On the Relation between the Mind and the Brain: A Neuroscience Perspective

Edmund T. Rolls[*]

Oxford Centre for Computational Neuroscience, Oxford (UK)

Résumé : Dans cet article, je montre que les neurosciences computationnelles fournissent une nouvelle approche pertinente à des problèmes traditionnels en philosophie tels que la relation entre les états mentaux et cérébraux (le problème esprit–corps ou corps–esprit), le déterminisme et le libre arbitre, et peut nous aider à traiter le problème « difficile » des aspects phénoménaux de la conscience. Un des thèmes de cet article et de mon livre *Neuroculture: on the Implications of Brain Science* ([Rolls 2012c]) est qu'en comprenant les calculs effectués par les neurones et les réseaux neuronaux, et les effets du bruit dans le cerveau sur ceux-ci, nous gagnerons une vraie compréhension des mécanismes qui sous-tendent le fonctionnement du cerveau. Une partie de notre solution au problème esprit–corps est que l'esprit et le cerveau sont différents niveaux d'explication du traitement de l'information, leur relation pouvant être appréhendée par la compréhension des mécanismes en jeu à l'aide de l'approche fournie par les neurosciences computationnelles. Mais cette solution ne traite pas certains problèmes « difficiles » tels que le problème de la conscience phénoménale, et, même si j'ai fourni de nouvelles suggestions sur ce point dans cet article, il faut reconnaître qu'il y a toujours une brèche dans notre compréhension entre les événements dans le cerveau et les expériences subjectives qui peuvent les accompagner. L'explication que je propose est que, lorsque cela « *fait quelque chose* », il ne s'agit que d'une propriété d'un processus computationnel qui a des pensées sur ses propres pensées (pensées d'ordre supérieur), les pensées étant ancrées dans le monde.

Abstract: In this paper I show that computational neuroscience provides an important new approach to traditional problems in philosophy such as the relation between mental states and brain states (the mind-body or mind-brain problem), to determinism and free will, and helps one with the 'hard' problem, the phenomenal aspects of consciousness.

∗. www.oxcns.org.

Philosophia Scientiæ, 17 (2), 2013, 31–70.

One of the themes of the paper and of my book *Neuroculture: on the Implications of Brain Science* ([Rolls 2012c]) is that by understanding the computations performed by neurons and neuronal networks, and the effects of noise in the brain on these, we will gain a true understanding of the mechanisms that underlie brain function. Part of the solution proposed to the mind-body problem is that the mind and the brain are different levels of explanation of information processing, the correspondence between which can be understood by understanding the mechanisms involved using the approach of computational neuroscience.

But this does leave some 'hard' problems, such as the problem of phenomenal consciousness, and while I have provided new suggestions about this in this paper, one must recognise that there is still somewhat of a gap in our understanding of events in the brain and the subjective experiences that may accompany them. The explanation I offer is that when it 'feels like something' this is just a property of a computational process that has thoughts about its own thoughts (higher order thoughts), and with the thoughts grounded in the world.

1 Introduction

We consider here a neuroscience-based approach to the following issues. What is the relation between the mind and the brain? Do mental, mind, events cause brain events? Do brain events cause mental effects? What can we learn from the relation between software and hardware in a computer about mind–brain interactions and how causality operates? The hard problem of consciousness: why does some mental processing feel like something, and other mental processing does not? What type of processing is occurring when it does feel like something? Is consciousness an epiphenomenon, or is it useful? Are we conscious of the action at the time it starts, or later? How is the world represented in the brain?

2 The mind–brain problem

The relation between the mind and the brain is the mind–brain or mind–body problem. Do mental, mind, events cause brain events? Do brain events cause mental effects? What can we learn from the relation between software and hardware in a computer about mind–brain interactions and how causality operates? Neuroscience shows that there is a close relation between mind and matter (captured by the following inverted saying: 'Never matter, no mind').

My view is that the relationship between mental events and neurophysiological events is similar (apart from the problem of consciousness) to the relationship between the program running in a computer and the hardware of

the computer. In a sense, the program (the software loaded onto the computer usually written in a high-level language such as C or Matlab) causes the logic gates (TTL, transistor-transistor logic) of the hardware to move to the next state. This hardware state change causes the program to move to its next step or 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). This is the solution I propose to this aspect of the mind–body (or mind–brain) problem.

There are alternative ways of treating the mind–brain issue. Another is to consider the process as a mechanism with different levels of explanation, in the following way. We can now understand brain processing from the level of ion channels in neurons, through neuronal biophysics, to neuronal firing, through the computations performed by populations of neurons, and how their activity is reflected by functional neuroimaging, to behavioural and cognitive effects [Rolls 2008b, Rolls & Deco 2010, Rolls 2012c]. Activity at any one level can be used to understand activity at the next. This raises the philosophical issue of how we should consider causality with these different levels. Does the brain cause effects in the mind, or do events at the mental, mind, level influence brain activity? Here the analogy with a computer described in the previous paragraph is helpful. The view we have of the relation between a computer program and its implementation on the computer hardware provides a foundation for understanding the relation between the mind and the brain. Of course brain computation and computation in a digital computer are implemented in different ways, which are fascinating to understand (see [Rolls 2012c, section 2.15]), but that does not alter the point.

Overall, understanding brain activity at these different levels provides a unifying approach to understanding brain function, which is proving to be so powerful that the fundamental operations involved in many aspects of brain function can be understood in principle, though with of course many details still to be discovered. These functions include many aspects of perception including visual face and object recognition, and taste, olfactory and related processing; short-term memory; long-term memory; attention; emotion; and decision-making [Rolls 2008b], [Rolls & Deco 2010], [Rolls 2012c, 2014, 2010a, 2012b]. Predictions made at one level can be tested at another. Conceptually this is an enormous advance. But it is also of great practical importance, in medicine. For example, we now have new ways of predicting effects of possible pharmacological treatments for brain diseases by a developing understanding of how drugs affect synaptic receptors, which in turn affect neuronal activity, which in turn affect the stability of the whole network of neurons and hence cognitive symptoms such as attention vs. distractibility (see [Rolls 2012c, chap. 10], and [Rolls 2012a]). Perhaps the great computational unknown at present is how syntax for language is implemented in the brain.

