Pergamon

Neural Networks, Vol. 10, No. 7, pp. 1227–1240, 1997

Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00049-X

1997 SPECIAL ISSUE

Consciousness in Neural Networks?

EDMUND T. ROLLS

Department of Experimental Psychology, University of Oxford

(*Received* 31 *August* 1996; *accepted* 15 *January* 1997)

Abstract—*A combined neurophysiological and computational approach is reviewed that leads to a proposal for how neural networks in the temporal cortical visual areas of primates could function to produce invariant object representation and identification. A similar approach is then reviewed which leads to a theory of how the hippocampus could rapidly store memories, especially episodic memories including spatial context, and how later recall of the information to the neocortex could occur. Third, it is argued that the visual and memory mechanisms described could operate without consciousness, and that a different type of processing is related to consciousness. It is suggested that the type of processing related to consciousness involves higher-order thoughts ("thoughts about thoughts"), and evolved to allow plans, formulated in a language, with many steps, to be corrected. It is suggested that it would feel like something to be a system that can think linguistically (using syntax) about its own thoughts, and that the subjective or phenomenal aspects of consciousness arise in this way. It is further suggested that "raw sensory feels" arise in evolution because once some types of processing feel like something by virtue of a system capable of higher-order thoughts, it is then parsimonious to postulate that sensory and related processing, which has to be taken into account in that processing system, should feel like something. It is suggested that it is this type of processing, which must be implemented in neural networks, which is related to consciousness.* © *1997 Elsevier Science Ltd.*

Keywords—Consciousness, Hippocampus, Memory, Invariance, Visual recognition, Higher-order thoughts, Visual cortex.

1. INTRODUCTION

Advances are being made in understanding *how* the brain could perform some of the processing involved in perception and memory. These advances come in part from neurophysiological experiments in which the processing involved in vision and memory is analysed by recording the activity of single neurons in primates during these types of processing, and incorporating this information into computational models at the neuronal network level which provide an account of the ways in which many neurons in the networks found in different brain regions could perform the required computations. Examples of this approach are described first in this paper.

Having considered brain mechanisms involved in visual object recognition and memory, I then consider whether, once this processing is fully understood, we will have produced an account of the brain mechanisms underlying consciousness. I argue that we will not, and that it is a different type of information processing that is involved in consciousness. I outline a theory of what the processing is that is involved in consciousness, of its adaptive value in an evolutionary perspective, and of how processing in our visual and other sensory systems can result in subjective or phenomenal states, the "raw feels" of conscious awareness. These processes involved in consciousness must themselves be implemented in neural networks, but before considering how these processes are implemented, it is useful to be clear about what processing must be implemented.

2. NEURONAL NETWORKS INVOLVED IN INVARIANT VISUAL OBJECT RECOGNITION

2.1. Neurophysiology

The visual pathways project in primates by a number of

Acknowledgements: The author has worked on some of the experiments described here with G. C. Baylis, M. Booth, M. J. Burton, P. Georges-François, M. E. Hasselmo, C. M. Leonard, F. Mora, D. I. Perrett, R. G. Robertson, M. K. Sanghera, T. R. Scott, S. J. Thorpe, and F. A. W. Wilson, and their collaboration, and helpful discussions with or communications from M. Davies and C. C. W. Taylor (Corpus Christi College, Oxford), and M. Stamp Dawkins, are sincerely acknowledged. Some of the research described was supported by the Medical Research Council (PG8513579), and by The Human Frontier Science Program.

Requests for reprints should be sent to Professor E. T. Rolls, University of Oxford, Department of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK; Tel.: +44-1865-271348; Fax: +44-1865-310447; e-mail: Edmund.Rolls@psy.ox.ac.uk

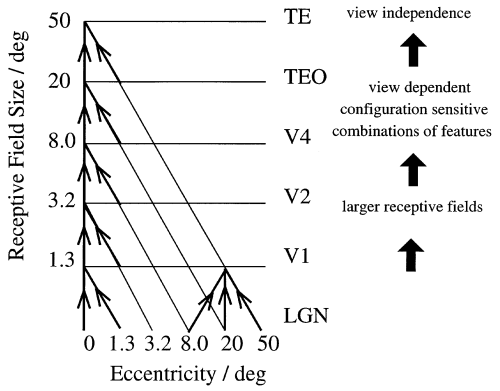

FIGURE 1. Schematic diagram showing convergence achieved by the forward projections in the visual system, and the types of representation that may be built by competitive networks operating at each stage of the system from the primary visual cortex (V1) to the inferior temporal visual cortex (area TE) (see text). LGN, Lateral geniculate nucleus. Area TEO forms the posterior inferior temporal cortex. The receptive fields in the inferior temporal visual cortex (e.g. in the TE areas) cross the vertical midline (not shown). (Reprinted from Wallis & Rolls, 1997.)

cortico-cortical stages from the primary visual cortex until they reach the temporal lobe visual cortical areas (see Figure 1, and for details of the neurophysiology summarized next, see Rolls, 1991, 1992, 1994b, 1995b, 1997). Along these pathways the receptive fields of neurons gradually become larger, as shown in Figure 1. (The receptive field of a neuron is the part of visual space within which appropriate visual stimuli can activate the neuron.) Part of the basis for this is the convergence onto neurons at any one stage of processing from a limited area of the preceding cortical area (see Figure 1). For this to result in neurons at the final stages of visual processing responding to the same object or stimulus independently of position on the retina, the appropriate connections must be set up in the hierarchy. Ways in which the appropriate synaptic weights to achieve this translation invariance could be learned are considered below.

The encoding that is provided of objects and faces at the end of this processing in the inferior temporal visual cortex is distributed, in the sense that the representation is not local or "grandmother cell" like, but instead many neurons are active to encode any one object (Rolls & Tovee, 1995; Rolls et al., 1996). Using an information-theoretic approach, it has been shown that the information available from the firing rates of a population of neurons about which visual stimulus (which of 20 equiprobable faces) has been shown on a single 500 ms presentation increases linearly with the number of neurons in the sample (Abbott et al., 1996; Rolls et al., 1997b). Because information is a logarithmic measure, this indicates that the number of stimuli encoded rises approximately exponentially, as the number of neurons in the sample increases. The consequence of this is that large numbers of stimuli, and fine discriminations between them, can be represented without (a receiving neuron) having to measure the activity of an enormous number of neurons. For example, the results of the experiments of Rolls et al. (1997b) indicate that the activity of 15 neurons would be able to encode 192 face stimuli (at 50% accuracy), of 20 neurons 768 stimuli, and of 25 neurons 3072 stimuli (Abbott et al., 1996; the values are for an optimal decoding case). This is strong evidence for distributed encoding. This type of encoding makes brain connectivity possible, in the sense that a receiving neuron can gain a great deal of information even when it does not receive vast numbers of inputs. Another interesting aspect of this encoding is that the information just described is available from the firing rates of the neurons, without taking into account the relative time at which the neurons fire. Thus temporal encoding is not an essential part of the code at this stage at least of visual information processing (see further Rolls et al., 1997b; Tovee & Rolls, 1995; Tovee et al., 1993). Another interesting aspect of the encoding is that much of the information from a population of neurons is available when the decoding is a simple neuronally plausible decoding involving a dot product of the neuronal activity in the current 500 ms (or 100 ms or 50 ms) presentation with that which occurred previously in the population of neurons to a particular stimulus (Rolls et al., 1997b). Such decoding could be performed by neurons which calculate their activation by a weighted sum of their input activity, which is common in neural network modelling. The fact that the information is available in a form in which it can be read out by this simple neuronally plausible dot product decoding with sampling from a limited set of neurons, and at the same time having the properties of a constant sparseness of the representation, and providing for generalization and graceful degradation, is probably what accounts for the fact that neurophysiologically interpretable information is available in the responses of *single* neurons about which stimulus has been seen (Rolls et al., 1997a; Tovee & Rolls, 1995; Tovee et al., 1993)[1]. This is one of the factors that allows single neuron recording to be so useful in understanding brain function—a correlation can frequently be found between the activity of even a single neuron and a subset of the stimuli being shown, of the motor responses being made, etc.

