states were normalized by the maximum possible mutual information. In the integration part of the description, neurons that were not included in any complex were assigned a Φ value of zero.

In order to compare the different theories' predictions about the distribution of consciousness associated with the network, the amount of predicted consciousness per neuron was averaged over Analysis Run 1 and Noise Run 1, normalized to the range 0-1 and used to highlight the network in figures 7.10 - 7.15. I have only shown the *relative* distribution of consciousness in the network because the assignment of absolute values to predicted

[26] See Section 7.3.2 for the calculation of this value. In practice the normalized values occasionally strayed over 1.0 due to noise in the data.

consciousness is largely meaningless without some form of *calibration* on humans – a problem that is discussed in Section 7.9.9.

7.9.4 Predictions about the Consciousness of the Network According to Tononi's Theory

Tononi's theory predicts that the main complex is the only conscious part of a system and that the amount of consciousness in the main complex can be measured by its Φ value. In this network the main complex has a Φ of 103 and it includes all of the neurons highlighted in Figure 7.7. The predicted consciousness of the network at each point in time is therefore the intersection of the neuron activity with the main complex. In Noise Run 1 there is fairly uniform activity across the network, and so the distribution of consciousness for Noise Run 1 is an extract from the average activity shown in Figure 7.9 that is shaped like the main complex (see Figure 7.10).

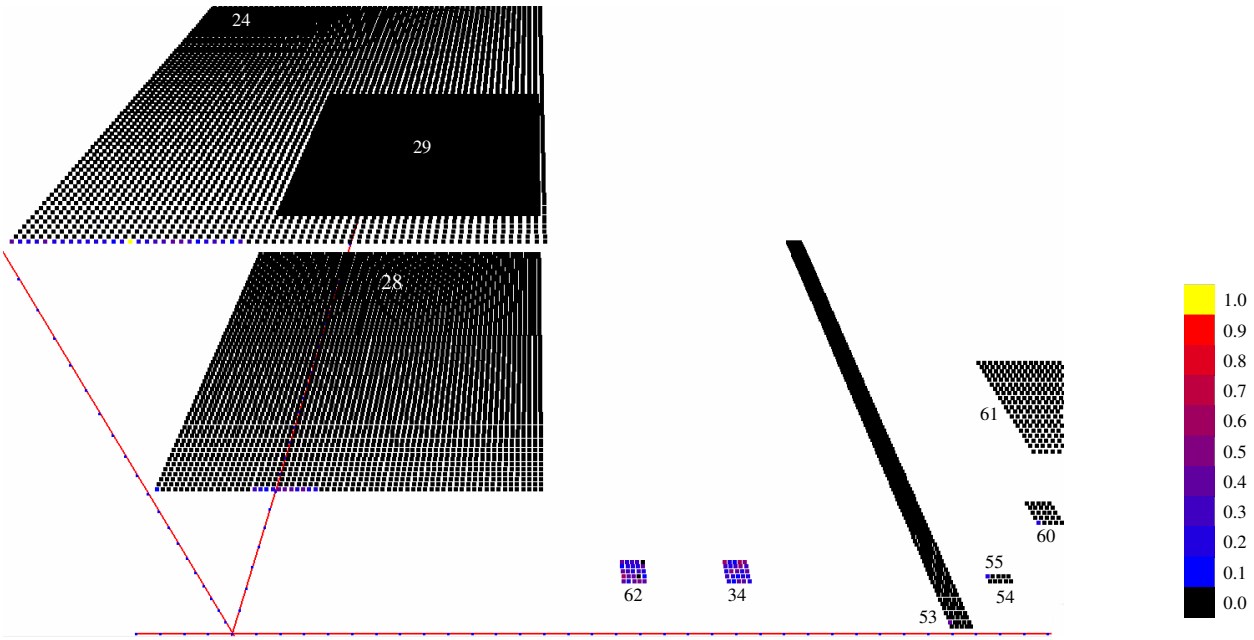

Figure 7.10. Predicted distribution of consciousness during Noise Run 1 according to Tononi's theory