The whole processing in the brain can now be specified in principle from the mechanistic level of neuronal firings, etc., up through the computational

level to the cognitive and behavioural level. Sometimes the cognitive effects seem remarkable, for example the recall of a whole memory from a part of it, and we describe this as an 'emergent property', but once understood from the mechanistic level upwards, the functions implemented are elegant and wonderful, but understandable and not magical or poorly understood [Rolls 2008b], [Rolls & Deco 2010], [Rolls 2012c]. Different philosophers may choose or not to say that causality operates between these different levels of explanation, but the point I make is that however they speak about causality in such a mechanistic system with interesting 'emergent' computational properties, the system is now well-defined, is no longer mysterious or magical, and we have now from a combination of neuroscience and analyses of the type used in theoretical physics a clear understanding of the properties of neural systems and how cognition emerges from neural mechanisms. There are of course particular problems that remain to be resolved with this approach, such as that of how language is implemented in the brain, but my point is that this mechanistic approach, supported by parsimony, appears to be capable of leading us to a full understanding of brain function, cognition, and behaviour.

A possible exception where a complete explanation may not emerge from the mechanistic approach is phenomenal consciousness, which is treated next.

Before embarking on a consideration of consciousness, I note that much behaviour can be performed without apparently being conscious [Rolls 2003], [Brooks, Savov, Allzen *et al.* 2012], [Prabhakaran & Gray 2012] (an example of which might be driving a car for a short distance), and that conscious processing may actually interfere with some motor skill non-conscious operations of the brain, such as a golfer's swing. Much of the information processing of the brain can be understood in terms of computations without having to consider consciousness. The representations by the firing of populations of neurons in the brain of events in the world (such as visual, taste, and olfactory stimuli) do provide accurate representations of those events, as of course they need to in order to be useful. The code is based in large part on the changes in the firing rates of neurons that are produced in specialized brain areas by these stimuli [Rolls 2008b], [Rolls & Treves 2011]. These representations by neurons not only reflect information in the world, but also our subjective (i.e., phenomenal, what it feels like) experience [Rolls 2005], [Kadohisa, Rolls & Verhagen 2005], [Rolls & Grabenhorst 2008], [Grabenhorst & Rolls 2011]. Again, that must be the case if the conscious processing is to be useful in dealing with events in the world. With that setting of the scene, I now turn to consider phenomenal consciousness, and later in the paper issues such as determinism, and free will.

3 Consciousness

3.1 Introduction

It might be possible to build a computer that would perform the functions of emotions described elsewhere [Rolls 2005, 2012c, 2014], and yet we might not want to ascribe emotional feelings to the computer. We might even build the computer with some of the main processing stages present in the brain, and implemented using neural networks that simulate the operation of the real neural networks in the brain (see [Rolls & Treves 1998], [Rolls & Deco 2002], and [Rolls 2008b]), yet we might not still wish to ascribe emotional feelings to this computer. This point often arises in discussions with undergraduates, who may say that they follow the types of point made about emotion [Rolls 2005, 2012c], yet believe that almost the most important aspect of emotions, the feelings, have not been accounted for, nor their neural basis described. In a sense, the functions of reward and punishment in emotional behaviour have been described [Rolls 2005, 2012c], but what about the subjective aspects of emotion, what about the pleasure?

A similar point also arises when parts of the taste, olfactory, and visual systems in which the reward value of the taste, smell, and sight of food is represented are described [Rolls 2005, 2012c]. Although the neuronal representation in the orbitofrontal cortex is clearly related to the reward value of food, and in humans the activations found with functional neuroimaging are directly correlated with the reported subjective pleasantness of the stimuli, is this where the pleasantness (the subjective hedonic aspect) of the taste, smell, and sight of food is represented and produced? Again, we could (in principle at least) build a computer with neural networks to simulate each of the processing stages for the taste, smell, and sight of food [Rolls 2005, 2008b], and yet would probably not wish to ascribe feelings of subjective pleasantness to the system we have simulated on the computer.

What is it about neural processing that makes it feel like something when some types of information processing are taking place? It is clearly not a general property of processing in neural networks, for there is much processing, for example that in the autonomic nervous system concerned with the control of our blood pressure and heart rate, of which we are not aware. Is it then that awareness arises when a certain type of information processing is being performed? If so, what type of information processing? And how do emotional feelings, and sensory events, come to feel like anything? These 'feels' are called qualia. These are great mysteries that have puzzled philosophers for centuries. They are at the heart of the problem of consciousness, for why it should feel like something at all is the great mystery, the 'hard' problem.

Other aspects of consciousness may be easier to analyse, such as the fact that often when we 'pay attention' to events in the world, we can process those events in some better way. These are referred to as 'process' or 'access' aspects of consciousness, as opposed to the 'phenomenal' or 'feeling' aspects of

consciousness referred to in the preceding paragraph [Block 1995a], [Chalmers 1996], [Allport 1988], [Koch 2004], [Block 1995b].

The puzzle of qualia, that is of the phenomenal aspect of consciousness, seems to be rather different from normal investigations in science, in that there is no agreement on criteria by which to assess whether we have made progress. So, although the aim of this section is to address the issue of consciousness, especially of qualia, what is written cannot be regarded as being as firmly scientific as most research relating to brain function [Rolls 2008b], [Rolls & Deco 2010]. For most brain research, there is good evidence for most of the points made, and there would be no hesitation or difficulty in adjusting the view of how things work as new evidence is obtained. However, in the work on qualia, the criteria are much less clear. Nevertheless, the reader may well find these issues interesting, because although not easily solvable, they are very important issues to consider if we wish to really say that we understand some of the very complex and interesting issues about brain function, and ourselves.

With these caveats in mind, I consider in this section the general issue of consciousness and its functions, and how feelings, and pleasure, come to occur as a result of the operation of our brains. A view on consciousness, influenced by contemporary cognitive neuroscience, is outlined next. 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. However, this view on consciousness that I describe is only preliminary, and theories of consciousness are likely to develop considerably. Partly for these reasons, this theory of consciousness, at least, should not be taken to have practical implications.

3.2 A theory of consciousness

3.2.1 Conscious and unconscious routes to action

A starting point is that many actions can be performed relatively automatically, without apparent conscious intervention [Rolls 2003], [Brooks, Savov, Allzen *et al.* 2012], [Prabhakaran & Gray 2012]. Such actions could involve control of behaviour by brain systems that 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 [Rolls 2005]). 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 2008b]. 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 that 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 (see [Rolls 2008b]). 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). 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 role of the language areas in the syntactic manipulation of semantic entities that may make long-term planning possible. A schematic diagram of this suggestion is provided in Fig. 1.