Some neurons in the temporal cortical visual areas have responses which are invariant not only for position on the retina, but also for the size, contrast, spatial frequency, position on the retina, and even angle of view

[1] The fact that the information increases approximately linearly with the number of neurons in the sample implies that the neurons convey almost independent information (if the stimulus set size is sufficiently large). If local encoding were used, the information would increase in proportion to the logarithm of the number of cells. If, for example, binary encoding were used (as, for example, numbers are encoded in a computer word), then the sparseness of the representation would fluctuate wildly, any receiving neuron would need to receive from all the input neurons, and generalization and graceful degradation would not occur.

(see Rolls, 1992, 1994b, 1995b, 1997; Rolls et al., 1996). It is clearly important that invariance in the visual system is made explicit in the neuronal responses, for this simplifies greatly the output of the visual system to memory systems such as the hippocampus and amygdala, which can then remember or form associations about *objects*. The function of these memory systems would be almost impossible if there were no consistent output from the visual system about objects (including faces), for then the memory systems would need to learn about all possible sizes, positions, etc., of each object, and there would be no easy generalization from one size or position of an object to that object when seen with another retinal size or position.

Other aspects of the neurophysiological findings which provide constraints on and guide the development of neural network theories about how the visual cortical areas involved in visual object recognition operate is that learning of new faces or objects can occur rapidly, within a few seconds; that the processing within any one cortical area is fast, with sufficient processing being completed within 30 ms in each cortical area in the hierarchy to subserve recognition; and that neurons in intermediate stages of processing (e.g. V2 and V4) respond to combinations of features present at earlier stages of processing (see Figure 1 and Rolls, 1992, 1994b, 1995b, 1997).

2.2. Computational Processes Involved in Invariant Visual Object Recognition

Cortical visual processing for object recognition is considered to be organized as a set of hierarchically connected cortical regions consisting at least of V1, V2, V4, posterior inferior temporal cortex (TEO), inferior temporal cortex (e.g. TE3, TEa and TEm), and anterior temporal cortical areas (e.g. TE2 and TE1). (This stream of processing has many connections with a set of cortical areas in the anterior part of the superior temporal sulcus, including area TPO.) There is convergence from each small part of a region to the succeeding region (or layer in the hierarchy) in such a way that the receptive field sizes of neurons (e.g. 1° near the fovea in V1) become larger by a factor of approximately 2.5 with each succeeding stage (and the typical parafoveal receptive field sizes found would not be inconsistent with the calculated approximations of, e.g. 8° in V4, 20° in TEO, and 50° in inferior temporal cortex; Boussaoud et al., 1991) (see Figure 1). Such zones of convergence would overlap continuously with each other (see Figure 1). This connectivity would be part of the architecture by which translation invariant representations are computed (see Rolls, 1992, 1994b, 1995b, 1996a; Wallis & Rolls, 1997). Each layer is considered to act partly as a set of local self-organizing competitive neuronal networks with overlapping inputs. (The region within which competition would be implemented would depend on the spatial properties of inhibitory interneurons, and might operate over distances of 1–2 mm in the cortex.) These competitive nets operate by a single set of forward inputs leading to (typically non-linear, e.g. sigmoid) activation of output neurons; of competition between the output neurons mediated by a set of feedback inhibitory interneurons which receive from many of the principal (in the cortex, pyramidal) cells in the net and project back to many of the principal cells, which serves to decrease the firing rates of the less active neurons relative to the rates of the more active neurons (i.e. soft competition); and then of synaptic modification by a modified Hebb rule, such that synapses to strongly activated output neurons from active input axons strengthen, and from inactive input axons weaken (see Rolls, 1989c; Rolls & Treves, 1997). (A biologically plausible form of this learning rule that operates well in such networks is

$$\delta w_{ij} = k \cdot m_i (r_j' - w_{ij})$$

where k is a constant, δw_{ij} is the change of synaptic weight, r_j' is the firing rate of the jth axon, and m_i is a non-linear function of the output activation of neuron i which mimics the operation of the NMDA receptors in learning; see Rolls, 1989a, b, c; Rolls & Treves, 1997). Related approaches to self-organization in the visual system are described by Linsker (1986, 1988) and MacKay & Miller (1990).

Translation invariance would be computed in such a system by utilizing competitive learning to detect regularities in inputs when real objects are translated in the physical world. The hypothesis is that because objects have continuous properties in space and time in the world, an object at one place on the retina might activate feature analysers at the next stage of cortical processing, and when the object was translated to a nearby position, because this would occur in a short period (e.g. 0.5 s), the membrane of the postsynaptic neuron would still be in its "Hebb-modifiable" state, and the presynaptic afferents activated with the object in its new position would thus become strengthened on the still-activated postsynaptic neuron. It is suggested (Rolls, 1992) that the short temporal window (e.g. 0.5 s) of Hebb-modifiability helps neurons to learn the statistics of objects moving in the physical world, and at the same time to form different representations of different feature combinations or objects, as these are physically discontinuous and present less regular correlations to the visual system. Foldiak (1991) has proposed computing an average activation of the postsynaptic neuron to assist with the same problem. Another suggestion is that a memory trace for what has been seen in the last 300 ms appears to be implemented by a mechanism as simple as continued firing of inferior temporal neurons after the stimulus has disappeared, as we have shown in masking experiments (see Rolls & Tovee, 1994; Rolls et al., 1994b). This continued firing could be implemented by local attractor networks in columns or modules in the cerebral cortex implemented by the local recurrent collaterals of the cortical

pyramidal cells (Rolls & Treves, 1997). Other invariances, for example, size, spatial frequency, and rotation invariance, could be learned by a comparable process. It is suggested that this process takes place at each stage of the multiple-layer cortical processing hierarchy, so that invariances are learned first over small regions of space, and then over successively larger regions. This limits the size of the connection space within which correlations must be sought. It is suggested that view-independent representations could be formed by the same type of computation, operating to combine a limited set of views of objects. Increasing complexity of representations could also be built in such a multiple layer hierarchy by similar mechanisms. At each stage or layer the self-organizing competitive nets would result in combinations of inputs becoming the effective stimuli for neurons.

To test and clarify these hypotheses (see further Rolls, 1992, 1994b, 1995b, 1997) about how the visual system may operate to learn invariant object recognition, we have performed a simulation which implements many of the ideas just described, and is consistent with and based on much of the neurophysiology summarized above. The network simulated can perform object, including face, recognition in a biologically plausible way, and after training shows, for example, translation and view invariance (Wallis & Rolls, 1997; Wallis et al., 1993).

In the four-layer network, the successive layers correspond approximately to V2, V4, the posterior temporal cortex, and the anterior temporal cortex. The forward connections to a cell in one layer are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities to determine the exact neurons in the preceding layer to which connections are made. This schema is constrained to preclude the repeated connection of any cells. Each cell receives 100 connections from the 32 $\times$ 32 cells of the preceding layer, with a 67% probability that a connection comes from within four cells of the distribution centre. Figure 2 shows the general convergent network architecture used, and may be compared with Figure 1. Within each layer, lateral inhibition between neurons has a radius of effect just greater than the radius of feedforward convergence just defined. The lateral inhibition is simulated via a linear local contrast enhancing filter active on each neuron. (Note that this differs from the global 'winner-take-all' paradigm implemented by Foldiak, 1991). The cell activation is then passed through a non-linear activation function (e.g. sigmoid), which also produces contrast enhancement of the firing rates.

