EDMUND T. ROLLS

REPRESENTATIONS IN THE BRAIN

ABSTRACT. The representation of objects and faces by neurons in the temporal lobe visual cortical areas of primates has the property that the neurons encode relatively independent information in their firing rates. This means that the number of stimuli that can be encoded increases exponentially with the number of neurons in an ensemble. Moreover, the information can be read by receiving neurons that perform just a synaptically weighted sum of the firing rates being received. Some ways in which these representations become grounded in the world are described. The issue of syntactic binding in representations, and of its value for a higher order thought system, is discussed.

1. THE ENCODING OF INFORMATION IN THE BRAIN

How is information encoded in the cerebral cortex? Can we read the code being used by the cortex? What are the advantages of the encoding scheme used for the computations being performed by neurons in different areas of the cortex? These are some of the key issues considered here. Because information is exchanged between the computing elements of the cortex, the neurons, by their spiking activity, the appropriate level of analysis is how single neurons, and populations of single neurons, encode information in their firing. More global measures which reflect the averaged activity of large numbers of neurons (for example PET (positron emission tomography) and fMRI (functional magnetic resonance imaging), EEG (electroencephalographic recording), and ERPs (event-related potentials)) cannot reveal how the information is represented, or *how* the computation is being performed.

Some of the types of representation that might be found at the neuronal level are as follows (see Rolls and Treves 1998). A **local** representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of one of the neurons. This is sometimes called a grandmother cell representation, because in a famous example, a single neuron might be active only if one's grandmother was being seen (see Barlow 1995). A **fully distributed** representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of the full set of neurons. If the neurons are binary (for

Synthese **129:** 153–171, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

example either active or not), the most distributed encoding is when half the neurons are active for any one stimulus or event. A **sparse distributed** representation is a distributed representation in which a small proportion of the neurons is active at any one time.

1.1. *Sparse Distributed Representations of Visual Stimuli*

In the higher parts of the visual system in the temporal lobe visual cortical areas, there are neurons that are tuned to respond to faces (see Rolls 1992; Wallis and Rolls 1997) or to objects (Booth and Rolls 1998). (Both classes of neuron are described as being tuned to provide information about faces or objects, in that their responses can be view-invariant; see Rolls and Treves 1998, Chapter 8.) Neurons that respond to faces can regularly be found on tracks into the temporal cortical visual areas, and they therefore provide a useful set of cells for systematic studies about how information about a large set of different visual stimuli, in this case different faces, is represented (Rolls 1992). First, it has been shown that the representation of which particular object (face) is present is rather distributed. Baylis et al. (1985) showed this with the responses of temporal cortical neurons that typically responded to several members of a set of five faces, with each neuron having a different profile of responses to each face (see examples in Figure 1). It would be difficult for most of these single cells to tell which of even five faces, let alone which of hundreds of faces, had been seen. (At the same time, the neurons discriminated between the faces reliably, as shown by analyses of variance.)

In a more recent study, the responses of another set of temporal cortical neurons to 23 faces and 42 non-face natural images were measured, and again a distributed representation was found (Rolls and Tovee 1995a). The tuning was typically graded, with a range of different firing rates to the set of faces, and very little response to the non-face stimuli (see Figure 2). The spontaneous firing rate of the neuron in Figure 2 was 20 spikes/s, and the histogram bars indicate the change of firing rate from the spontaneous value produced by each stimulus. Stimuli which are faces are marked F, or P if they are in profile. B refers to images of scenes which included either a small face within the scene, sometimes as part of an image which included a whole person, or other body parts, such as hands (H) or legs. The non-face stimuli are unlabelled. The neuron responded best to three of the faces (profile views), had some response to some of the other faces, and had little or no response, and sometimes had a small decrease of firing rate below the spontaneous firing rate, to the non-face stimuli.

The implications of this sparse distributed representation of visual stimuli found in higher order visual cortical areas are considered below.