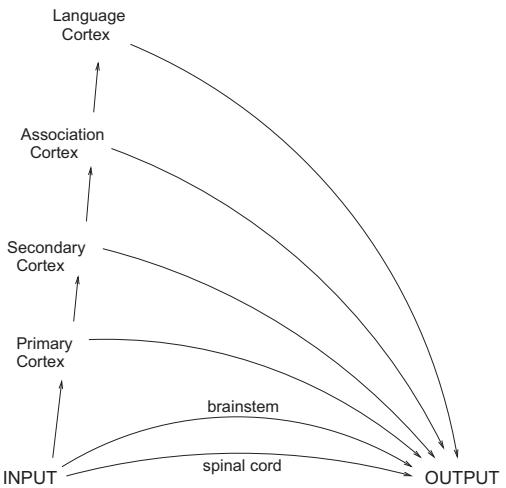

FIGURE 1: Schematic illustration indicating many possible routes from input systems to action (output) systems. Cortical information-processing systems are organized hierarchically, and there are routes to output systems from most levels of the hierarchy.

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 & LeDoux 1978], [Gazzaniga 1988, 1995]. 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, Hornak, Wade *et al.* 1994a], [Hornak, Bramham, Rolls *et al.* 2003]. 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 environmental event that actually triggered the action [Gazzaniga & LeDoux 1978], [Gazzaniga 1988, 1995].

It is accordingly 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 in Fig. 1). (It is of interest that mathematicians may get 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.)

3.2.2 Higher-order syntactic thoughts and consciousness

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 she thinks that..., involving 'theory of mind'), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others. (Second-order thoughts are thoughts about thoughts. Higher-order thoughts refer to second-order, third-order, etc., thoughts about thoughts...) This capability for higher-order thoughts would also enable 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.

(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 that reflect the evolution of an arbitrary symbol-manipulation system that could be useful for flexible planning, cf. [Cheney & Seyfarth 1990], [Byrne & Whiten 1988], and [Whiten & Byrne 1997].) (For an earlier view that is close to this part of the argument see [Humphrey 1980].)

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 that is the sense in which language is defined and used here. The type of syntactic processing need not be at the natural language level (which implies a universal grammar), but could be at the level of mentalese [Rolls 2005, 2004], [Fodor 1994], [Rolls 2011].

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 that arises in a system that can think about (or reflect on) her own (or other peoples') thoughts, that is in a system capable of second- or higher-order thoughts [Rosenthal 1986, 1990, 1993], [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]. Another position that is close in some respects to that of Carruthers and the present position is that of [Chalmers 1996], that awareness is something that has direct availability for behavioural control. This amounts effectively for him in humans to saying that consciousness is what we can report about verbally. 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 that is present when linguistic processing that involves second- or higher-order thoughts is being performed.

It might be objected that this hypothesis captures some of the process aspects of consciousness, that is, what is useful 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 her own first- or lower-order thoughts. But the return point is the following: if a human with second-order thoughts is thinking about its own first-order thoughts, surely it is very difficult for us to conceive that this would not feel like something? (Perhaps the higher-order thoughts in thinking about the first-order thoughts would need to have in doing this some sense of continuity or self, so that the first-order thoughts would be related to the same system that had thought of something else a few minutes ago. But even this continuity aspect may not be a requirement for consciousness. Humans with anterograde amnesia cannot remember what they felt a few minutes ago, yet their current state does feel like something.)

It is suggested that part of the evolutionary adaptive significance of this type of higher-order thought is that it enables 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 (see also [Squire, Stark & Clark 2004]). Its close relation to 'conscious' processing in humans (Squire has classified it as a declarative memory system) may be simply that it enables 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 understand better the thoughts of another individual in a similar situation, and therefore predict that individual's behaviour better ([Humphrey 1980], [Humphrey 1986], cf. [Barlow 1997]).

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 (which 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 that 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 (or in American usage, 'one-time', meaning used once) set of steps, as in a sentence or in a particular plan used just once, 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, and this type of cognitive processing is thought to be a primary basis for human skilled performance [Anderson 1996]. It is partly because complex lower-order thoughts such as these that 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 steps or links in the chain of reasoning at the lower-order level, and having detected the weak link or step, 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? [In multilayer neural networks, the credit assignment problem is that if errors are being specified at the output layer, the problem arises about how to propagate back the error to earlier, hidden, layers of the network to assign credit or blame to individual synaptic connection; see [Rumelhart, Hinton & Williams 1986] and [Rolls 2008b].] **My suggestion is that this solution to the credit assignment problem for a one-off syntactic plan 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 issue of symbol grounding is considered further in Section 3.3.

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, or the application of fixed rules by a supervisory part of a computer program.

3.2.3 Qualia

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 (given that the processing being performed by this system is inherently conscious, as suggested above). 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 system, 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, which 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 enable 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 sensory, emotional, and motivational qualia are thus accounted for by the evolution of a linguistic (i.e., syntactic) 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 (i.e., syntactic, not necessarily verbal) 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. (This reinforcement-guided learning could be implemented using for example stimulus–reinforcer association learning in the amygdala and orbitofrontal cortex followed by action-outcome learning in the cingulate cortex [Rolls 2005, 2009], [Grabenhorst & Rolls 2011]; or rule-following using habit or stimulus–response learning in the basal ganglia [Rolls 2005].) Such behaviours might appear very similar to human behaviour performed in similar circumstances, but need 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.

In order 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 [Rolls 1989, 1997b, 2005, 2013, 2014]. Now the quality of the tastant (sweet, salt, etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is reduced 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 that 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) [Rolls 2008b], [Grabenhorst & Rolls 2011].

The same argument applies 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 [Rolls 2005, 2014].

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 [Rolls 2005, 2014]. 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, e.g., 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 [Rolls 2008b], [Grabenhorst & Rolls 2011], to the system that has the higher-order thoughts, which as I have argued above, requires language (understood as syntactic manipulation of symbols). 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 [Rolls 1989, 2005, 2008b, 2013]. 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 that receive backprojections from the hippocampus [Rolls 2008b], 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 Fig. 2.

3.2.4 Consciousness and causality

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

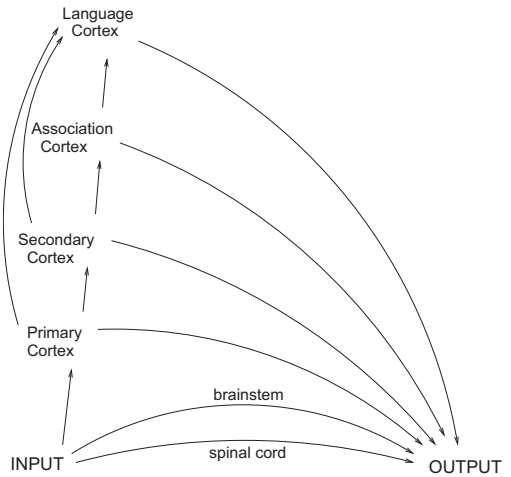

FIGURE 2: Schematic illustration indicating that early cortical stages in information processing may need access to language areas that bypass subsequent levels in the hierarchy, so that consciousness of what is represented in early cortical stages, and which may not be represented in later cortical 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 & Treves 1998], [Rolls & Deco 2002], [Treves & Rolls 1994], probably reciprocate all the connections shown.