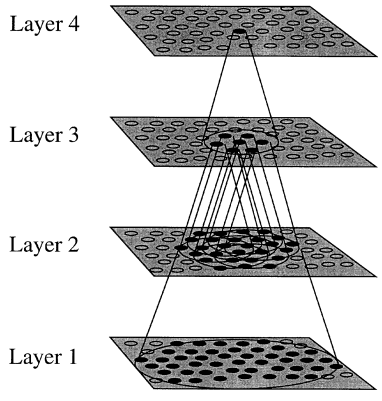

FIGURE 2. Hierarchical network structure used in the model of invariant visual object recognition. (Reprinted from Wallis & Rolls, 1997.)

So that the results of the simulation might be made particularly relevant to understanding processing in higher cortical visual areas, the inputs to layer 1 come from a separate input layer which provides an approximation to the encoding found in cortical visual area 1 (V1) of the primate visual system. These response characteristics of neurons in the input layer are provided by a series of spatially tuned filters with image contrast sensitivities chosen to accord with the general tuning profiles observed in the simple cells of V1.

The synaptic learning rule used in these simulations (VisNet) can be summarized as follows:

$$\delta w_{ij} = k m_i r_j'$$

and

$$m_i^t = (1-\eta) r_i^{(t)} + \eta m_i^{(t-1)}$$

where r_j' is the jth input to the neuron, r_i is the output of the ith neuron, w_{ij} is the jth weight on the ith neuron, η governs the relative influence of the trace and the new input (typically 0.4–0.6), and $m_i^{(t)}$ represents the value of the ith cell's memory trace at time t. In the simulations the neuronal learning was bounded by normalization of each cell's dendritic weight vector.

To train the network to produce a translation invariant representation, one stimulus was placed successively in a sequence of nine positions across the input, then the next stimulus was placed successively in the same sequence of nine positions across the input, and so on through the set of stimuli. The idea was to enable the network to learn whatever was common at each stage of the network about a stimulus shown in different positions. To train on view invariance, different views of the same object were shown in succession, then different views of the next object were shown in succession, and so on. It has been shown that this network, inspired by Fukushima's (Fukushima, 1980) neocognitron as well as by the neurophysiological data, can form cells in its final layer with translation, size and view invariant responses to stimuli presented on the 'retina' (Wallis & Rolls, 1997; Wallis et al., 1993).

These results show that the proposed learning mechanism and neural architecture can produce cells with responses selective for stimulus type with considerable

position, size or view invariance. The ability of the network to be trained with natural scenes is currently helping to advance our understanding of how representations of objects are built and encoded in the primate visual system.

This combined neurophysiological and computational approach is thus leading to biologically plausible theories about how the brain operates when it performs face or object recognition. In addition, there is now considerable evidence about what happens in our higher cortical visual areas when we recognize faces, and about how information about at least some classes of object in the world is represented in the visual system. Yet does this understanding of visual object recognition help us directly with the problem of consciousness, of why it is that it feels the way it does when we recognize a face? Would a computer which operated in the way described above be conscious during object recognition? I suggest that it would not be, and that for the object recognition processes to be conscious, including to feel like anything, the information from the type of visual processing system I describe would have to be projected to a different brain system, the nature of which will be described below. Before turning to that, some recent advances in understanding the brain processing that occurs when we store and then recall later everyday events are described, and I ask whether these memory processes are closer to consciousness.

3. THE HIPPOCAMPUS AND MEMORY

The hippocampus is implicated in a particular type of memory, the memory for recent events and episodes, in which there is frequently a spatial aspect or context (see for details Rolls, 1996b, d, 1997). In monkeys, a prototypical memory task impaired by damage to the hippocampal system is object-place memory, in which the locations of objects in space must be remembered (see Gaffan, 1994). This impairment is analogous to that shown by anterograde amnesic patients with damage to

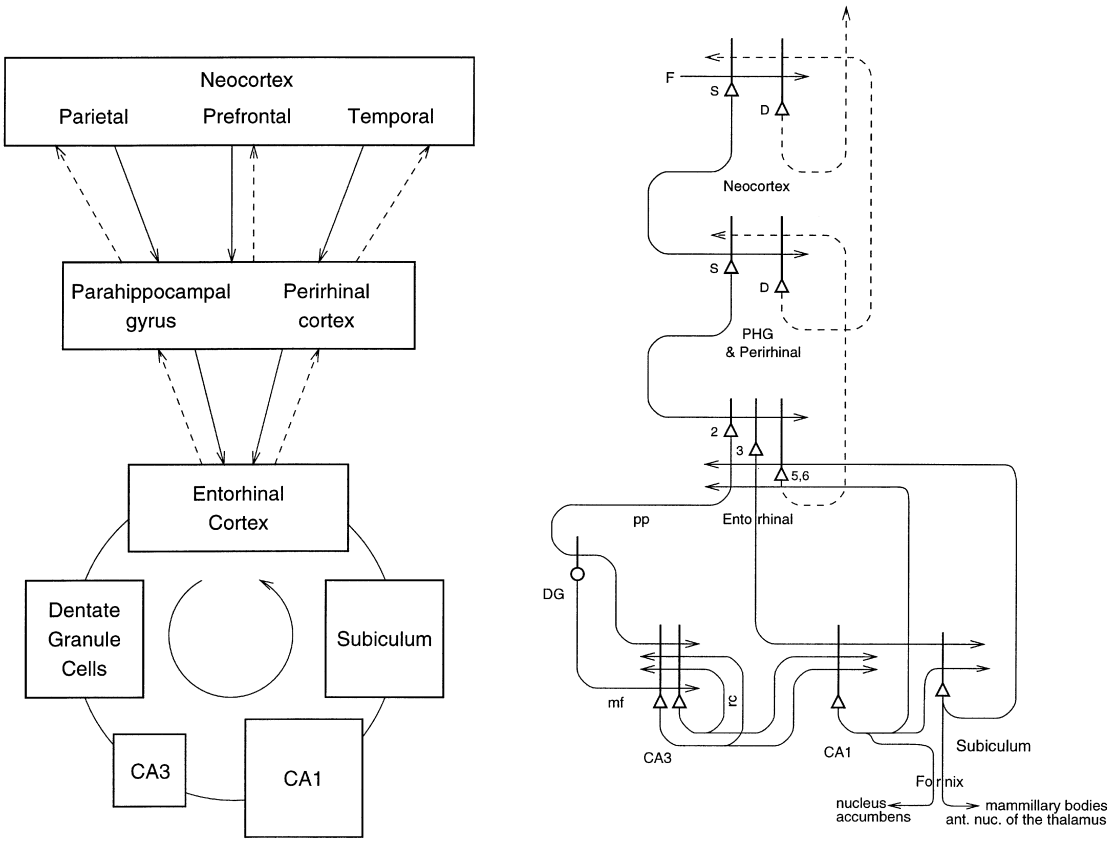

FIGURE 3. Forward connections (continuous lines) from areas of cerebral association neocortex via the parahippocampal gyrus and perirhinal cortex, and entorhinal cortex, to the hippocampus; and backprojections (dashed lines) via the hippocampal CA1 pyramidal cells, subiculum, and parahippocampal gyrus to the neocortex. There is great convergence in the forward connections down to the single network implemented in the CA3 pyramidal cells, and great divergence again in the backprojections. Left: block diagram. Right: more detailed representation of some of the principal excitatory neurons in the pathways. D, Deep pyramidal cells; DG, dentate granule cells; F, forward inputs to areas of the association cortex from preceding cortical areas in the hierarchy; mf, mossy fibres; PHG, parahippocampal gyrus and perirhinal cortex; pp, perforant path; rc, recurrent collateral of the CA3 hippocampal pyramidal cells; S, superficial pyramidal cells; 2, pyramidal cells in layer 2 of the entorhinal cortex; 3, pyramidal cells in layer 3 of the entorhinal cortex. The thick lines above the cell bodies represent the dendrites.

the hippocampus and nearby areas who cannot remember the locations of objects on a tray.