The more specific neuron activity during Analysis Run 1 did not include any of the main complex neurons outside of Emotion and Inhibition, and so the predicted distribution of consciousness in Figure 7.11 only includes neurons from Emotion and Inhibition, with the

pattern closely matching the average firing frequencies shown in Figure 7.8. The network would not have been conscious *of* anything during Analysis Run 1 because none of the conscious mental states were representational.[27]

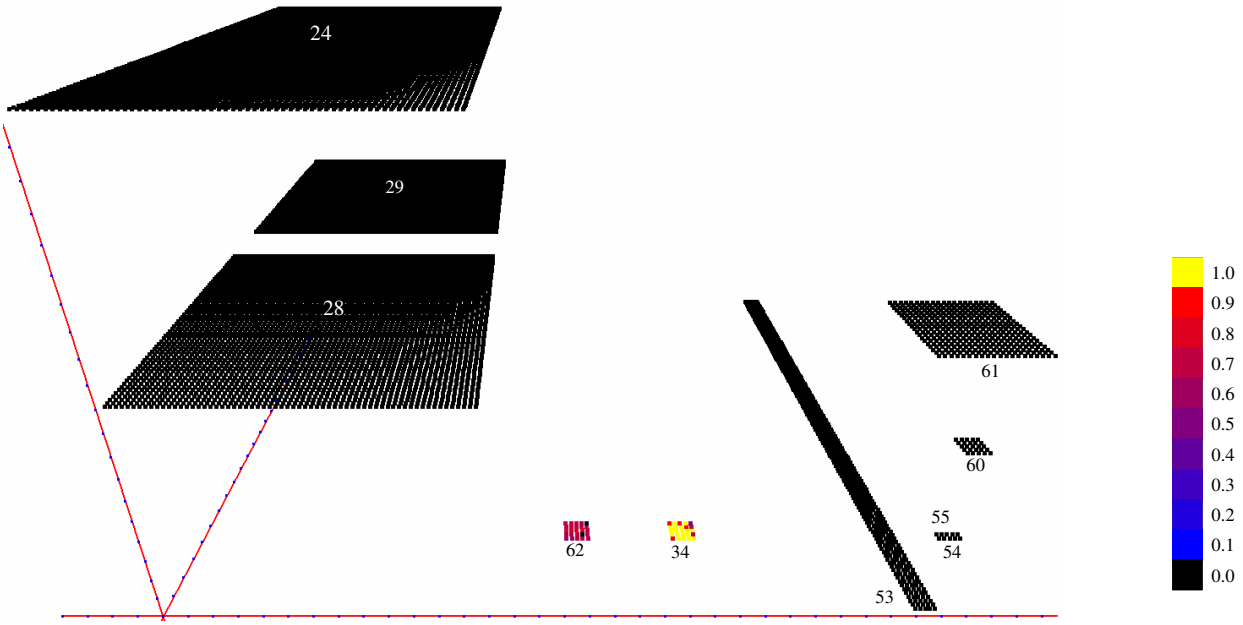

Figure 7.11. Predicted distribution of consciousness during Analysis Run 1 according to Tononi's theory

These results highlight a major problem with a simplistic link between the main complex and consciousness. This network has a number of overlapping complexes with approximately the same value of Φ and it seems somewhat arbitrary to interpret just one of these as the main complex, when it is 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 the different neurons in Vision Input. This appears to reflect our own phenomenology since we seem 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

[27] Tononi's (2004) suggestion that the qualitative character of mental states is determined by their informational relationships might lead to different predictions about what the network was conscious of during Analysis Run 1.

overlapping complexes: a high rate of change of Φ could be used indicate a boundary between the conscious and unconscious parts of the system.

7.9.5 Predictions about the Consciousness of the Network According to Aleksander's Theory

Aleksander's emphasis on depiction led to a prediction about phenomenal states that was based on the minimum amount of mutual information shared with both sensory input and proprioception/ motor output. In this network only Red Sensorimotor and Blue Sensorimotor share mutual information with both Vision Input and Motor Integration, and so these were the only layers that were capable of consciousness according to Aleksander's theory. Whilst there are homogenous connections between Vision Input and Red/ Blue Sensorimotor, the connections between Motor Integration and Red/ Blue Sensorimotor reflect the learnt associations between motor output and visual input, which are stronger whenever motor output consistently resulted in red or blue visual input. This variation in connection strength affects the mutual information between Motor Integration and Red/ Blue Sensorimotor, producing a pattern in the predicted distribution of consciousness for Noise Run 1, which is shown in Figure 7.12.