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 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 that can later be recalled (see Fig. 1).

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 [Rolls 1990, 1995, 2005]. 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 that 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, negative consequences may not be as bad as feared.)

3.2.5 Consciousness and free will

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.

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.) The issue of whether the brain operates deterministically (Section 4) is not therefore I suggest the central or most interesting question about free will. Instead, analysis of which brain processing systems are engaged when we are taking decisions [Rolls & Deco 2010], [Deco, Rolls, Albantakis *et al.* 2012], and which processing systems are inextricably linked to feelings as suggested above, may be more revealing about free will.

3.2.6 Consciousness and self-identity

Before leaving these thoughts, it may be worth commenting on the feeling of continuing self-identity that is characteristic of humans. Why might this arise? One suggestion is that if one is an organism that can think about its own long-term multistep plans, then for those plans to be consistently and thus adaptively executed, the goals of the plans would need to remain stable, as would memories of how far one had proceeded along the execution path of each plan. If one felt each time one came to execute, perhaps on another day, the next step of a plan, that the goals were different, or if one did not remember which steps had already been taken in a multistep plan, the plan would never be usefully executed. So, given that it does feel like something to be doing this type of planning using higher-order thoughts, it would have to feel as if one were the same agent, acting towards the same goals, from day to day, for which autobiographical memory would be important.

Thus it is suggested that the feeling of continuing self-identity falls out of a situation in which there is an actor with consistent long-term goals, and

long-term recall. If it feels like anything to be the actor, according to the suggestions of the higher-order thought theory, then it should feel like the same thing from occasion to occasion to be the actor, and no special further construct is needed to account for self-identity. Humans without such a feeling of being the same person from day to day might be expected to have, for example, inconsistent goals from day to day, or a poor recall memory. It may be noted that the ability to recall previous steps in a plan, and bring them into the conscious, higher-order thought system, is an important prerequisite for long-term planning which involves checking each step in a multistep process.

Conscious feelings of self will be likely to be of value to the individual. Indeed, it would be maladaptive if feelings of self-identity, and continuation of the self, were not wanted by the individual, for that would lead to the brain's capacity for feelings about self-identity to leave the gene pool, due for example to suicide. This wish for feelings and thoughts about the self to continue may lead to the wish and hope that this will occur after death, and this may be important as a foundation for religions [Rolls 2012c].

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. However, as stated above, one does not feel that there are straightforward criteria in this philosophical field of enquiry for knowing whether the suggested theory is correct; so it is likely that theories of consciousness will continue to undergo rapid development; and current theories should not be taken to have practical implications.

3.3 Content and meaning in representations: How are representations grounded in the world?

In Section 3.2 I suggested that representations need to be grounded in the world for a system with higher-order thoughts to be conscious. I therefore now develop somewhat what I understand by representations being grounded in the world.

It is possible to analyse how the firing of populations of neurons encodes information about stimuli in the world [Rolls 2008b], [Rolls & Treves 2011]. 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, Rolls & Tovee 1996], [Rolls, Treves & Tovee 1997]. 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 & 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 eight odours was presented [Rolls, Critchley & Treves 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, Treves,

Robertson *et al.* 1998]. 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 [Abbott, Rolls & Tovee 1996], [Rolls, Treves & Tovee 1997], [Rolls & Treves 1998], [Rolls, Aggelopoulos, Franco *et al.* 2004], [Franco, Rolls, Aggelopoulos *et al.* 2004], [Aggelopoulos, Franco & Rolls 2005], [Rolls 2008b], [Rolls & Treves 2011]. 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, 1994], [Wallis & Rolls 1997], [Rolls 2000], [Rolls & Milward 2000], [Stringer & Rolls 2000], [Rolls & Stringer 2001], [Rolls & Deco 2002], [Elliffe, Rolls & Stringer 2002], [Stringer & Rolls 2002], [Deco & Rolls 2004], [Rolls 2008b, 2012b]. Now competitive networks categorize their inputs without the use of a teacher [Kohonen 1989], [Hertz, Krogh & Palmer 1991], [Rolls 2008b]. So which particular neurons fire as a result of the self-organization 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 between the firing of the ensemble of cells in the brain and a stimulus or event in the world [Rolls 2008b], [Rolls & Treves 2011] does not fully 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 that enables different views of objects to be associated together [Wallis & Rolls 1997], [Rolls & Milward 2000], [Rolls & Stringer 2001], [Rolls 2008b, 2012b]. 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 for 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 neurons when hungry, and to escape from stimuli that cause the neurons representing pain to be activated. Evolution has built the brain so that genes specify these primary reinforcing stimuli, and so that their representations in the brain should be the targets for actions [Rolls 2005, 2014]. In the case of other ensembles of

neurons in, for example, the visual cortex that 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–reinforcer 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 might 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 might have 'meaning' only 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. These reinforcers have 'meaning' to the animal in that evolution has built animals as machines designed to do everything that they can to obtain these reinforcers, so that they can eventually reproduce successfully and pass their genes onto the next generation. (The fact that some stimuli are reinforcers but may not be adaptive as goals for action is no objection. Genes are limited in number, and can not allow for every eventuality, such as the availability to humans of (non-nutritive) saccharin as a sweetener. The genes can just build reinforcement systems the activation of which is generally likely to increase the fitness of the genes specifying the reinforcer (or may have increased their fitness in the recent past).) In this sense, obtaining reinforcers may have life-threatening 'meaning' for animals, though of course the use of the word 'meaning' here does not imply any subjective state, just that the animal is built as a survival for reproduction machine. This is a novel, Darwinian, approach to the issue of symbol grounding.

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, and obstructing movement if they are in the way (see Section 3.2).

This section thus adds to Section 3.2 on a higher-order syntactic thought (HOST) theory of consciousness, by addressing the sense in which the thoughts may need to be grounded in the world. The HOST theory holds that the thoughts 'mean' something to the individual, in the sense that they may be about the survival of the individual (the phenotype) in the world, which the rational, thought, system aims to maximize [Rolls 2012c].

3.4 Other related approaches to consciousness

Some ways in which the current theory may be different from other related theories [Rosenthal 2004], [Gennaro 2004], [Carruthers 2000] follow.