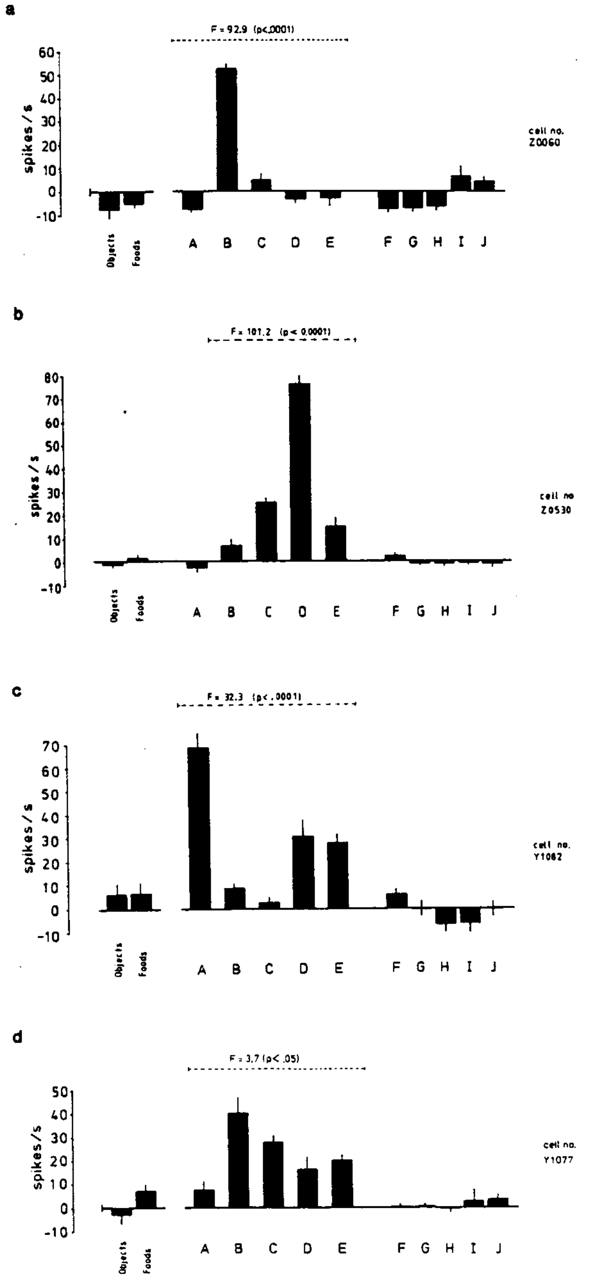

Figure 1. Responses of four different temporal cortex visual neurons to a set of five faces (A–E), and for comparison to a wide range of non-face objects and foods. F–J are non-face stimuli. The means and standard errors of the responses computed over 8–10 trials are shown. (After Baylis et al. 1985; and Rolls and Treves, 1998, Fig. 10.10.)

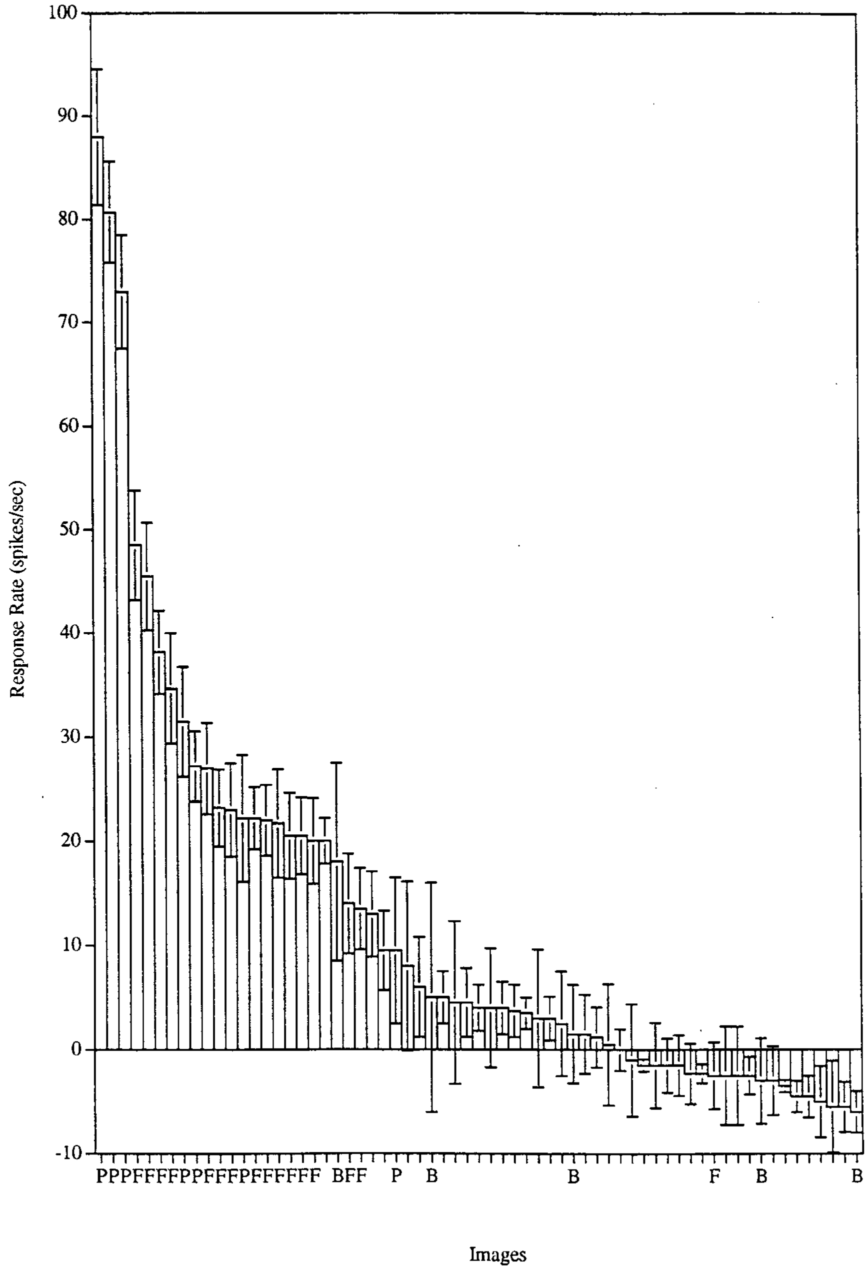

Figure 2. The firing rates of a temporal visual cortex neuron to a set of 23 face stimuli and 42 non-face stimuli. Neuron am242. The firing rate of the neuron is shown on the ordinate, the spontaneous firing rate of the neuron was 20 spikes/s, and the histogram bars are drawn to show changes of firing rate from the spontaneous rate (i.e. neuronal responses) produced by each stimulus. Stimuli which are faces are marked F, or P if they are in profile. B refers to images of scenes which included either a small face within the scene, sometimes as part of an image which included a whole person, or other body parts, such as legs; and H is used if hands were a prominent part of such images. The non-face stimuli are unlabelled. (After Rolls et al. 1997, Fig. 2a; and Rolls and Treves 1998, Fig. 10.11.)