On the basis of these findings in humans and other animals, the hypothesis is suggested that the importance of the hippocampus in spatial and other memories is that it can rapidly form event or "episodic" representations of information originating from many areas of the cerebral cortex. In rats, hippocampal pyramidal cells (e.g. CA3 and CA1 neurons) respond when the rat is in a particular place in a spatial environment. In monkeys, it has been shown that there is a rich representation of space outside the monkey implemented by "spatial view" cells (see Rolls, 1996b, 1996d). These would provide an excellent representation of the spatial information needed to form a memory of where an object had been seen in space. It is suggested that an autoassociation network implemented by the CA3 cells of the hippocampus brings together the object information represented in temporal cortical visual areas, and spatial information represented in parietal areas, so that associations can be formed between objects and places (see Figure 3 and Rolls, 1989a, b, c, 1990a, 1996a,b).

A theory of how the networks shown in Figure 3 could operate, not only to store memories of events, but also to recall them to the neocortex via the backprojection pathways, has been developed (see Rolls, 1989a, b, 1996a; Rolls & Treves, 1997; Treves & Rolls, 1992, 1994). A way in which such recall could be useful in the cortex for building long-term semantic memories has been described by McClelland et al. (1995). A comparison of these approaches with others (for example by Burgess et al., 1994; and Hasselmo & Bower, 1993) is provided by Rolls (1996a), Rolls & Treves (1997) and Treves & Rolls (1994).

4. CONSCIOUSNESS

It would be possible to build a computer which would perform all the above functions of visual object recognition, memory storage and recall to the neocortex, and even emotion (Rolls, 1990b, 1995c), using the same computational principles described above, and yet we might not want to ascribe subjective or phenomenal states, which I shall call qualia, to this computer. We might not want to say that it feels like something to the computer when the computer is performing these functions. This raises the issue of in which networks in the brain would consciousness be represented. Because the topic of subjective or phenomenal feels or feelings (that it feels like something to be in that state) is of considerable current interest, and is for the present purposes the defining aspect of consciousness, one view on consciousness, influenced by contemporary cognitive neuroscience, is outlined next. However, this view is only preliminary, and theories of consciousness are likely to develop considerably. A reason for describing this view of consciousness is that we need to be clear about *what* must be implemented before considering *how* it could be implemented in neural networks.

A starting point is that many actions can be performed relatively automatically, without apparent conscious intervention. An example sometimes given is driving a car. Such actions could involve control of behaviour by brain systems which are old in evolutionary terms such as the basal ganglia. It is of interest that the basal ganglia (and cerebellum) do not have backprojection systems to most of the parts of the cerebral cortex from which they receive inputs (see, e.g. Rolls, 1994a; Rolls & Johnstone, 1992). In contrast, parts of the brain such as the hippocampus and amygdala, involved in functions such as episodic memory and emotion respectively, about which we can make (verbal) declarations (hence declarative memory, Squire, 1992) do have major backprojection systems to the high parts of the cerebral cortex from which they receive forward projections (Rolls, 1992; Rolls & Treves, 1997; Treves & Rolls, 1994; see Figure 3). It may be that evolutionarily newer parts of the brain, such as the language areas and parts of the prefrontal cortex, are involved in an alternative type of control of behaviour, in which actions can be planned with the use of a (language) system which allows relatively arbitrary (syntactic) manipulation of semantic entities (symbols).

The general view that there are many routes to behavioural output is supported by the evidence that there are many input systems to the basal ganglia (from almost all areas of the cerebral cortex), and that neuronal activity in each part of the striatum reflects the activity in the overlying cortical area (Rolls, 1994a; Rolls & Johnstone, 1992). The evidence is consistent with the possibility that different cortical areas, each specialized for a different type of computation, have their outputs directed to the basal ganglia, which then select the strongest input, and map this into action (via outputs directed, for example, to the premotor cortex) (Rolls & Johnstone, 1992; Rolls & Treves, 1997). Within this scheme, the language areas would offer one of many routes to action, but a route particularly suited to planning actions, because of the syntactic manipulation of semantic entities which may make long-term planning possible. A schematic diagram of this suggestion is provided in Figure 4. Consistent with the hypothesis of multiple routes to action, only some of which utilize language, is the evidence that split-brain patients may not be aware of actions being performed by the "non-dominant" hemisphere (Gazzaniga, 1988, 1995; Gazzaniga & LeDoux, 1978). Also consistent with multiple including non-verbal routes to action, patients with focal brain damage, for example to the prefrontal cortex, may emit actions, yet comment verbally that they should not be performing those actions (Rolls et al., 1994a). In both these types of patient, confabulation may occur, in that a verbal account of why the action was performed may be given, and this may not be related at all to the

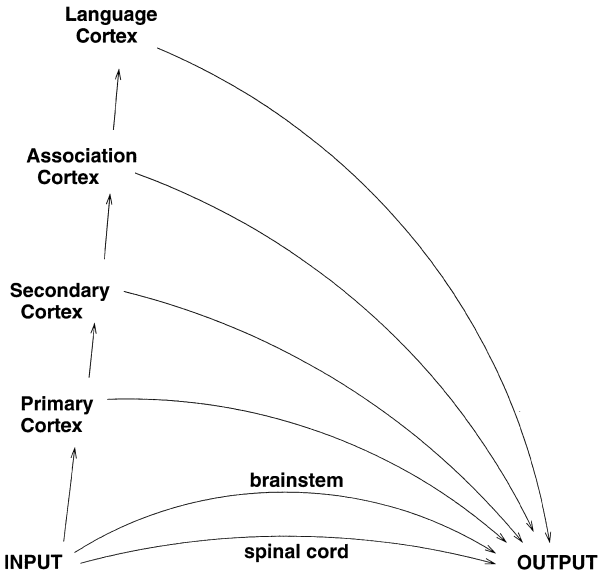

FIGURE 4. Schematic illustration indicating many possible routes from input systems to action (output) systems. Cortical information processing systems are organized hierarchically, and there are routes to output systems from most levels of the hierarchy.

environmental event which actually triggered the action (Gazzaniga, 1988, 1995; Gazzaniga & LeDoux, 1978). It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region such as those involved in some types of rewarded behaviour, the language system may subsequently elaborate a coherent account of why that action was performed (i.e. confabulate). This would be consistent with a general view of brain evolution in which as areas of the cortex evolve, they are laid on top of existing circuitry connecting inputs to outputs, and in which each level in this hierarchy of separate input–output pathways may control behaviour according to the specialized function it can perform (see schematic diagram in Figure 4). (It is of interest that mathematicians may have a hunch that something is correct, yet not be able to verbalize why. They may then resort to formal, more serial and language-like theorems to prove the case, and these seem to require conscious processing. This is a further indication of a close association between linguistic processing and consciousness. The linguistic processing need not, as in reading, involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements, and to perform syntactic operations on the semantic representations of states in the world, would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g. I think that he thinks that···), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others. This capability for higher-order thoughts would also allow reflection on past events, which would also be useful in planning. In contrast, non-linguistic behaviour would be driven by learned reinforcement associations, learned rules, etc., but not by flexible planning for many steps ahead involving a model of the world including others' behaviour. (For an earlier view which is close to this part of the argument, see Humphrey, 1980.) (The examples of behaviour from non-humans that may reflect planning may reflect much more limited and inflexible planning. For example, the dance of the honey-bee to signal to other bees the location of food may be said to reflect planning, but the symbol manipulation is not arbitrary. There are likely to be interesting examples of non-human primate behaviour, perhaps in the great apes, that reflect the evolution of an arbitrary symbol-manipulation system that could be useful for flexible planning; see Cheney & Seyfarth, 1990.) It is important to state that the language ability referred to here is not necessarily human verbal language (though this would be an example). What it is suggested is important to planning is the syntactic manipulation of symbols, and it is this syntactic manipulation of symbols which is the sense in which language is defined and used here.