The current theory holds that it is higher-order syntactic thoughts, HOSTs, [Rolls 1997a, 2004, 2006, 2007a,b, 2008a, 2010b, 2011] that are closely associated with consciousness, and this might differ from Rosenthal's higher-order thoughts (HOTs) theory [Rosenthal 1986, 1990, 1993, 2004, 2005] in the emphasis in the current theory on language. Language in the current theory is defined by syntactic manipulation of symbols, and does not necessarily imply verbal (or natural) language. The reason that strong emphasis is placed on language is that it is as a result of having a multistep, flexible, 'one-off', reasoning procedure that errors can be corrected by using 'thoughts about thoughts'. This enables correction of errors that cannot be easily corrected by reward or punishment received at the end of the reasoning, due to the credit assignment problem. That is, there is a need for some type of supervisory and monitoring process, to detect where errors in the reasoning have occurred. It is having such a HOST brain system, and it becoming engaged (even if only a little), that according to the HOST theory is associated with phenomenal consciousness.

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 is 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. 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., networks trained by back-propagation of errors or by reinforcement learning) mean that after training they operate according to fixed rules, and are often quite impenetrable and inflexible [Rumelhart, Hinton & Williams 1986], [Rolls 2008b]. 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 that is most

likely to be in error can be identified. This may be part of the reason why there is a close relationship 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 that 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 one-off linguistic thoughts with symbols grounded in the real world (as described at the end of Section 3.3), then prima facie this process would 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. [This is a functionalist position. Apparently Damasio does not subscribe to this view, for he suggests that there is something in the 'stuff' (the 'natural medium') that the brain is made of that is also important [Damasio 2003]. It is difficult for a person with this view to make telling points about consciousness from neuroscience, for it may always be the 'stuff' that is actually important.] 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 render it difficult to implement more than a few syntactic bindings of symbols simultaneously [Rolls & Treves 1998], [McLeod, Plunkett & Rolls 1998], [Rolls 2008b]. 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 relationship between, and the limited capacity of, both the stream of consciousness, and auditory–verbal short-term working memory, may be that both implement the capacity for syntax in neural networks.

The suggestion that syntax in real neuronal networks is implemented by temporal binding [Malsburg 1990], [Singer 1999] seems unlikely [Rolls 2008b], [Deco & Rolls 2011], [Rolls & Treves 2011]. (For example, the code about which visual stimulus has been shown can be read off from the end of the visual system without taking the temporal aspects of the neuronal firing into account; 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 [Tovee, Rolls, Treves *et al.* 1993], [Rolls & Tovee 1994], [Tovee & Rolls 1995], [Rolls & Treves 1998], [Rolls & Deco 2002], [Rolls 2003], [Rolls 2006] (these are rather short time-windows for the expression of multiple separate populations of synchronized neurons); and stimulus-dependent synchronization of firing between neurons is not a quantitatively important way of encoding information in the primate temporal cortical visual areas involved in the representation of objects and faces [Tovee & Rolls 1992], [Rolls & Treves 1998], [Rolls & Deco 2002], [Rolls, Franco, Aggelopoulos *et al.* 2003], [Rolls, Aggelopoulos, Franco *et al.* 2004], [Franco, Rolls, Aggelopoulos *et al.* 2004], [Aggelopoulos, Franco & Rolls 2005], [Rolls 2008b], [Deco & Rolls 2011], [Rolls & Treves 2011].)

However, the hypothesis that syntactic binding is necessary for consciousness is one of the postulates of the theory I am describing (for the system I describe must be capable of correcting its own syntactic thoughts). The fact that the binding must be implemented in neuronal networks may well place limitations on consciousness that lead to some of its properties, such as its unitary nature. The postulate of [Crick & Koch 1990] that oscillations and synchronization are necessary bases of consciousness could thus be related to the present theory if it turns out that oscillations or neuronal synchronization are the way the brain implements syntactic binding. However, the fact that oscillations and neuronal synchronization are especially evident in anaesthetized cats does not impress as strong evidence that oscillations and synchronization are critical features of consciousness, for most people would hold that anaesthetized cats are not conscious. The fact that oscillations and stimulus-dependent neuronal synchronization are much more difficult to demonstrate in the temporal cortical visual areas of awake behaving monkeys [Tovee & Rolls 1992], [Franco, Rolls, Aggelopoulos *et al.* 2004], [Aggelopoulos, Franco & Rolls 2005], [Rolls 2008b], [Rolls & Treves 2011] might just mean that during the evolution of primates the cortex has become better able to avoid parasitic oscillations, as a result of developing better feedforward and feedback inhibitory circuits [Rolls 2008b].

The theory [Rolls 1997a, 2004, 2006, 2007a,b, 2008a, 2010b, 2011] 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 might 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 discussion of the *causal role of consciousness* (Section 3.2.4). Does consciousness cause our behaviour? The view that I currently hold is that the information processing that 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. 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, which may be useful in correcting plans.

The hypothesis that it does feel like something when this processing is taking 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, for most humans, to be necessarily conscious.

Finally, I provide a short specification of what might have to be implemented in a neuronal network to implement conscious processing. First, a linguistic system, not necessarily verbal, but implementing syntax between symbols implemented in the environment would be needed. This system would be necessary for a multi-step one-off planning system. Then a higher-order thought system also implementing syntax and able to think about the representations in the first-order linguistic system, and able to correct the reasoning in the first-order linguistic system in a flexible manner, would be needed. The system would also need to have its representations grounded in the world, as discussed in Section 3.3. So my view is that consciousness can be implemented in neuronal networks (and that this is a topic worth discussing), but that the neuronal networks would have to implement the type of higher-order linguistic processing described in this paper, and also would need to be grounded in the world.

3.5 Monitoring and consciousness

An attractor network in the brain with positive feedback implemented by excitatory recurrent collateral connections between the neurons can implement decision-making [Wang 2002], [Deco & Rolls 2006], [Wang 2008], [Rolls & Deco 2010], [Deco, Rolls, Albantakis *et al.* 2012]. As explained in detail elsewhere [Rolls & Deco 2010], if the external evidence for the decision is consistent with the decision taken (which has been influenced by the noisy neuronal firing times), then the firing rates in the winning attractor are supported by the external evidence, and become especially high. If the external evidence is contrary to the noise-influenced decision, then the firing rates of the neurons in the winning attractor are not supported by the external evidence, and are lower than expected. In this way the confidence in a decision is reflected in, and encoded by, the firing rates of the neurons in the winning attractor population of neurons [Rolls & Deco 2010].

If we now add a second attractor network to read the firing rates from the first decision-making network, the second attractor network can take a decision based on the confidence expressed in the firing rates in the first network [Insabato, Pannunzi, Rolls *et al.* 2010]. The second attractor network allows decisions to be made about whether to change the decision made by the first network, and for example abort the trial or strategy (see Fig. 3). The second network, the confidence decision network, is in effect monitoring the decisions taken by the first network, and can cause a change of strategy or behaviour if the assessment of the decision taken by the first network does not seem a confident decision. This is described in detail elsewhere [Insabato, Pannunzi, Rolls *et al.* 2010], [Rolls & Deco 2010], but Fig. 3 shows the simple system of two attractor networks that enables confidence-based (second-level) decisions to be made, by monitoring the output of the first, decision-making, network.