1.2. *The Representation of Information in the Responses of Populations of Cortical Visual Neurons*

Quantitative evidence about the nature of the code used comes from applying information theory to analyse how information is represented by a population of these neurons. Figure 3 shows the responses of a typical single neuron in the inferior temporal cortex to four different faces. Peristimulus time histograms and rastergrams show the responses on different trials (originally in random order) of a face-selective neuron to four different faces. (In the rastergrams each vertical line represents one spike from the neuron, and each row is a separate trial.) To analyse how information is represented, we need to know what we would learn from any single trial taken from many such cells as that shown on Figure 4 about which stimulus was shown. Figure 4 shows that if we know the average firing rate of each cell in a population to each stimulus, then on any single trial we can guess the stimulus that was present by taking into account the response of all the cells. We can expect that the more cells in the sample, the more accurate may be the estimate of the stimulus. If the encoding was local, the number of stimuli encoded by a population of neurons would be expected to rise approximately linearly with the number of neurons in the population. In contrast, with distributed encoding, provided that the neuronal responses are sufficiently independent, and are sufficiently reliable (not too noisy), information from the ensemble would be expected to rise linearly with the number of cells in the ensemble, and (as information is a log measure) the number of stimuli encodable by the population of neurons might be expected to rise exponentially as the number of neurons in the sample of the population was increased.

The information available about which of 20 equiprobable faces had been shown that was available from the responses of different numbers of these neurons is shown in Figure 5 (Rolls et al. 1997a; Abbott et al. 1996). First, it is clear that some information is available from the responses of just one neuron: on average approximately 0.34 bits. Thus, knowing the activity of just one neuron in the population does provide some evidence about which stimulus was present. This evidence that information is available in the responses of individual neurons in this way, without having to know the state of all the other neurons in the population, indicates that information is made explicit in the firing of individual neurons in a way that will allow neurally plausible decoding, involving computing a sum of input activities each weighted by synaptic strength, to work (see Rolls and Treves 1998, Section 10.4.4.2). Second, it is clear (Figure 5) that the information rises approximately linearly, and the number of stimuli encoded thus rises approximately exponentially, as the number of cells in

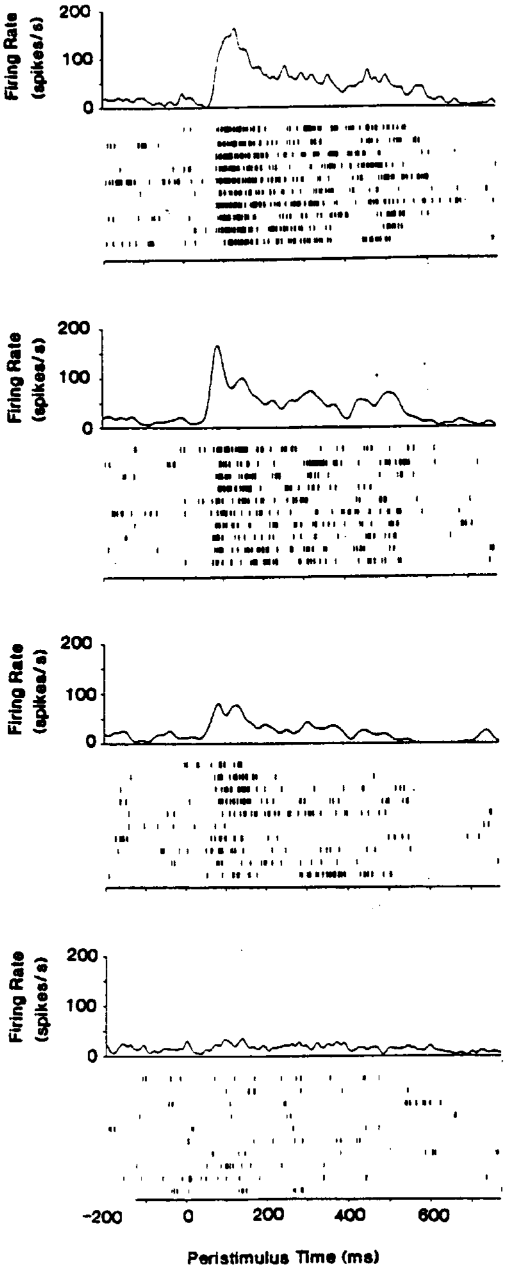

Figure 3. Peristimulus time histograms and rastergrams showing the responses on different trials (originally in random order) of a face-selective neuron to four different faces. Each set of rastergrams is for a different face. (In the rastergrams each vertical line represents one spike from the neuron, and each row is a separate trial.) (After Rolls and Treves 1998, Fig. 10.12.)