It is next suggested that this arbitrary symbol-manipulation using important aspects of language processing and used for planning but not in initiating all types of behaviour is close to what consciousness is about. In particular, consciousness may *be* the state which arises in a system that can think about (or reflect on) its own (or other peoples') thoughts, that is, in a system capable of second- or higher-order thoughts (Rosenthal, 1986, 1990, 1993; compare Dennett, 1991). On this account, a mental state is non-introspectively (i.e. non-reflectively) conscious if one has a roughly simultaneous thought that one is in that mental state. Following from this, introspective consciousness (or reflexive consciousness, or self consciousness) is the attentive, deliberately focused consciousness of one's mental states. It is noted that not all of the higher-order thoughts need themselves be conscious (many mental states are not). However, according to the analysis, having a higher-order thought about a lower-order thought is necessary for the lower-order thought to be conscious. (A slightly weaker position than Rosenthal's on this is that a conscious state corresponds to a first-order thought that has the *capacity* to cause a second-order thought or judgement about it—Carruthers, 1996). This analysis is consistent with the points made above that the brain systems that are required for consciousness and language are similar. In particular, a system that can have second- or higher-order thoughts about its own operation, including its planning and linguistic operation, must itself be a language

processor, in that it must be able to bind correctly to the symbols and syntax in the first-order system. According to this explanation, the feeling of anything is the state which is present when linguistic processing that involves second- or higher-order thoughts is being performed.

It might be objected that this captures some of the process aspects of consciousness, what it is good for in an information processing system, but does not capture the phenomenal aspect of consciousness. I agree that there is an element of "mystery" that is invoked at this step of the argument, when I say that it feels like something for a machine with higher-order thoughts to be thinking about its own first- or lower-order thoughts. But the return point is the following: *if a human with second-order thoughts is thinking about his or her first-order thoughts, surely it is very difficult for us to conceive that this would* NOT *feel like something*? This is especially the case when the first-order thoughts are linguistic, and are about (grounded in) the real world.

It is suggested that part of the evolutionary adaptive significance of this type of higher-order thought is that it allows correction of errors made in first-order linguistic or in non-linguistic processing. Indeed, the ability to reflect on previous events is extremely important for learning from them, including setting up new long-term semantic structures. It was shown above that the hippocampus may be a system for such "declarative" recall of recent memories. Its close relation to "conscious" processing in humans (Squire has classified it as a declarative memory system) may be simply that it allows the recall of recent memories, which can then be reflected upon in conscious, higher-order, processing. Another part of the adaptive value of a higher-order thought system may be that by thinking about its own thoughts in a given situation, it may be able to better understand the thoughts of another individual in a similar situation, and therefore predict that individual's behaviour better (Humphrey, 1980).

As a point of clarification, I note that according to this theory, a language processing system is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have higher-order thoughts, and a first-order language processor (that might be perfectly competent at language) would not be conscious, in that it could not think about its own or others' thoughts. One can perfectly well conceive of a system which obeyed the rules of language (which is the aim of much connectionist modelling), and implemented a first-order linguistic system, that would not be conscious. (Possible examples of language processing that might be performed non-consciously include computer programs implementing aspects of language, or ritualized human conversations, e.g. about the weather. These might require syntax and correctly grounded semantics, and yet be performed non-consciously. A more complex example, illustrating that syntax could be used, might be: "If A does X, then B will probably do Y, and then C would be able to do Z." A first-order language system could process this statement. Moreover, the first-order language system could apply the rule usefully in the world, provided that the symbols in the language system (A, B, X, Y, etc.) are grounded (have meaning) in the world.) In line with the argument on the adaptive value of higher-order thoughts and thus consciousness given above, that they are useful for correcting lower-order thoughts, I now suggest that correction using higher-order thoughts of lower-order thoughts would have adaptive value primarily if the lower-order thoughts are sufficiently complex to benefit from correction in this way. The nature of the complexity is specific: that it should involve syntactic manipulation of symbols, probably with several steps in the chain, and that the chain of steps should be a one-off set of steps, as in a particular plan or sentence, rather than a set of well-learned rules. The first- or lower-order thoughts might involve a linked chain of "if⋯then" statements that would be involved in planning, an example of which has been given above. It is partly because complex lower-order thoughts such as these, which involve syntax and language, would benefit from correction by higher-order thoughts, that I suggest that there is a close link between this reflective consciousness and language. The hypothesis is that by thinking about lower-order thoughts, the higher-order thoughts can discover what may be weak links in the chain of reasoning at the lower-order level, and having detected the weak link, might alter the plan, to see if this gives better success. In our example above, if it transpired that C could not do Z, how might the plan have failed? Instead of having to go through endless random changes to the plan to see if by trial and error some combination does happen to produce results, what I am suggesting is that by thinking about the previous plan, one might, for example, using knowledge of the situation and the probabilities that operate in it, guess that the step where the plan failed was that B did not in fact do Y. So by thinking about the plan (the first- or lower-order thought), one might correct the original plan, in such a way that the weak link in that chain, that "B will probably do Y", is circumvented. To draw a parallel with neural networks: there is a "credit assignment" problem in such multistep syntactic plans, in that if the whole plan fails, how does the system assign credit or blame to particular steps of the plan? The suggestion is that this is the function of higher-order thoughts and is why systems with higher-order thoughts evolved. The suggestion I then make is that if a system were doing this type of processing (thinking about its own thoughts), it would then be very plausible that it should feel like something to be doing this. I even suggest to the reader that it is not plausible to suggest that it would not feel like anything to a system if it were doing this.

Two other points in the argument should be emphasized for clarity. One is that the system that is having

syntactic thoughts about its own syntactic thoughts would have to have its symbols grounded in the real world for it to feel like something to be having higher-order thoughts. The intention of this clarification is to exclude systems such as a computer running a program when there is in addition some sort of control or even overseeing program checking the operation of the first program. We would want to say that in such a situation it would feel like something to be running the higher-level control program only if the first-order program was symbolically performing operations on the world and receiving input about the results of those operations, and if the higher-order system understood what the first-order system was trying to do in the world. The second clarification is that the plan would have to be a unique string of steps, in much the same way as a sentence can be a unique and one-off string of words. The point here is that it is helpful to be able to think about particular one-off plans, and to correct them; and that this type of operation is very different from the slow learning of fixed rules by trial and error.

This analysis does not yet give an account for sensory qualia ("raw sensory feels"; for example, why "red" feels red), for emotional qualia (e.g. why a rewarding touch produces an emotional feeling of pleasure), or for motivational qualia (e.g. why food deprivation makes us *feel* hungry). The view I suggest on such qualia is as follows. Information processing in and from our sensory systems (e.g. the sight of the colour red) may be relevant to planning actions using language and the conscious processing thereby implied. Given that these inputs must be represented in the system that plans, we may ask whether it is more likely that we would be conscious of them or that we would not. I suggest that it would be a very special-purpose system that would allow such sensory inputs, and emotional and motivational states, to be part of (linguistically based) planning, and yet remain unconscious. It seems to be much more parsimonious to hold that we would be conscious of such sensory, emotional and motivational qualia because they would be being used (or are available to be used) in this type of (linguistically based) higher-order thought processing, and this is what I propose.

The explanation of emotional and motivational subjective feelings or qualia that this discussion has led towards is thus that they should be felt as conscious because they enter into a specialized linguistic symbol-manipulation system that is part of a higher-order thought system that is capable of reflecting on and correcting its lower-order thoughts involved, for example, in the flexible planning of actions. It would require a very special machine to allow this higher-order linguistically based thought processing, which is conscious by its nature, to occur without the sensory, emotional and motivational states (which must be taken into account by the higher-order thought system) becoming felt qualia. The qualia are thus accounted for by the evolution of the linguistic system that can reflect on and correct its own lower-order processes, and thus has adaptive value.