Now this is the type of description, and language used, to describe 'monitoring' functions, taken to be a high-level cognitive process, possibly related to consciousness [Block 1995a], [Lycan 1997]. For example, in an experiment performed by Hampton [Hampton 2001] (experiment 3), a monkey had to remember a picture over a delay. He was then given a choice of a 'test flag', in which case he would be allowed to choose from one of four pictures the one seen before the delay, and if correct earn a large reward (a peanut). If he was not sure that he remembered the first picture, he could choose an 'escape flag', to start another trial. With longer delays, when memory strength might be lower partly due to noise in the system, and confidence therefore in the memory on some trials might be lower, the monkey was more likely to choose the escape flag. The experiment is described as showing that the monkey is thinking about his own memory, that is, is a case of meta-memory, which may be related to consciousness [Heyes 2008]. However, the decision about whether to escape from a trial can be taken just by adding a second decision network to the first decision network. Thus we can account for what seem like complex

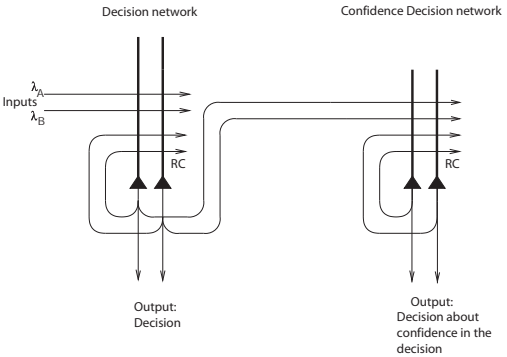

FIGURE 3: Network architecture for decisions about confidence estimates. The first network is a decision-making network, and its outputs are sent to a second network that makes decisions based on the firing rates from the first network, which reflect the decision confidence. In the first network, high firing of neuronal population (or pool) DA represents decision A, and high firing of population DB represents decision B. Pools DA and DB receive a stimulus-related input (respectively λ_A and λ_B), the evidence for each of the decisions, and these bias the attractor networks, which have internal positive feedback produced by the recurrent excitatory connections (RC). Pools DA and DB compete through inhibitory interneurons. The neurons are integrate-and-fire spiking neurons with random spiking times (for a given mean firing rate) which introduce noise into the network and influence the decision-making, making it probabilistic. The second network is a confidence decision attractor network, and receives inputs from the first network. The confidence decision network has two selective pools of neurons, one of which (C) responds to represent confidence in the decision, and the other of which responds when there is little or a lack of confidence in the decision (LC). The C neurons receive the outputs from the selective pools of the (first) decision-making network, and the LC neurons receive $\lambda_{\text{Reference}}$ which is from the same source but saturates at 40 spikes/s, a rate that is close to the rates averaged across correct and error trials of the sum of the firing in the selective pools in the (first) decision-making network. (After [Insabato, Pannunzi, Rolls *et al.* 2010].)

cognitive phenomena with a simple system of two attractor decision-making networks (Fig. 3) [Rolls & Deco 2010].

The implication is that some types of 'self-monitoring' can be accounted for by simple, two attractor network, computational processes. But what of more complex 'self-monitoring', such as is described as occurring in a commentary that might be based on reflection on previous events, and appears to be closely related to consciousness [Weiskrantz 1997]. This approach has been developed into my higher-order syntactic theory (HOST) of consciousness (Section 3.2 [Rolls 1997a, 2004, 2005, 2007a, 2008a, 2007b, 2010b, 2011]), in which there is a credit assignment problem if a multi-step reasoned plan fails, and it may be unclear which step failed. Such plans are described as syntactic as there

are symbols at each stage that must be linked together with the syntactic relationships between the symbols specified, but kept separate across stages of the plan. It is suggested that in this situation being able to have higher-order syntactic thoughts will enable one to think and reason about the first-order plan, and detect which steps are likely to be at fault.

Now this type of 'self-monitoring' is much more complex, as it requires syntax. The thrust of the argument is that some types of 'self-monitoring' are computationally simple, for example in decisions made based on confidence in a first decision [Rolls & Deco 2010], and may have little to do with consciousness; whereas higher-order thought processes are very different in terms of the type of syntactic computation required, and may be more closely related to consciousness [Rolls 1997a, 2003, 2004, 2005, 2007a, 2008a, 2007b, 2010b, 2011].

3.6 Conclusions on consciousness, and comparisons

It is suggested that it feels like something to be an organism or machine that can think about its own (syntactic and semantically grounded) thoughts.

It is suggested that qualia, raw sensory, and emotional, 'feels', arise secondarily to having evolved such a higher-order thought system, and that sensory and emotional processing feels like something because once this emotional processing has entered the planning, higher-order thought, system, it would be unparsimonious for it not to feel like something, given that all the other processing in this system I suggest does feel like something.

The adaptive value of having sensory and emotional feelings, or qualia, is thus suggested to be that such inputs are important to the long-term planning, explicit, processing system. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution.

Reasons why the ventral visual system is more closely related to explicit than implicit processing include the fact that representations of objects and individuals need to enter the planning, hence conscious, system, and are considered in more detail by [Rolls 2003] and by [Rolls 2008b].

Evidence that explicit, conscious, processing may have a higher threshold in sensory processing than implicit processing is considered by [Rolls 2003] and [Rolls 2006], based on neurophysiological and psychophysical investigations of backward masking [Rolls & Tovee 1994], [Rolls, Tovee, Purcell *et al.* 1994b, Rolls, Tovee & Panzeri 1999], [Rolls 2003, 2006]. It is suggested there that part of the adaptive value of this is that if linguistic processing is inherently serial and slow, it may be maladaptive to interrupt it unless there is a high probability that the interrupting signal does not arise from noise in the system. In the psychophysical and neurophysiological studies, it was found that face stimuli presented for 16 ms and followed immediately by a masking stimulus were not consciously perceived by humans, yet produced above chance

identification, and firing of inferior temporal cortex neurons in macaques for approximately 30 ms. If the mask was delayed for 20 ms, the neurons fired for approximately 50 ms, and the test face stimuli were more likely to be perceived consciously. In a similar backward masking paradigm, it was found that happy vs. angry face expressions could influence how much beverage was wanted and consumed even when the faces were not consciously perceived [Winkielman & Berridge 2005], [Winkielman & Berridge 2003]. This is further evidence that unconscious emotional stimuli can influence behaviour.