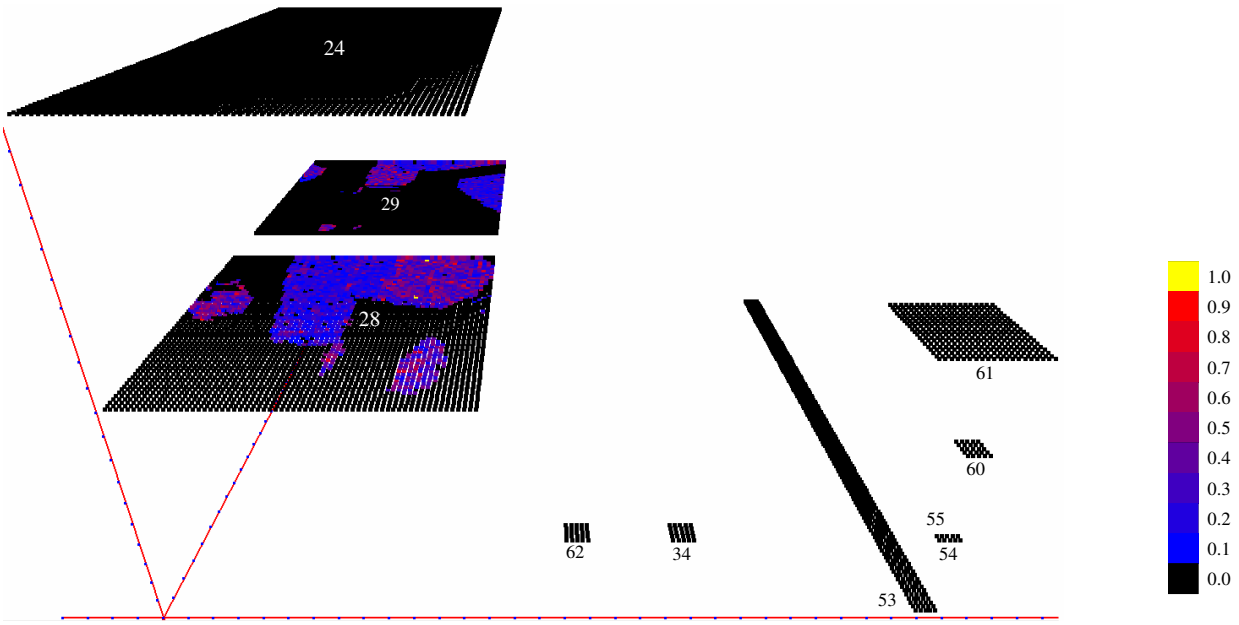

Figure 7.12. Predicted distribution of consciousness during Noise Run 1 according to Aleksander's theory

The predicted distribution of consciousness for Analysis Run 1 reflects the fact that visual activity was concentrated in the top left corner of the red visual field with the occasional 'imagined' blue image (see Figure 7.13). According to Aleksander's definition of depiction, the red and blue data that is represented by these conscious mental states would have been experienced by the system as part of an out there world.