This account implies that it may be especially animals with a higher-order belief and thought system and with linguistic symbol manipulation that have qualia. It may be that much non-human animal behaviour, provided that it does not require flexible linguistic planning and correction by reflection, could take place according to reinforcement-guidance (using, e.g. stimulus-reinforcement association learning in the amygdala and orbitofrontal cortex, Rolls, 1990b, 1996c), and rule-following (implemented, e.g. using habit or stimulus-response learning in the basal ganglia, Rolls, 1994a; Rolls & Johnstone, 1992). Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals close to humans, with demonstrable syntactic manipulation of symbols, and the ability to think about these linguistic processes) would be different from other animals, and would have evolved qualia.

For processing in a part of our brain to be able to reach consciousness, appropriate pathways must be present. Certain constraints arise here. For example, in the sensory pathways, the nature of the representation may change as it passes through a hierarchy of processing levels, and in order to be conscious of the information in the form in which it is represented in early processing stages, the early processing stages must have access to the part of the brain necessary for consciousness. An example is provided by processing in the taste system. In the primate primary taste cortex, neurons respond to taste independently of hunger, yet in the secondary taste cortex, food-related taste neurons (e.g. responding to sweet taste) only respond to food if hunger is present, and gradually stop responding to that taste during feeding to satiety (see Rolls, 1989d, 1993, 1995a). Now the quality of the tastant (sweet, salt, etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety. The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex), and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it). The same argument arises for reward in general, and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is

reflected in the responses of neurons to visual stimuli (see Rolls, 1990b, 1995a, c). It is also of interest that reward signals (e.g. the taste of food when we are hungry) are associated with subjective feelings of pleasure (see Rolls, 1990b, 1993, 1995a, c). I suggest that this correspondence arises because pleasure is the subjective state that represents in the conscious system a signal that is positively reinforcing (rewarding), and that inconsistent behaviour would result if the representations did not correspond to a signal for positive reinforcement in both the conscious and the non-conscious processing systems.

Do these arguments mean that the conscious sensation of, for example, taste quality (i.e. identity and intensity) is represented or occurs in the primary taste cortex, and of the pleasantness of taste in the secondary taste cortex, and that activity in these areas is sufficient for conscious sensations (qualia) to occur? I do not suggest this at all. Instead, the arguments I have put forward above suggest that we are only conscious of representations when we have high-order thoughts about them. The implication then is that pathways must connect from each of the brain areas in which information is represented about which we can be conscious, to the system which has the higher-order thoughts, which, as I have argued above, requires language. Thus, in the example given, there must be connections to the language areas from the primary taste cortex, which need not be direct, but which must bypass the secondary taste cortex, in which the information is represented differently (see Rolls, 1989d, 1995a). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have higher-order thoughts about the pleasantness of the representation in the secondary taste cortex. There would also need to be pathways from the hippocampus, implicated in the recall of declarative memories, back to the language areas of the cerebral cortex (at least via the cortical areas which receive backprojections from the hippocampus, see Figure 3, which would in turn need connections to the language areas). A schematic diagram incorporating this anatomical prediction about human cortical neural connectivity in relation to consciousness is shown in Figure 5.

One question that has been discussed is whether there is a causal role for consciousness (e.g. Armstrong & Malcolm, 1984). The position to which the above arguments lead is that indeed conscious processing does have a causal role in the elicitation of behaviour, but only under the set of circumstances when higher-order thoughts play a role in correcting or influencing lower-order thoughts. The sense in which the consciousness is causal is then, it is suggested, that the higher-order thought is causally involved in correcting the lower-order thought; and that it is a property of the higher-order thought system that it feels like something when it is operating. As we have seen, some behavioural responses can be elicited when there is not this type of

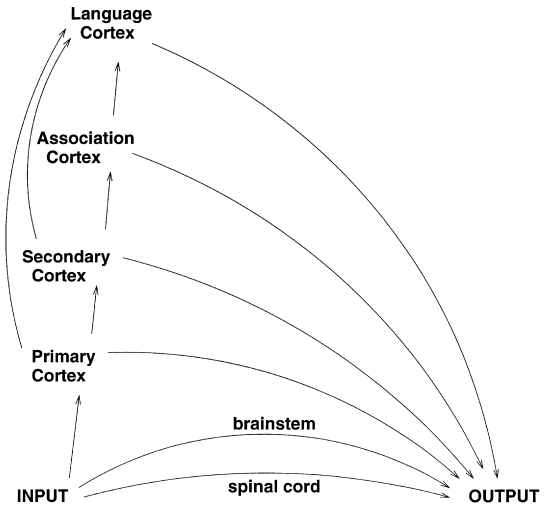

FIGURE 5. Schematic illustration indicating that early cortical stages in information processing may need access to language areas which bypass subsequent levels in the hierarchy, so that consciousness of what is represented in early cortical stages, and which may not be represented in later cortical stages, can occur. Higher-order linguistic thoughts (HOLTs) could be implemented in the language cortex itself, and would not need a separate cortical area. Backprojections, a notable feature of cortical connectivity, with many probable functions including recall (Rolls, 1989a, b, 1996a), probably reciprocate all the connections shown.

reflective control of lower-order processing, nor indeed any contribution of language. There are many brain processing routes to output regions, and only one of these involves conscious, verbally represented processing which can later be recalled (see Figure 4).

Some of the brain systems involved in this type of conscious processing that it is suggested has evolved to help the correction of plans are as follows. One module is a system that can implement syntax, because the many symbols (e.g. names of people) that are part of the plan must be correctly linked or bound. Such linking might be of the form: "if A does this, then B is likely to do this, and this will cause C to do this⋯". The requirement of syntax for this type of planning implies that an output to language systems in the brain is required for this type of planning (see Figure 4). Another building block for such planning operations in the brain may be the type of short-term memory in which the prefrontal cortex is involved. This short-term memory may be, for example, in non-human primates of where in space a response has just been made. A development of this type of short-term response memory system in humans to allow multiple short-term memories to be held in place correctly, preferably with the temporal order of the different items in the short-term memory coded correctly, may be another building block for the multiple step "if⋯then" type of computation so as to form a multiple step plan. Such short-term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of

non-human primates and humans (see Goldman-Rakic, 1996; Petrides, 1996), and may be part of the reason why prefrontal cortex damage impairs planning (see Shallice & Burgess, 1996).

It is of interest to comment on how the evolution of a system for flexible planning might affect emotions. Consider grief which may occur when a reward is terminated and no immediate action is possible (see Rolls, 1990b, 1995c). It may be adaptive by leading to a cessation of the formerly rewarded behaviour and thus facilitating the possible identification of other positive reinforcers in the environment. In humans, grief may be particularly potent because it becomes represented in a system which can plan ahead, and understand the enduring implications of the loss. (Thinking about or verbally discussing emotional states may also in these circumstances help, because this can lead towards the identification of new or alternative reinforcers, and of the realization that, for example, the negative consequences may not be as bad as feared.)

This account of consciousness also leads to a suggestion about the processing that underlies the feeling of free will. Free will would in this scheme involve the use of language to check many moves ahead on a number of possible series of actions and their outcomes, and then with this information to make a choice from the likely outcomes of different possible series of actions. (If, in contrast, choices were made only on the basis of the reinforcement value of immediately available stimuli, without the arbitrary syntactic symbol manipulation made possible by language, then the choice strategy would be much more limited, and we might not want to use the term free will, as all the consequences of those actions would not have been computed.) It is suggested that when this type of reflective, conscious, information processing is occurring and leading to action, the system performing this processing and producing the action would have to believe that it could cause the action, for otherwise inconsistencies would arise, and the system might no longer try to initiate action. This belief held by the system may partly underlie the feeling of free will. At other times, when other brain modules are initiating actions, the conscious processor may confabulate and believe that it caused the action, or at least give an account (possibly wrong) of why the action was initiated. The fact that the conscious processor may have the belief even in these circumstances that it initiated the action may arise as a property of it being inconsistent for a system which can take overall control using conscious verbal processing to believe that it was overridden by another system.