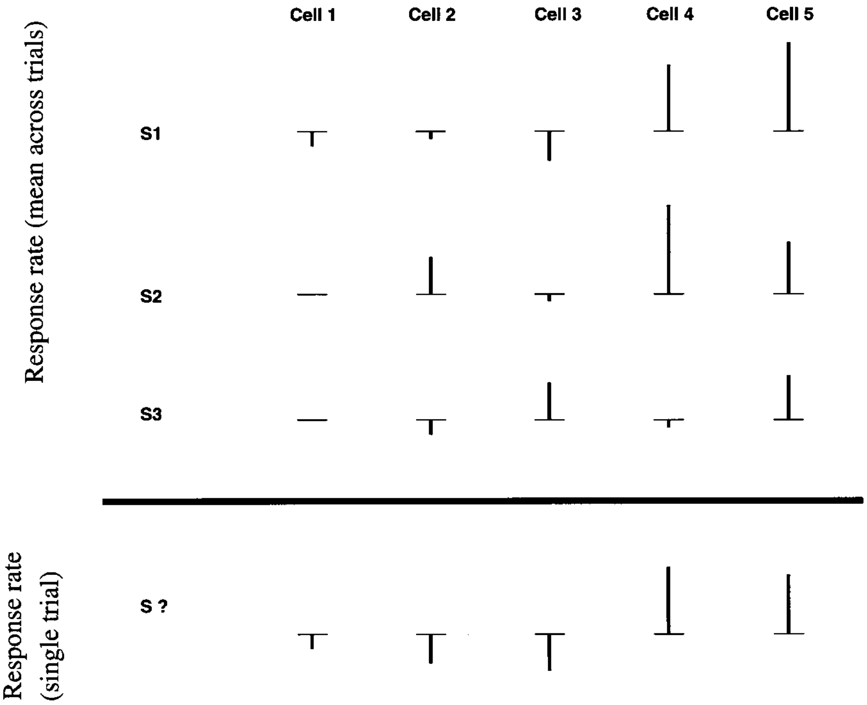
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
Response rate (mean across trials)
S1
S2
S3
Response rate (single trial)
S ?

Figure 4. This diagram shows the average response for each of several cells (Cell 1 etc.) to each of several stimuli (S1 etc.). The change of firing rate from the spontaneous rate is indicated by the vertical line above or below the horizontal line which represents the spontaneous rate. We can imagine guessing from such a table the stimulus S? that was present on any one trial (see text). (After Rolls and Treves 1998, Fig. 10.17.)

the sample increases (Rolls et al. 1997a). Although the data just described were from neurons recorded non-simultaneously, there is a preliminary indication that for simultaneously recorded pairs of temporal cortex visual neurons, the information they convey is relatively independent (Gawne and Richmond 1993).This has now been confirmed when recordings are made from up to four neurons recorded simultaneously, and moreover, almost all the information was available in the firing rates rather than the relative time of firing of different simultaneously recorded neurons (Panzeri et al. 1999).

This direct neurophysiological evidence thus demonstrates that the encoding is distributed, and the responses are sufficiently independent and reliable that the representational capacity increases exponentially. The consequence of this is that large numbers of stimuli, and fine discriminations between them, can be represented without having to measure the activity of an enormous number of neurons.

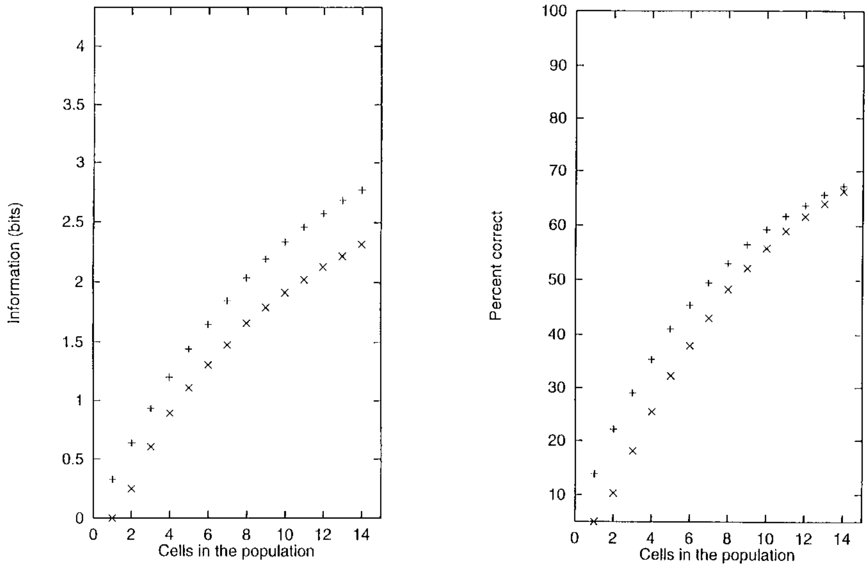

Figure 5. (a) The values for the average information available in the responses of different numbers of neurons in the temporal cortical visual areas on each trial in a 500 ms period, about which of a set of 20 face stimuli has been shown. The decoding method was Dot Product (×) or Probability Estimation (+). (b) The percentage correct for the corresponding data to those shown in Figure 5a. (After Fig. 4 of Rolls et al. 1997; Rolls and Treves 1998, Fig. 8.8.)

We believe that the same type of ensemble encoding of what stimulus is present (i.e. stimulus identity) is likely to be used in other sensory systems, and have evidence that this is the case for the primate taste and olfactory systems, in particular for the cortical taste and olfactory areas in the orbitofrontal cortex (Rolls et al. 1996). This type of ensemble encoding is also used in the primate hippocampus, in that the information about which spatial view is being seen rises approximately linearly with the number of hippocampal neurons in the sample (Rolls et al. 1998).

1.3. *Advantages of the Distributed Representations Found of Objects for Brain Processing*

Three key types of evidence that the visual representation provided by neurons in the temporal cortical areas, and the olfactory and taste representations in the orbitofrontal cortex, are distributed have been provided, and reviewed above. One is that the coding is not sparse (Baylis et al. 1985; Rolls and Tovee 1995). The second is that different neurons have different response profiles to a set of stimuli, and thus have at least partly independent responses (Baylis et al. 1985; Rolls and Tovee 1995a; Rolls et al. 1997a). The third is that the capacity of the representations rises exponentially with the number of neurons (Rolls et al. 1997a). The advantages of such distributed encoding are now considered, and apply to both fully distributed and to more sparse (but not to local) encoding schemes.