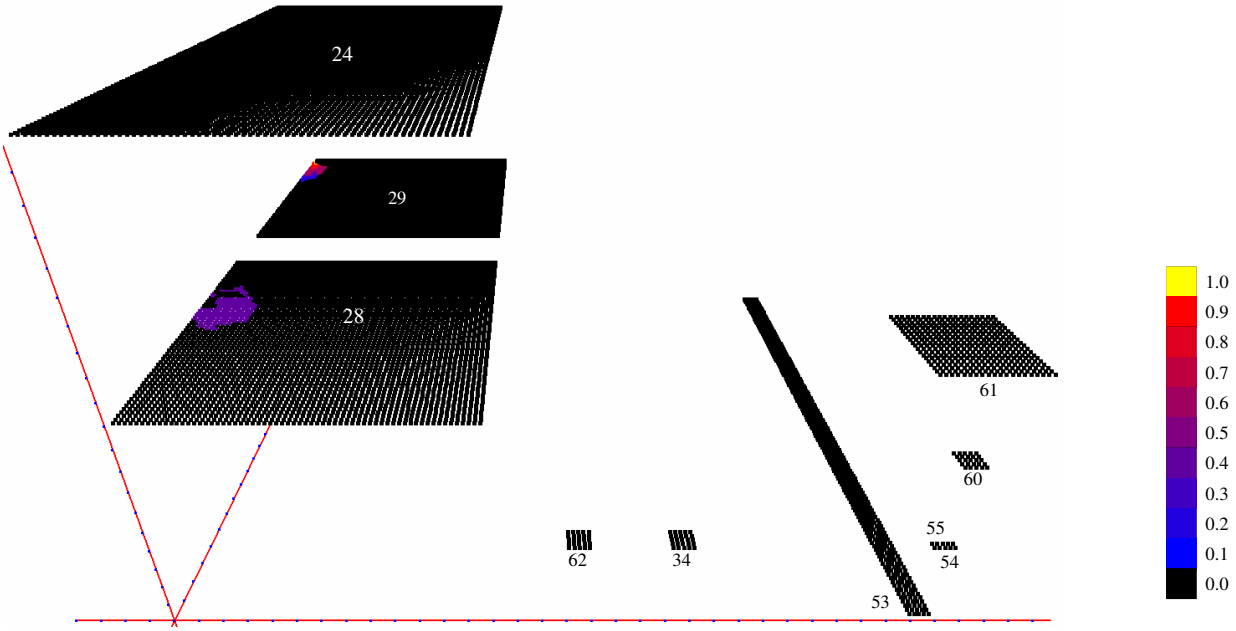

Figure 7.13. Predicted distribution of consciousness during Analysis Run 1 according to Aleksander's theory

7.9.6 Predictions about the Consciousness of the Network According to Metzinger's Theory

Predictions about consciousness based on Metzinger's theory used a combination of spatial and temporal integration, with the former measured using Φ and the latter marked by the presence of a recurrent neuron group in the highest Φ complex. It turned out that almost all of the neurons' highest Φ complexes included one of the three recurrent layers (Motor Cortex, Emotion and Inhibition), and so almost all of the network was predicted to be minimally conscious according to Metzinger. This is shown in the predicted distribution of consciousness for Noise Run 1 (Figure 7.14) and Analysis Run 1 (Figure 7.15), which closely match the distribution of firing frequencies depicted in Figure 7.9 and Figure 7.8.