The theory is different from some other higher-order theories of consciousness [Rosenthal 1990, 1993, 2004], [Carruthers 2000], [Gennaro 2004] in that it provides an account of the evolutionary, adaptive, value of a higher-order thought system in helping to solve a credit assignment problem that arises in a multistep syntactic plan, links this type of processing to consciousness, and therefore emphasizes a role for syntactic processing in consciousness.

The theory described here is also different from other theories of consciousness and affect. James and Lange [James 1884], [Lange 1885] held that emotional feelings arise when feedback from the periphery (about for example heart rate) reach the brain, but had no theory of why some stimuli and not others produced the peripheral changes, and thus of why some but not other events produce emotional feelings.

Moreover, the evidence that feedback from peripheral autonomic and proprioceptive systems is essential for emotions is very weak, in that for example blocking peripheral feedback does not eliminate emotions, and producing peripheral, e.g., autonomic, changes does not elicit emotion [Reisenzein 1983], [Schachter & Singer 1962], [Rolls 2005].

Damasio's theory of emotion [Damasio 1994, 2003] is a similar theory to the James–Lange theory (and is therefore subject to some of the same objections), but holds that the peripheral feedback is used in decision-making rather than in consciousness. He does not formally define emotions, but holds that body maps and representations are the basis of emotions. When considering consciousness, he assumes that all consciousness is self-consciousness [Damasio 2003, 184], and that the foundational images in the stream of the mind are images of some kind of body event, whether the event happens in the depth of the body or in some specialized sensory device near its periphery [Damasio 2003, 197] . His theory does not appear to be a fully testable theory, in that he suspects that "the ultimate quality of feelings, a part of why feelings feel the way they feel, is conferred by the neural medium" [Damasio 2003, 131]. Thus presumably if processes he discusses [Damasio 1994, 2003] were implemented in a computer, then the computer would not have all the same properties with respect to consciousness as the real brain. In this sense he appears to be arguing for a non-functionalist position, and something crucial about consciousness being related to the particular biological machinery from which the system is made. In this respect the theory seems somewhat intangible.

LeDoux's approach to emotion [LeDoux 1992, 1995, 1996] is largely (to quote him) one of automaticity, with emphasis on brain mechanisms involved in the rapid, subcortical, mechanisms involved in fear. LeDoux, in line with [Johnson-Laird 1988] and [Baars 1988], emphasizes the role of working memory in consciousness, where he views working memory as a limited-capacity serial processor that creates and manipulates symbolic representations [LeDoux 1996, 280]. He thus holds that much emotional processing is unconscious, and that when it becomes conscious it is because emotional information is entered into a working memory system. However, LeDoux concedes that consciousness, especially its phenomenal or subjective nature, is not completely explained by the computational processes that underlie working memory [LeDoux 1996, 281].

Panksepp's approach to emotion has its origins in neuroethological investigations of brainstem systems that when activated lead to behaviours like fixed action patterns, including escape, flight and fear behaviour [Panksepp 1998]. His views about consciousness include the postulate that "feelings may emerge when endogenous sensory and emotional systems within the brain that receive direct inputs from the outside world as well as the neurodynamics of the SELF (a Simple Ego-type Life Form) begin to reverberate with each other's changing neuronal firing rhythms" [Panksepp 1998, 309].

Thus the theory of consciousness described in this paper is different from some other theories of consciousness.

4 Determinism

There are a number of senses in which our behaviour might be deterministic. One sense might be genetic determinism, and we have already seen that there are far too few genes to determine the structure and function of our brains, and thus to determine our behaviour [Rolls 2012c]. Moreover, development, and the environment with the opportunities it provides for brain self-organization and learning, play a large part in brain structure and function, and thus in our behaviour.

Another sense might be that if there were random factors that influence the operation of the brain, then our behaviour might be thought not to be completely predictable and deterministic. It is this that I consider here, a topic developed in *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function* [Rolls & Deco 2010], in which we show that there is noise or randomness in the brain, and argue that this can be advantageous.

Neurons emit action potentials, voltage spikes, which transmit information along axons to other neurons. These all-or-none spikes are a safe way to transmit information along axons, for they do not lose amplitude and degrade along a long axon. In most brain systems, an increase in the firing rate of the spikes carries the information. For example, taste neurons in the taste cortex

fire faster if the particular taste to which they respond is present, and neurons in the inferior temporal visual cortex fire faster if for example one of the faces to which they are tuned is seen [Rolls 2005, 2008b]. However, for a given mean firing rate (e.g., 50 spikes/s), the exact timing of each spike is quite random, and indeed is close to a Poisson distribution which is what is expected for a random process in which the timing of each spike is independent of the other spikes. Part of the neuronal basis of this randomness of the spike firing times is that each cortical neuron is held close to its threshold for firing and even produces occasional spontaneous firing, so that when an input is received, some at least of the cortical neurons will be so close to threshold that they emit a spike very rapidly, allowing information processing to be rapid [Rolls 2008b], [Rolls & Deco 2010].

This randomness in the firing time of individual neurons results in probabilistic behaviour of the brain [Rolls & Deco 2010]. For example, in decision-making, if the population of neurons that represents decision 1 has by chance more randomly occurring spikes in a short time, that population may win the competition (implemented through inhibitory interneurons) with a different population of neurons that represents decision 2. Decision-making is by this mechanism probabilistic. For example, if the odds are equal for decision 1 and decision 2, each decision will be taken probabilistically on 50% of the occasions or trials. This is highly adaptive, and is much better than getting stuck between two equally attractive rewards and unable to make a decision, as in the medieval tale of Duns Scotus about the donkey who starved because it could not choose between two equally attractive foods [Rolls & Deco 2010].

However, given that the brain operates with some degree of randomness due to the statistical fluctuations produced by the random spiking times of neurons, brain function is to some extent non-deterministic, as defined in terms of these statistical fluctuations. That is, the behaviour of the system, and of the individual, can vary from trial to trial based on these statistical fluctuations, in ways that are described by [Rolls & Deco 2010]. Indeed, given that each neuron has this randomness, and that there are sufficiently small numbers of synapses on the neurons in each network (between a few thousand and 20,000) that these statistical fluctuations are not smoothed out, and that there are a number of different networks involved in typical thoughts and actions each one of which may behave probabilistically, and with 10^{11} neurons in the brain each with this number of synapses, the system has so many degrees of freedom that it operates effectively as a non-deterministic system. (Philosophers may wish to argue about different senses of the term deterministic, but it is being used here in a precise, scientific, and quantitative way, which has been clearly defined.)