Exponentially High Coding Capacity

This property arises from a combination of the encoding being sufficiently close to independent by the different neurons (i.e. factorial), and sufficiently distributed. The independence or factorial requirement is simply to ensure that the information $I(n)$ from the population of neurons rises linearly with the number of neurons in the population. We note that if local encoding were used, the information would increase in proportion to the logarithm of the number of cells, which is not what has been found.

Part of the biological significance of such exponential encoding capacity is that a receiving neuron or neurons can obtain information about which one of a very large number of stimuli is present by receiving the rate of firing of relatively small numbers of inputs from each of the neuronal populations from which it receives. In particular, if neurons received from something in the order of 100 inputs from the population described here, they would have a great deal of information about which stimulus was in the environment. In particular, the characteristics of the actual visual cells described here indicate that the activity of 15 would be able to encode 192 face stimuli (at 50% accuracy), of 20 neurons 768 stimuli, of 25 neurons 3072 stimuli, of 30 neurons 12288 stimuli, and of 35 neurons 49152 stimuli (Abbott et al. 1996; the values are for the optimal decoding case). Given that most neurons receive a limited number of synaptic contacts, in the order of several thousand, this type of encoding is ideal. It would enable, for example, neurons in the amygdala and orbitofrontal cortex to form pattern associations of visual stimuli with reinforcers such as the taste of food when each neuron received a reasonable number, perhaps in the order of hundreds, of randomly assigned inputs from the visually responsive neurons in the temporal cortical visual areas which specify which visual stimulus or object is being seen (see Rolls and Treves 1998). Such a representation would also be appropriate for interfacing to the hippocampus, to allow an episodic memory to be formed, for example that a particular visual object was seen in a particular place in the environment (Rolls and Treves 1998).

One of the underlying themes here is the neural representation of objects. How would one know that one has found a neuronal representation of objects in the brain? The criterion we suggest that arises from this research is that when one can identify the object or stimulus that is present (from a large set of stimuli, perhaps thousands or more) with a realistic number of neurons, say in the order of 100, then one has a representation of the object. This criterion appears to imply exponential encoding, for only then could such a large number of stimuli be represented with a relatively small number of units, at least for units with the response characteristics of actual neurons. (In artificial systems a few multilevel errorless units could represent a much larger set of objects, but in a non-neural-like way.) Equivalently, we can say that there is a representation of the object when the information required to specify which of many stimuli or objects is present can be decoded from the responses of a limited number of neurons.

We may note at this point that an additional criterion for an object representation is that the representation of the stimulus or object readable from the ensemble of neurons should show at least reasonable invariance with respect to a number of transforms which do not affect the identity of the object. In the case of the visual representation these invariances include translation (shift), size, and even view invariance. These are transforms to which the responses of some neurons in the temporal cortical visual areas are robust or invariant (see Wallis and Rolls 1997; Tovee et al. 1994; Booth and Rolls 1998; Rolls 1999a). To complete the example, we can make it clear that although information about visual stimuli passes through the optic nerve from the retina, the representation at this level of the visual system is not of objects, for no decoding of a small set of neurons in the optic nerve would provide information in an invariant way about which of many objects was present on the retina.

One question which arises as a result of this demonstration of exponential encoding capacity is that of why there are so many neurons in the temporal cortical visual areas, if so few can provide such a high capacity representation of stimuli. One answer to this question is that high information capacity is needed for fine discrimination. Another point is that the 14 cells analysed to provide the data shown in Figure 4, or the 38 olfactory cells analysed by Rolls et al. (1996), were a selected subset of the cells in the relevant brain regions. The subsets were selected on the basis of the cells individually providing significant information about the stimuli in the set of visual, olfactory, or taste stimuli presented. If a random sample of say temporal cortical visual neurons had been taken, then that sample would have needed to be one to two orders of magnitude larger to include the subset of neurons in the sample. It is likely that the ensemble of neurons that projects to any particular cell in a receiving area is closer to a random sample than to our selected sample.

Ease with which the Code Can Be Read by Receiving Neurons: The Compactness of the Distributed Representation

For brain plausibility, it would also be a requirement that the decoding process should itself not demand more than neurons are likely to be able to perform. This is why when we have estimated the information from populations of neurons, we have used in addition to a probability estimation (optimal, in the Bayesian sense) method, a dot product measure, which is a way of specifying that all that is required of decoding neurons would be the property of adding up postsynaptic potentials produced through each synapse as a result of the activity of each incoming axon (Rolls et al. 1997a). More formally, the way in which the activation h of a neuron would be produced is by the following principle:

$$h = \sum_j r'_j w_j$$

where r'_j is the firing of the jth axon, and w_j is the strength of its synapse. The firing r of the neuron is a function of the activation

$$r = f(h).$$

This activation function f may be linear, sigmoid, binary threshold, etc.

It was found that with such a neurally plausible algorithm (the dot product, DP, algorithm), which calculates which average response vector the neuronal response vector on a single test trial was closest to by performing a normalized dot product (equivalent to measuring the angle between the test and the average vector), the same generic results were obtained, with only at most a 40% reduction of information compared to the more efficient (optimal) algorithm. This is an indication that the brain could utilize the exponentially increasing capacity for encoding stimuli as the number of neurons in the population increases. For example, by using the representation provided by the neurons described here as the input to an associative or autoassociative memory, which computes effectively the dot product on each neuron between the input vector and the synaptic weight vector, most of the information available would in fact be extracted (see Rolls and Treves 1998).