In the operation of such a free will system, the uncertainties introduced by the limited information possible about the likely outcomes of series of actions, and the inability to use optimal algorithms when combining conditional probabilities, would be much more important factors than whether the brain operates deterministically or not. (The operation of brain machinery must be relatively deterministic, for it has evolved to provide reliable outputs for given inputs.)

These are my initial thoughts on why we have consciousness, and are conscious of sensory, emotional and motivational qualia, as well as qualia associated with first-order linguistic thoughts. It is likely that theories of consciousness will continue to undergo rapid development, and current theories should not be taken to have practical implications.

5. DISCUSSION

Some ways in which the current theory may be different from other related theories follow. The current theory holds that it is higher-order *linguistic* thoughts (HOLTs) that are closely associated with consciousness, and this may differ from Rosenthal's higher-order thoughts (HOTs) theory (Rosenthal, 1986, 1990, 1993), in the emphasis in the current theory on language. Similarly, the theory differs from suggestions for a function of consciousness in "monitoring" (e.g. Marcel, 1988), in that a specification is given in the present theory of the type of correction being performed of first-order linguistic thought processes, and of the computational advantages of this. Language in the current theory is defined by syntactic manipulation of symbols, and does not necessarily imply verbal language. The reason that strong emphasis is placed on language is that it is as a result of having a multistep flexible "on the fly" reasoning procedure that errors which cannot be easily corrected by reward or punishment received at the end of the reasoning, need 'thoughts about thoughts', that is, some type of supervisory and monitoring process, to detect where errors in the reasoning have occurred. This suggestion on the adaptive value in evolution of such a higher-order linguistic thought process for multistep planning ahead, and correcting such plans, may also be different from earlier work. Put another way, this point is that credit assignment when reward or punishment are received is straightforward in a one layer network (in which the reinforcement can be used directly to correct nodes in error, or responses), but is very difficult in a multistep linguistic process executed once "on the fly". Very complex mappings in a multilayer network can be learned if hundreds of learning trials are provided. But once these complex mappings are learned, their success or failure in a new situation on a given trial cannot be evaluated and corrected by the network. Indeed, the complex mappings achieved by such networks (e.g. back-propagation nets) mean that after training they operate according to fixed rules, and are often impenetrable and inflexible. In contrast, to correct a multistep, single occasion, linguistically based plan or procedure, recall of the steps just made in the reasoning or planning, and perhaps related episodic material, needs to occur, so that the link in the chain which is most likely to be in error

can be identified. This may be part of the reason why there is a close relation between declarative memory systems, which can explicitly recall memories, and consciousness.

Some computer programs may have supervisory processes. Should these count as higher-order linguistic thought processes? My current response to this is that they should not, to the extent that they operate with fixed rules to correct the operation of a system which does not itself involve linguistic thoughts about symbols grounded semantically in the external world. If, on the other hand, it were possible to implement on a computer such a high-order linguistic thought supervisory correction process to correct first-order linguistic thoughts with symbols grounded in the real world, then this process would prima facie be conscious. If it were possible in a thought experiment to reproduce the neural connectivity and operation of a human brain on a computer, then prima facie it would also have the attributes of consciousness. It might continue to have those attributes for as long as power was applied to the system.

Another possible difference from earlier theories is that raw sensory feels are suggested to arise as a consequence of having a system that can think about its own thoughts. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution.

A property often attributed to consciousness is that it is unitary. The current theory would account for this by the limited syntactic capability of neuronal networks in the brain, which renders it difficult to implement more than a few syntactic bindings of symbols simultaneously (see McLeod et al., 1998; Rolls & Treves, 1997). This limitation makes it difficult to run several "streams of consciousness" simultaneously. In addition, given that a linguistic system can control behavioural output, several parallel streams might produce maladaptive behaviour (apparent as, e.g. indecision), and might be selected against. The close relation between, and the limited capacity of, both the stream of consciousness, and auditory–verbal short-term memory, may be that both implement the capacity for syntax in neural networks. Whether syntax in real neuronal networks is implemented by temporal binding (see von der Malsburg, 1990) is still an unresolved issue (see Rolls & Treves, 1997). (For example, the code can be read off from the end of the visual system without taking the temporal aspects of the neuronal firing into account, as described above; much of the information about which stimulus is shown is available in short times of 30–50 ms, and cortical neurons need fire for only this long during the identification of objects (Rolls & Tovee, 1994; Rolls et al., 1994b; Tovee & Rolls, 1995; Tovee et al., 1993) (these are rather short time windows for the expression of multiple separate populations of synchronized neurons); and oscillations, at least, are not an obvious property of neuronal firing in the primate temporal cortical visual areas involved in the representation of faces and objects (Tovee & Rolls, 1992).)

The current theory holds that consciousness arises by virtue of a system that can think linguistically about its own linguistic thoughts. The advantages for a system of being able to do this have been described, and this has been suggested as the reason why consciousness evolved. The evidence that consciousness arises by virtue of having a system that can perform higher-order linguistic processing is, however, and I think may remain, circumstantial. (Why must it feel like something when we are performing a certain type of information processing? The evidence described here suggests that it does feel like something when we are performing a certain type of information processing, but does not produce a strong reason for why it has to feel like something. It just does, when we are using this linguistic processing system capable of higher-order thoughts.) The evidence, summarized above, includes the points that we think of ourselves as conscious when, for example, we recall earlier events, compare them with current events, and plan many steps ahead. Evidence also comes from neurological cases, from, for example, split brain patients (who may confabulate conscious stories about what is happening in their other, non-language, hemisphere), and from cases such as frontal lobe patients who can tell one consciously what they should be doing, but nevertheless may be doing the opposite. (The force of this type of case is that much of our behaviour may normally be produced by routes about which we cannot verbalize, and are not conscious about.) This raises the issue of the causal role of consciousness. Does consciousness cause our behaviour?[2] The view that I currently hold is that the information processing which is related to consciousness (activity in a linguistic system capable of higher-order thoughts, and used for planning and correcting the operation of lower-order linguistic systems) can play a causal role in producing our behaviour (see Figure 4). It is, I postulate, a property of processing in this system (capable of higher-order thoughts) that it feels like something to be performing that type of processing. It is in this sense that I suggest that consciousness can act causally to influence our behaviour: consciousness is the property that occurs when a linguistic system is thinking about its lower-order thoughts. The hypothesis that it does feel like something when this processing is taking

[2] This raises the issue of the causal relation between mental events and neurophysiological events, part of the mind–body problem. My view is that the relation between mental events and neurophysiological events is similar (apart from the problem of consciousness) to the relation between the program running in a computer and the hardware in the computer. In a sense, the program causes the logic gates to move to the next state. This move causes the program to move to its next state. Effectively, we are looking at different levels of what is overall the operation of a *system*, and causality can usefully be understood as operating both within levels (causing one step of the program to move to the next), as well as between levels (e.g. software to hardware and vice versa).

place is at least to some extent testable: humans performing this type of higher-order linguistic processing, for example, recalling episodic memories and comparing them with current circumstances, who denied being conscious, would prima facie constitute evidence against the theory. Most humans would find it very implausible though to posit that they could be thinking about their own thoughts, and reflecting on their own thoughts, without being conscious. This type of processing does appear to be for most humans to be necessarily conscious.

Finally, I provide a short specification of what might have to be implemented in a neural network to implement conscious processing. First, a linguistic system, not necessarily verbal, but implementing syntax between symbols implemented in the environment would be needed. Then a higher-order thought system also implementing syntax and able to think about the representations in the first-order language system, and able to correct the reasoning in the first-order linguistic system in a flexible manner, would be needed. So my answer to the title of this paper is that consciousness can be implemented in neural networks (and that this is a topic worth discussing), but that the neural networks would have to implement the type of higher-order linguistic processing described in this paper.

REFERENCES

Abbott, L. A., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex, 6,* 498–505.