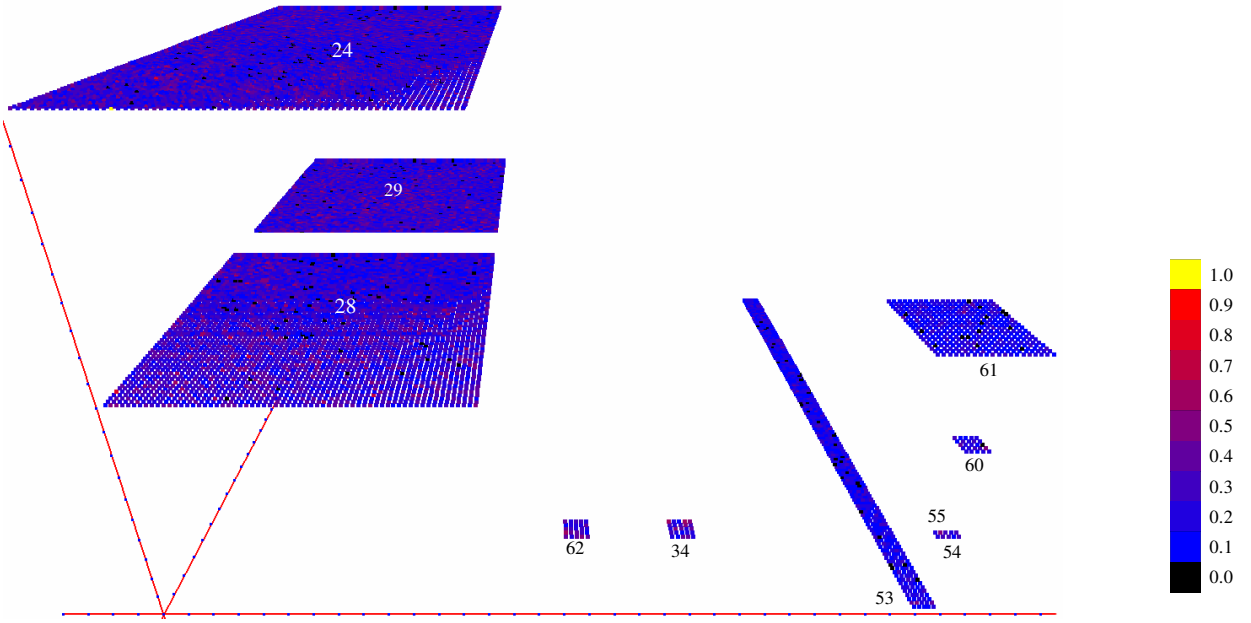

Figure 7.14. Predicted distribution of consciousness during Noise Run 1 according to Metzinger's theory

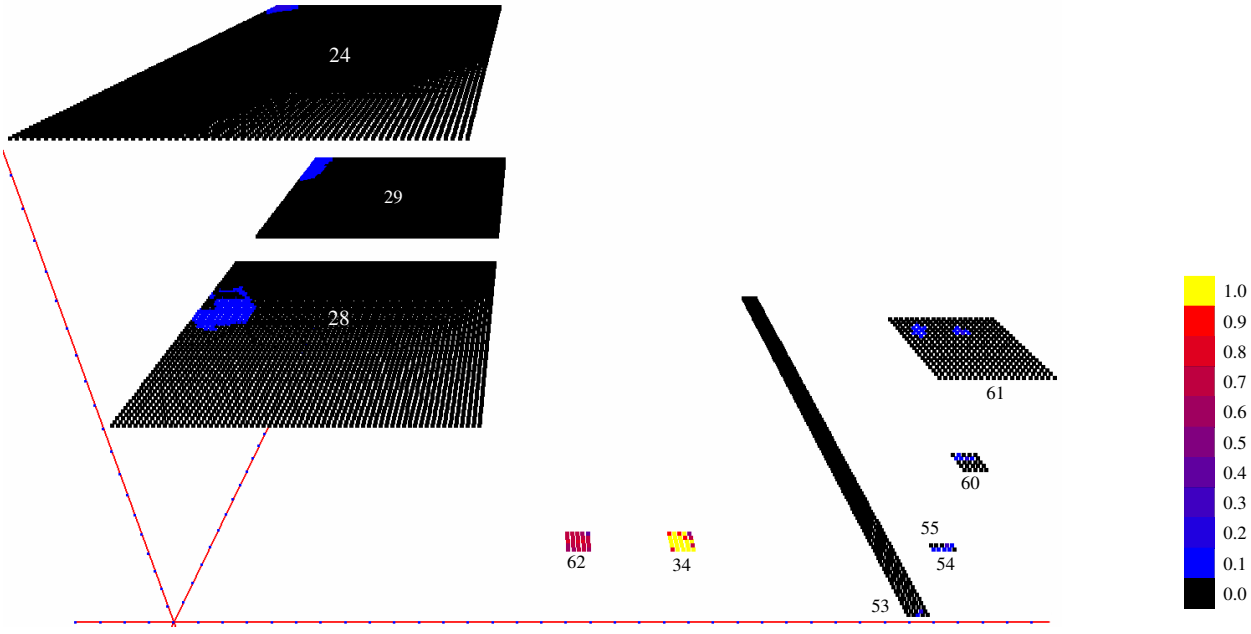

Figure 7.15. Predicted distribution of consciousness during Analysis Run 1 according to Metzinger's theory

During Analysis Run 1 the network would have been conscious of all the active visual and proprioception/ motor output information. This prediction of uniform potential for consciousness throughout the network is likely to change if more of Metzinger's constraints were taken into account. For example, if the mental states associated with consciousness had to be

capable of offline activation (constraint 8), then the neurons in Vision Input and Motor Output would no longer be predicted to be associated with consciousness.