Higher Resistance to Noise

This, like the next few properties, is in general an advantage of distributed over local representations, which applies to artificial systems as well, but is presumably of particular value in biological systems in which some of the elements have an intrinsic variability in their operation. Because the decoding of a distributed representation involves assessing the activity of a whole population of neurons, and computing a dot product or correlation, a distributed representation provides more resistance to variation in individual components than does a local encoding scheme.

Generalization

Generalization to similar stimuli is again a property that arises in neuronal networks if distributed but not if local encoding is used, and when dot product operation is involved (e.g. see Rolls and Treves 1998, Chapters 2 and 3). The distributed encoding found in the cerebral cortex allows this generalization to occur.

Completion

Completion occurs in associative memory networks by a similar process (e.g. see Rolls and Treves 1998, Chapters 2 and 3). The distributed encoding found in the cerebral cortex is appropriate for allowing such completion to occur.

Graceful Degradation or Fault Tolerance

This also arises only if the input patterns have distributed representations, and not if they are local (e.g. see Rolls and Treves 1998, Chapters 2 and 3). The distributed encoding found in the cerebral cortex is appropriate for allowing such graceful degradation including tolerance to missing connections to occur.

Speed of Readout of the Information

The information available in a distributed representation can be decoded by an analyzer more quickly than can the information from a local representation, given comparable firing rates. Within a fraction of an interspike interval, with a distributed representation, much information can be extracted (Rolls et al. 1997a; Rolls and Treves 1998, Appendix A2).

2. CONTENT AND MEANING IN REPRESENTATIONS: HOW ARE REPRESENTATIONS GROUNDED IN THE WORLD?

The research just described shows that the firing of populations of neurons encodes information about stimuli in the world (see Rolls and Treves 1998). For example, from the firing rates of small numbers of neurons in the primate inferior temporal visual cortex, it is possible to know which of 20 faces has been shown to the monkey (Abbott et al. 1996; Rolls et al. 1997a). Similarly, a population of neurons in the anterior part of the macaque temporal lobe visual cortex has been discovered that has a view invariant representation of objects (Booth and Rolls 1998). From the firing of a small ensemble of neurons in the olfactory part of the orbitofrontal cortex, it is possible to know which of 8 odours was presented (Rolls et al. 1996). From the firing of small ensembles of neurons in the hippocampus, it is possible to know where in allocentric space a monkey is looking (Rolls et al. 1998; Rolls 1999b). In each of these cases, the number of stimuli that is encoded increases exponentially with the number of neurons in the ensemble, so this is a very powerful representation (Rolls and Treves 1998). What is being measured in each example is the mutual information between the firing of an ensemble of neurons and which stimuli are present in the world. In this sense, one can read off the code that is being used at the end of each of these sensory systems. However, what sense does the representation make to the animal? What does the firing of each ensemble of neurons "mean"? What is the content of the representation? In the visual system for example it is suggested that the representation is built by a series of appropriately connected competitive networks, operating with a modified Hebb learning rule (Rolls 1992; Wallis and Rolls 1997; Rolls and Treves 1998). Now competitive networks categorise their inputs without the use of a teacher. So which neurons fire to represent a particular object or stimulus is arbitrary. What meaning therefore does the particular ensemble that fires to an object have? How is the representation grounded in the real world? The fact that there is mutual information (see Rolls and Treves 1998, Appendix 2) between the firing of the ensemble of cells in the brain and a stimulus or event in the world does not answer this question.

One answer to this question is that there may be meaning in the case of objects and faces that it is an object or face, and not just a particular view. This is the case in that the representation may be activated by any view of the object or face. This is a step, suggested to be made possible by a short term memory in the learning rule which enable different views of objects to be associated together (see Rolls and Treves 1998). But it still does not provide the representation with any meaning in terms of the real world. What actions might one make, or what emotions might one feel, if that arbitrary set of temporal cortex visual cells was activated?

This leads to one of the answers I propose. I suggest that one type of meaning of representations in the brain is provided by their reward (or punishment) value: activation of these representations is the goal of actions. In the case of primary reinforcers such as the taste of food or pain, the activation of these representations would have meaning in the sense that the animal would work to obtain the activation of the taste of food cells when hungry, and to escape from stimuli that cause the cells representing pain to be activated. Evolution has built the brain is such a way that the animal will work to produce activation of the “taste of food reward” cells when hungry, and to escape from situations which cause the activation of the cells that respond to pain. In the case of other ensembles of cells in for example the visual cortex which respond to objects with the colour and shape of a banana, and which “represent” the sight of a banana in that their activation is always and uniquely produced by the sight of a banana, such representations come to have meaning only by association with a primary reinforcer, involving the process of stimulus-reinforcement association learning.

The second sense in which a representation may be said to have meaning is by virtue of sensory-motor correspondences in the world. For example, the touch of a solid object such as a table may become associated with evidence from the motor system that attempts to walk through the table result in cessation of movement. The representation of the table in the inferior temporal visual cortex may only have “meaning” in the sense that there is mutual information between the representation and the sight of the table until the table is seen just before and while it is touched, when sensory-sensory association between inputs from different sensory modalities will be set up that will enable the visual representation to become associated with its correspondences in the touch and movement worlds. In this second sense, meaning will be conferred on the visual sensory representation because of its associations in the sensory-motor world. Thus it is suggested that there are two ways by which sensory representations can be said to be grounded, that is to have meaning, in the real world.

It is suggested that the symbols used in language become grounded in the real world by the same two processes. In the first, a symbol such as the word banana has meaning because it is associated with primary reinforcers such as the flavour of the banana and with secondary reinforcers such as the sight of the banana. In the second process, the word “table” may have meaning because it is associated with sensory stimuli produced by tables such as their touch, shape and sight, as well as other functional properties, such as for example being load-bearing.