Armstrong, D. M., & Malcolm, N. (1984). *Consciousness and Causality*. Oxford: Blackwell.

Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *Journal of Computational Neurology, 306,* 554–575.

Burgess, N., Recce, M., & O'Keefe, J. (1994). A model of hippocampal function. *Neural Networks, 7,* 1065–1081.

Carruthers, P. (1996). *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.

Cheney, D. L., & Seyfarth, R. M. (1990). *How Monkeys See the World.* Chicago: University of Chicago Press.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3,* 193–199.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Gaffan, D. (1994). Scene-specific memory for objects: a model of episodic memory impairment in monkeys with fornix transection. *Journal of Cognitive Neuroscience, 6,* 305–320.

Gazzaniga, M. S. (1988). Brain modularity: towards a philosophy of conscious experience. In A.J. Marcel & E. Bisiach (Eds.), *Consciousness in Contemporary Science* (Ch. 10, pp. 218–238). Oxford: Oxford University Press.

Gazzaniga, M. S. (1995). Consciousness and the cerebral hemispheres. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1392–1400). Cambridge, MA: MIT Press.

Gazzaniga, M. S., & LeDoux, J. (1978). *The Integrated Mind.* New York: Plenum.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society, Series B, 351,* 1445–1453.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neuroscience, 16,* 218–222.

Humphrey, N. K. (1980). Nature's psychologists. In B. D. Josephson & V. S. Ramachandran (Eds.), *Consciousness and the Physical World* (pp. 57–80). Oxford: Pergamon.

Linsker, E. (1986). From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences of the USA* **83**, 7508–7512, 8390–8394, 8779–8783.

Linsker, E. (1988). Self-organization in a perceptual network. *Computer, March 1988,* 105–117.

MacKay, D. J. C., & Miller, K. D. (1990). Analysis of Linsker's simulation of Hebbian Rules. *Neural Computation, 2,* 173–187.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Marcel, A. J. (1988). Phenomenal experience and functionalism. In A. J. Marcel and E. Bisiach (Eds.), *Consciousness in Contemporary Science* (pp. 121–158). Oxford: Oxford University Press.

Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Philosophical Transactions of the Royal Society, Series B, 351,* 1455–1462.

Rolls, E. T. (1989a). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne & W. O. Berry (Eds.), *Neural Models of Plasticity: Experimental and Theoretical Approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rolls, E. T. (1989b). The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron* (pp. 125–159). Wokingham, UK: Addison Wesley.

Rolls, E. T. (1989c). Functions of neuronal networks in the hippocampus and cerebral cortex in memory. In R. M. J. Cotterill (Ed.), *Models of Brain Function* (pp. 15–33). Cambridge: Cambridge University Press.

Rolls, E.T. (1989d). Information processing in the taste system of primates. *Journal of Experimental Biology, 146,* 141–164.

Rolls, E. T. (1990a). Functions of the primate hippocampus in spatial processing and memory. In D. S. Olton & R. P. Kesner (Eds.), *Neurobiology of Comparative Cognition* (pp. 339–362). Hillsdale, NJ: Lawrence Erlbaum.

Rolls, E.T. (1990b). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion, 4,* 161–190.

Rolls, E.T. (1991). Neural organisation of higher visual functions. *Current Opinion in Neurobiology, 1,* 274–278.

Rolls, E.T. (1992). Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society, 335,* 11–21.

Rolls, E. T. (1993). The neural control of feeding in primates. In D. A. Booth (Ed.), *Neurophysiology of Ingestion* (pp. 137–169). Oxford: Pergamon.

Rolls, E.T. (1994a). Neurophysiology and cognitive functions of the striatum. *Revue Neurologique (Paris), 150,* 648–660.

Rolls, E.T. (1994b). Brain mechanisms for invariant visual recognition and learning. *Behavioural Processes, 33,* 113–138.

Rolls, E. T. (1995a). Central taste anatomy and neurophysiology. In R. L. Doty (Ed.), *Handbook of Olfaction and Gustation* (pp. 549–573). New York: Dekker.

Rolls, E.T. (1995b). Learning mechanisms in the temporal lobe visual cortex. *Behavioural Brain Research, 66,* 177–185.

Rolls, E. T. (1995c). A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1091–1106). Cambridge, MA: MIT Press.

Rolls, E.T. (1996a). A theory of hippocampal function in memory. *Hippocampus, 6,* 601–620.

Rolls, E. T. (1996b). The representation of space in the primate hippocampus, and episodic memory. In T. Ono, B. L. McNaughton, S. Molotchnikoff, E. T. Rolls, & H. Nishijo (Eds.), *Perception, Memory and Emotion: Frontier in Neuroscience* (pp. 567–579). Amsterdam: Elsevier.

Rolls, E.T. (1996c). The orbitofrontal cortex. *Philosophical Transactions of the Royal Society, Series B, 351,* 1433–1444.

Rolls, E. T. (1996d). The representation of space in the primate hippocampus, and its relation to memory. In K. Ishikawa, J. L. McGaugh & H. Sakata (Eds.), *Brain Processes and Memory* (pp. 203–227). Amsterdam: Elsevier.

Rolls, E. T. (1997). A neurophysiological and computational approach to the functions of the temporal lobe cortical visual areas in invariant object recognition. In L. Harris & M. Jenkin (Eds.), *Computational and Psychophysical Mechanisms of Visual Coding.* Cambridge: Cambridge University Press.

Rolls, E. T., & Johnstone, S. (1992). Neurophysiological analysis of striatal function. In G. Vallar, S. F. Cappa, & C. W. Wallesch (Eds.), *Neuropsychological Disorders Associated with Subcortical Lesions* (pp. 61–97). Oxford: Oxford University Press.

Rolls, E.T., & Tovee, M.J. (1994). Processing speed in the cerebral cortex, and the neurophysiology of visual masking. *Proceedings of the Royal Society, Series B, 257,* 9–15.

Rolls, E.T., & Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73,* 713–726.

Rolls, E. T., & Treves, A. (1997). *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Rolls, E. T., Booth, M. C. A., & Treves, A. (1996). View-invariant representations of objects in the inferior temporal visual cortex. *Society for Neuroscience Abstracts, 22,* 760.5.

Rolls, E.T., Hornak, J., Wade, D., & McGrath, J. (1994a). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry, 57,* 1518–1524.

Rolls, E.T., Tovee, M.J., Purcell, D.G., Stewart, A.L., & Azzopardi, P. (1994b). The responses of neurons in the temporal cortex of primates, and face identification and detection. *Experimental Brain Research, 101,* 474–484.

Rolls, E. T., Tovee, M., Treves, A., & Panzeri, S. (1997a). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience,* in press.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997b). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research, 114,* 149–162.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49,* 329–359.

Rosenthal, D. (1990). *A Theory of Consciousness* (ZIF Rep. 40). Bielefeld, Germany: Zentrum fur Interdisziplinaire Forschung.

Rosenthal, D. M. (1993). Thinking that one thinks. In M. Davies & G. W. Humphreys (Eds.), *Consciousness* (pp. 197–223). Oxford: Blackwell.

Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society, Series B, 351,* 1405–1411.

Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychological Review, 99,* 195–231.

Tovee, M. J., & Rolls, E. T. (1992). Oscillatory activity is not evident in the primate temporal visual cortex with static stimuli. *Neuroreport, 3,* 369–372.

Tovee, M. J., & Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cognition, 2,* 35–58.

Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology, 70,* 640–654.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus, 2,* 189–199.

Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus, 4,* 374–391.

von der Malsburg, C. (1990). A neural architecture for the representation of scenes. In J. L. McGaugh, N. M. Weinberger and G. Lynch (Eds.), *Brain Organization and Memory: Cells, Systems and Circuits* (pp. 356–372). New York: Oxford University Press.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

Wallis, G., Rolls, E. T., & Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *International Joint Conference on Neural Networks, 2,* 1087–1090