7.9.7 Predictions about Conscious and Unconscious Action

This section looks at how the predictions made about the consciousness of the network stand in relation to the discussion of consciousness and action in Section 2.7. As discussed in Section 5.7, the absence of a reactive layer in the network makes it incapable of conscious will, and this discussion focuses on whether it is capable of discrete conscious control according to the different theories of consciousness.

Tononi

The main complex includes only a small number of neurons from Vision Input, Blue Sensorimotor, Motor Integration, Eye Pan, Eye Tilt and Motor Output, and all of these were predicted to be unconscious during Analysis Run 1. However, under very specific conditions it is possible that these sensory and motor parts of the main complex could become active and 'imagine' an action prior to carrying it out, but this is unlikely to happen during normal operation, and most of the time it will be the unconscious parts of the network that decide an action, initiate it and unconsciously carry it out.

Aleksander

Aleksander's theory predicts that there will not be any conscious activity in Vision Input, Motor Integration, Eye Pan, Eye Tilt, Motor Output, Inhibition or Emotion during Analysis Run 1. Whilst the network might be experiencing red and blue in an out there world, the conscious parts do not have any way of differentiating between real and imagined visual input, and so the system cannot tell whether it is deciding to perform an action or actually carrying it out. If the network cannot consciously differentiate between planning and execution, then it cannot be said to be

making a conscious decision - it may be conscious of parts of the planning process, but it is not conscious *that* it is planning and it is unable to remember whether it is planning or executing an action. So this interpretation of Aleksander's theory predicts that the network unconsciously chooses an action, unconsciously initiates it and then consciously carries it out.

Metzinger

According to this interpretation of Metzinger's theory, the network is conscious of its planned motor actions and their 'imagined' sensory consequences, and when an action is chosen and initiated, the system becomes conscious of the actual sensory consequences. This suggests that the network is capable of discrete conscious control, in which it consciously plans actions that are initiated immediately and consciously carried out.

7.9.8 Extensions and Enhancements to the Predicted Consciousness of the Network

These predictions about the consciousness of the network suggest a number of ways in which it could be extended or enhanced.

Tononi

Before any thought can be given to extending the consciousness that was predicted to be associated with the network according to Tononi's theory, it is essential to get a more plausible picture of its consciousness by improving the way that consciousness is analyzed to take account of overlapping complexes in a more flexible way (see Section 7.9.4). Once this has been done, it might be possible to design a network in which the main complex has enough representational mental states for conscious decision making. The network's consciousness could also be increased by evolving connection patterns that give the main complex a higher value of Φ.

Aleksander

The mutual information between Vision Input and Red/ Blue Sensorimotor cannot be increased because it is already at its theoretical maximum, but it might be possible to increase the mutual information between Motor Integration and Red/ Blue Sensorimotor by fine tuning the training. The main direction of improvement for this network would be to extend the range of consciousness by making more parts depictive. For example, Emotion and Inhibition could become depictive if they were connected to proprioceptive data and internal sensory data from virtual organs in SIMNOS's body. Red/ Blue Sensorimotor could then change the state of the virtual organs, and when the system sensed this change it would become conscious of the difference between 'positive' and 'negative' body states.

However, consciousness of 'positive' and 'negative' states would not be enough for the network to differentiate between imagination and online perception – it would be conscious of seeing red and feeling good or conscious of seeing blue and feeling bad, but it would not know if it was imagining or perceiving the red or blue stimuli.[28] One solution to this problem would be to use a remembered context or image intensity to indicate whether the network is imagining or not, and in Aleksander's kernel architecture (see Section 3.5.1), the memory module in the awareness area could perform this function by remembering which state is the real world.

Metzinger

The entire network was predicted to be minimally conscious according to Metzinger's theory, and so it would not be possible to extend this predicted consciousness. The qualitative characteristics of the consciousness in the network could be greatly improved by extending the

[28] This problem is closely related to Metzinger's discussion of the world zero hypothesis: "one of both world-models has to be defined as the *actual* one for the system. One of both simulations has to be represented as the *real* world, in a way that is functionally nontranscendable for the system itself. One of both models has to become indexed as the *reference model,* by being internally defined as real, that is, as *given* and not as constructed." (Metzinger 2003, p. 61).

visual processing, adding other senses, such as touch and audition, and increasing the complexity of the actions.