3. SYNTACTICAL OPERATIONS, SYMBOLIC REPRESENTATIONS, AND HIGHER ORDER THOUGHTS

In *The Brain and Emotion* (Rolls 1999a; see also Rolls 2000), I describe evidence that there are two main types of route to action for emotional events. One is for instrumental behavior to primary reinforcers or to stimuli associated with them, and involves working for immediate goals (including goals expected as a result of stimulus-reinforcement association learning). There is evidence that we (and split brain subjects and other patients, see Chapter 9) can perform many of these actions automatically, without conscious awareness, and this is therefore sometimes described as an implicit system. The second type of route uses a "what... if" type of reasoning involving flexible ("on-line") syntactic operations on symbols and enables long-term planning for strategies to obtain goals, and immediate goals to be deferred. It is argued that there is a credit assignment problem in such a first order syntactic system in that if a plan does not result in the desired outcome, then which of the multiple steps in the plan leads to the error is not clear. To solve this problem, it is suggested that it would be useful to have a higher order thought system to enable each step of the plan to be thought about (evaluating for example the premises for each step of the plan), so that the plan could be corrected.

The symbols (or symbolic representations) in this system are symbols in the sense that they can take part in syntactic processing. The symbolic representations are grounded in the world in that they refer to events in the world. The symbolic representations must have a great deal of information about what is referred to in the world, including the quality and intensity of sensory events, emotional states, etc. The need for this is that the reasoning in the symbolic system must be about stimuli, events, and states, and remembered stimuli, events and states, and for the reasoning to be correct, all the information that can affect the reasoning must be represented in the symbolic system, including for example just how light or strong the touch was, etc. Indeed, it is pointed out in *The Brain and Emotion* (Rolls 1999a, 252–253) that it is no accident that the shape of the multidimensional phenomenal (sensory etc) space does map so clearly onto the space defined by neuronal activity in sensory systems, for if this were not the case, reasoning about the state of affairs in the world would not map onto the world, and would not be useful. Good examples of this close correspondence are found in the taste system, in which subjective space maps simply onto the multidimensional space represented by neuronal firing in primate cortical taste areas. In particular, if a three-dimensional space reflecting the distances between the representations of different tastes provided by macaque neurons in the cortical taste areas is constructed, then the distances between the subjective ratings by humans of different tastes are very similar (Yaxley et al. 1990; Smith-Swintowsky et al. 1991; Plata-Salaman et al. 1996). Similarly, the changes in human subjective ratings of the pleasantness of the taste, smell and sight of food parallel very closely the responses of neurons in the macaque orbitofrontal cortex (see *The Brain and Emotion*, Chapter 2). The representations in the first order linguistic processor that the HOLTs process include beliefs (for example "Food is available", or at least representations of this), and the HOLT system would then have available to it the concept of a thought (so that it could represent "I believe [or there is a belief] that food is available"). However, as summarised in the first paragraph of this section, representations of sensory processes and emotional states must be processed by the first order linguistic system, and HOLTs may be about these representations of sensory processes and emotional states capable of taking part in the syntactic operations of the first order linguistic processor. Such sensory and emotional information may reach the first order linguistic system from many parts of the brain, including those such as the orbitofrontal cortex and amygdala implicated in emotional states (see *The Brain and Emotion*, Rolls 1999a, Fig. 9.3 and p. 253). When the sensory information is about the identity of the taste, the inputs to the first order linguistic system must come from the primary taste cortex, in that the identity of taste, independent of its pleasantness (in that the representation is independent of hunger) must come from the primary taste cortex. In contrast, when the information that reaches the first order linguistic system is about the pleasantness of taste, it must come from the secondary taste cortex, in that there the representation of taste depends on hunger (see Fig. 9.3 of Rolls, 1999a).

Another issue is that of the type of syntax that is required. What is required for the first order linguistic symbol processing system is the ability to link together representations in multiple "if... then" steps, to form a flexible plan. The plan involves flexible linking in that one plan might be formulated now, and another one, using many of the same symbols or representations, might be formed two minutes later. (Such a system formally *requires* syntax to bind the symbols.) Thus no claim is made about human verbal language being required, and a number of non-human animals may be able to form this type of plan. The higher order thought system needs to be able to understand and correct the plans of the first order syntactic system, and for this reason itself needs to be able to process syntax, and in this sense is termed a higher order linguistic thought (HOLT) system.

4. SOME THOUGHTS ON CONSCIOUSNESS

It is suggested that it is difficult to imagine a higher order linguistic thought system thinking about its own thoughts grounded in the world without it feeling like something. That is, it is suggested that phenomenal experience ("what it feels like") arises as a property of the operation of such a higher order linguistic thought (HOLT) system (Rolls 1997, 1999a, 2000;

cf. Rosenthal 1993). (This type of system is sometimes described as the explicit system.) Having identified a computational advantage for a system to have thoughts about thoughts, and suggested that phenomenal experience arises by virtue of this system, I note that sometimes this system must include sensory processes, emotional states etc. in its operations, and suggest that such sensory events, and emotional and motivational processes, feel like something by virtue of participating in this system.