7.9.9 Phenomenal Predictions: Discussion and Future Work

The XML format that was used in these experiments is intended to be a simple example that illustrates the main ideas and a great deal more work is needed to turn this starting point into a usable method. As this approach develops there are likely to be a large number of changes and ambiguities, and although this might initially appear to be a weaknesses of the method, it is actually a strength because it indicates that synthetic phenomenology has the potential to become a paradigmatic science that can move forward by asking questions and resolving ambiguities. At the moment synthetic phenomenology is so unclear that even its lack of clarity is unclear to it, and this XML-based approach will enable synthetic phenomenology to ask and answer precise questions and move forward in a sustainable manner. As has been shown, different theories generate different predictions about the phenomenal states of a system and as brain scanning improves and robots become able to report their conscious states, we will be able to test these predictions and eliminate inaccurate theories.

This analysis presented the final results as the normalized average distribution of consciousness in the network during Noise Run 1 and Analysis Run 1. Whilst this did provide useful predictions about the consciousness of the network and suggestions for enhancing it, it did not address the question about how *much* consciousness was present. Ideally, this analysis would have stated that this network exhibited 5% of the consciousness of the average waking human brain, for example, but without *calibration* of the measurement scales it is impossible to say how much consciousness was associated with the system. 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 calculate or measure the Φ of the main complex in an average waking human brain, and

without this reference point, the Φ values quoted in this analysis are without absolute meaning. The values of mutual information that were used to measure depiction are equally problematic because we have no idea about how much mutual information is needed to make a mental state depictive.

In order to address this problem urgent work is needed to measure or estimate the Φ and mutual information of a waking human brain, in order to have some way of comparing the measurements of other systems with a system that can (at least to begin with) be taken as a reference standard of consciousness. 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 Φ and mutual information on more realistic simulations, such as the networks created by the Blue Brain project (Markram 2006). This would give some idea about the Φ and mutual information 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 this network. Better ways of quantifying the amount of consciousness in the 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, which we would unwilling to attribute much consciousness to.

In the future it might make sense to multiply the predicted levels of consciousness by the OMC rating to compensate for the type I differences between each system and the human brain. However, in this analysis it would have been pointless to multiply the uncalibrated predictions by a constant factor that would not have appeared in the relative distributions plotted in figures 7.10 - 7.15. Once calibration has been done on Φ and on the use of mutual information to measure depiction, it will be possible to use the OMC scale to compensate for the differences

between the system and the human brain, and say how much consciousness the network experienced during Analysis Run 1.

The sequence of XML files is a reasonably accurate description of the predicted phenomenology of the network that makes minimal assumptions about the nature of the phenomenal states. However, large XML files are almost impossible to read and digest and it is difficult to understand how the predicted consciousness of the system changes over time A logical extension of this work would be to investigate ways of presenting the content of these XML files in a more intuitive manner. If the system was experiencing a red spot in the left hand corner of its visual field, then it would be much easier to use virtual reality, for example, to show this to a human observer, instead of asking him or her to read an XML description. Such a 'debugger' for conscious states would also have applications in neurophenomenology.

Another direction of future work would be to move towards a common XML standard for neuro- and synthetic phenomenology that would facilitate collaboration between people working on machine consciousness and people from neuroscience and experimental psychology. This would enable phenomenal prediction methods that were developed in the biological sciences to be tested on artificial systems, and the methodology developed for synthetic phenomenology could be applied to fMRI data and used to make predictions about the consciousness of live human subjects.

Finally, in future work it would be worth making predictions about the consciousness of the network using other theories. For example, it would be particularly interesting to use some of the neural correlates of consciousness, such as neural synchronization (Crick 1994).

7.10 Conclusions

This chapter has demonstrated how the approach to synthetic phenomenology developed in Chapter 4 can be used to make predictions about the consciousness of an artificial neural

network. This analysis led to a number of suggestions about how the network's consciousness could be extended and enhanced and it showed how different theories of consciousness make different predictions about the relationship between consciousness and action. This work is at an extremely early stage and a great deal of research is needed to improve the accuracy of our predictions about phenomenal states. It is hoped that this will eventually lead to a more systematic science of consciousness that includes both natural and artificial systems within a single conceptual and experimental framework