An interesting issue is whether higher order thoughts (HOTs) are involved when we are conscious about stimuli and events that can be processed implicitly, for example secondary reinforcers. My hypothesis is that the HOLT system for explicit linguistic planning should usually be monitoring our behavior when it is being performed implicitly and automatically, in case the different types of computation made possible by the syntactic planning system would result in a better outcome by deferring an immediate goal and acting for a goal achievable only by multistep planning (see Rolls 1999a, Chapters 9 and 10); and that it is by virtue of the operation of the HOLT system that we are conscious of the secondary reinforcer. That is, some behavioral responses to the secondary reinforcer may be learned about in an implicit system (one to which there is no conscious access), but there may nevertheless be explicit access to the stimuli involved because they reach a HOLT linguistic system that is continually monitoring. This may lead to the explicit system confabulating sometimes about causes or reasons for actions, as described in *The Brain and Emotion* (Rolls 1999a).

For those who might wonder whether it is proposed that human verbal language is necessary for qualia and feelings, it should be clear that this is not implied by the proposal. In any case, theories of consciousness are not sufficiently developed that they should be taken to have practical implications.

REFERENCES

Abbott, L. F., E. T. Rolls, and M. J. Tovee: 1996, 'Representational Capacity of Face Coding in Monkeys', *Cerebral Cortex* **6**, 498–505.

Barlow, H. B.: 1972, 'Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology', *Perception* **1**, 371–394.

Baylis, G. C., E. T. Rolls, and C. M. Leonard: 1985, 'Selectivity Between Faces in the Responses of a Population of Neurons in the Cortex in the Superior Temporal Sulcus of the Monkey', *Brain Research* **342**, 91–102.

Booth, M. C. A. and E. T. Rolls: 1998, 'View-Invariant Representations of Familiar Objects by Neurons in the Inferior Temporal Visual Cortex', *Cerebral Cortex* **8**, 510–523.

Gawne, T. J. and B. J. Richmond: 1993, 'How Independent Are the Messages Carried by Adjacent Inferior Temporal Cortical Neurons?', *Journal of Neuroscience* **13**, 2758–2771.

Panzeri, S., S. R. Schultz, A. Treves, and E. T. Rolls: 1999, 'Correlations and the Encoding of Information in the Nervous System', *Proceedings of the Royal Society* B **266**, 1001–1012.

Plata-Salaman, C. R., V. L. Smith-Swintosky, and T. R. Scott: 1996, 'Gustatory Neural Coding in the Monkey Cortex: Mixtures', *Journal of Neurophysiology* **75**, 2369–2379.

Parga, N. and E. T. Rolls: 1998, 'Transform Invariant Recognition by Association in a Recurrent Network', *Neural Computation* **10**, 1507–1525.

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.: 1997, 'Consciousness in Neural Networks?', *Neural Networks* **10**, 1227–1240.

Rolls, E. T.: 1999a, *The Brain and Emotion*, Oxford University Press, Oxford.

Rolls, E. T.: 1999b, 'Spatial View Cells and the Representation of Place in the Primate Hippocampus', *Hippocampus* **9**, 467–480.

Rolls, E. T.: 2000, 'Précis of The Brain and Emotion', *Behavioral and Brain Sciences* **23**, 177–233.

Rolls, E. T. and M. J. Tovee: 1995, 'Sparseness of the Neuronal Representation of Stimuli in the Primate Temporal Visual Cortex', *Journal of Neurophysiology* **73**, 713–726.

Rolls, E. T., H. D. Critchley, and A. Treves: 1996, 'The Representation of Olfactory Information in the Primate Orbitofrontal Cortex', *Journal of Neurophysiology* **75**, 1982–1996.

Rolls, E. T., A. Treves, and M. J. Tovee: 1997a, '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.

Rolls, E. T., A. Trevee, M. Tovee, and S. Panzeri: 1997b, 'Information in the Neuronal Representation of Individual Stimuli in the Primate Temporal Visual Cortex', *Journal of Computational Neuroscience* **4**, 309–333.

Rolls, E. T. and A. Treves: 1998, *Neural Networks and Brain Function*, Oxford University Press, Oxford.

Rolls, E. T., A. Treves, R. G. Robertson, P. Georges-François, and S. Panzeri: 1998, 'Information about Spatial View in an Ensemble of Primate Hippocampal Cells', *Journal of Neurophysiology* **79**, 1797–1813.

Rolls, E. T. and T. Milward: 2000, 'A Model of Invariant Object Recognition in the Visual System: Learning Rules, Activation Functions, Lateral Inhibition, and Information-Based Performance Measures', *Neural Computation* **12**, 2547–2572.

Rolls, E. T. and S. M. Stringer: 2001, 'Invariant Object Recognition in the Visual System with Error Correction and Temporal Difference Learning', *Network*, in press.

Rosenthal, D. M.: 1993, 'Thinking that one Thinks', in M. Davies and G. W. Humphreys (eds), *Consciousness*, Chapter 10, Blackwell, Oxford, pp. 197–223.

Tovee, M. J., E. T. Rolls, and P. Azzopardi: 1994, 'Translation Invariance and the Responses of Neurons in the Temporal Visual Cortical Areas of Primates', *Journal of Neurophysiology* **72**, 1049–1060.

Smith-Swintosky, V. L., C. R. Plata-Salaman, and T. R. Scott: 1991, 'Gustatory Neural Coding in the Monkey Cortex: Stimulus Quality', *Journal of Neurophysiology* **66**, 1156–1165.

Wallis, G. and E. T. Rolls: 1997, 'Invariant Face and Object Recognition in the Visual System', *Progress in Neurobiology* **51**, 167–194.

Yaxley, S., E. T. Rolls, and Z. J. Sienkiewicz: 1990, ‘Gustatory Responses of Single Neurons in the Insula of the Macaque Monkey’, *Journal of Neurophysiology* **63**, 689–700.

Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford OX1 3UD
U.K.
Website: www.cns.ox.ac.uk