5 Free will

Do we have free will when we make a choice? Given the distinction made between the implicit system that seeks for gene-specified rewards, and the explicit system that can use reasoning to defer an immediate goal and plan many steps ahead for longer-term goals [Rolls 2012c], do we have free will when both the implicit and the explicit systems have made the choice?

Free will would in Rolls' view [Rolls 2005, 2008a,b, 2010b, 2011] 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 (in the implicit systems), the conscious processor (the explicit system) 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 that can take overall control using conscious verbal processing to believe that it was overridden by another system. This may be the underlying computational reason why confabulation occurs [Rolls 2012c].

The interesting view we are led to is thus that when probabilistic choices influenced by stochastic dynamics [Rolls & Deco 2010] are made between the implicit and explicit systems, we may not be aware of which system made the choice. Further, when the stochastic noise has made us choose with the implicit system, we may confabulate and say that we made the choice of our own free will, and provide a guess at why the decision was taken. In this scenario, the stochastic dynamics of the brain plays a role even in how we understand free will [Rolls 2010b].

The implication of this argument is that a good use of the term free will is when the term refers to the operation of the rational, planning, explicit (conscious) system that can think many moves ahead, and choose from a number of such computations the multistep strategy that best optimizes the goals of the explicit system with long-term goals. When on the other hand our implicit system has taken a decision, and we confabulate a spurious account with our explicit system, and pronounce that we took the decision for such

and such a (confabulated) reason of our own "free will", then my view is that the feeling of free will was an illusion [Rolls 2005, 2010b].

Bibliography

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

Aggelopoulos, N. C., Franco, L. & Rolls, E. T.
2005 Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons, *Journal of Neurophysiology*, 93, 1342–1357.

Allport, A.
1988 What concept of consciousness?, in *Consciousness in Contemporary Science*, edited by Marcel, A. J. & Bisiach, E., Oxford: Oxford University Press, 159–182.

Anderson, J. R.
1996 ACT: a simple theory of complex cognition, *American Psychologist*, 51, 355–365.

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

Baars, B. J.
1988 *A Cognitive Theory of Consciousness*, New York: Cambridge University Press.

Barlow, H.
1997 Single neurons, communal goals, and consciousness, in *Cognition, Computation, and Consciousness*, edited by Miyashita, Y. & Rolls, E. T., Oxford: Oxford University Press, chap. 7, 121–136.

Block, N.
1995a On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18, 22–47.

1995b Two neural correlates of consciousness, *Trends in Cognitive Sciences*, 9, 46–52.

Booth, M. C. A. & Rolls, E. T.
1998 View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex, *Cerebral Cortex*, 8, 510–523.

Brooks, S. J., Savov, V., Allzen, E., Benedict, C., Fredriksson, R. & Schioth, H. B.
2012 Exposure to subliminal arousing stimuli induces robust activation

in the amygdala, hippocampus, anterior cingulate, insular cortex and primary visual cortex: a systematic meta-analysis of fMRI studies, *Neuroimage*, 59, 2962–2673.

Byrne, R. W. & Whiten, A.
1988 *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Oxford: Clarendon Press.

Carruthers, P.
1996 *Language, Thought and Consciousness*, Cambridge: Cambridge University Press.
2000 *Phenomenal Consciousness*, Cambridge: Cambridge University Press.

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

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

Crick, F. H. C. & Koch, C.
1990 Towards a neurobiological theory of consciousness, *Seminar Neuroscience*, 2, 263–275.

Damasio, A. R.
1994 *Descartes' Error*, New York: Putnam.
2003 *Looking for Spinoza*, London: Heinemann.

Deco, G. & Rolls, E. T.
2004 A neurodynamical cortical model of visual attention and invariant object recognition, *Vision Research*, 44, 621–644.
2006 A neurophysiological model of decision-making and Weber's law, *European Journal of Neuroscience*, 24, 901–916.
2011 Reconciling oscillations and firing rates.

Deco, G., Rolls, E. T., Albantakis, L. & Romo, R.
2012 Brain mechanisms for perceptual and reward-related decision-making, *Progress in Neurobiology*, 2, Epub, doi:http://dx.doi.org/10.1016/j.pneurobio.2012.01.010.

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

Elliffe, M. C. M., Rolls, E. T. & Stringer, S. M.
2002 Invariant recognition of feature combinations in the visual system, *Biological Cybernetics*, 86, 59–71.

Fodor, J. A.
1994 *The Elm and the Expert: Mentalese and its Semantics*, Cambridge, MA: MIT Press.

Franco, L., Rolls, E. T., Aggelopoulos, N. C. & Treves, A.
2004 The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons, *Experimental Brain Research*, 155, 370–384.

Gazzaniga, M. S.
1988 Brain modularity: towards a philosophy of conscious experience, in *Consciousness in Contemporary Science*, edited by Marcel, A. J. & Bisiach, E., Oxford: Oxford University Press, chap. 10, 218–238.

1995 Consciousness and the cerebral hemispheres, in *The Cognitive Neurosciences*, edited by Gazzaniga, M. S., Cambridge, MA: MIT Press, chap. 92, 1392–1400.

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

Gennaro, R. J.
2004 *Higher Order Theories of Consciousness*, Amsterdam: John Benjamins.

Grabenhorst, F. & Rolls, E. T.
2011 Value, pleasure, and choice systems in the ventral prefrontal cortex, *Trends in Cognitive Sciences*, 15, 56–67.

Hampton, R. R.
2001 Rhesus monkeys know when they can remember, *Proceedings of the National Academy of Sciences of the USA*, 98, 5359–5362.

Hertz, J., Krogh, A. & Palmer, R. G.
1991 *Introduction to the Theory of Neural Computation*, Wokingham: Addison Wesley.

Heyes, C.
2008 Beast machines? Questions of animal consciousness, in *Frontiers of Consciousness*, edited by Weiskrantz, L. & Davies, M., Oxford: Oxford University Press, chap. 9, 259–274.

Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R. & Polkey, C. E.
2003 Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices, *Brain*, 126, 1691–1712.

Humphrey, N. K.
1980 Nature's psychologists, in *Consciousness and the Physical World*, edited by Josephson, B. D. & Ramachandran, V. S., Oxford: Pergamon, 57–80.

1986 *The Inner Eye*, London: Faber.

Insabato, A., Pannunzi, M., Rolls, E. T. & Deco, G.
2010 Confidence-related decision-making, *Journal of Neurophysiology*, 104, 539–547.

James, W.
1884 What is an emotion?, *Mind*, 9, 188–205.

Johnson-Laird, P. N.
1988 *The Computer and the Mind: An Introduction to Cognitive Science*, Cambridge, MA: Harvard University Press.

Kadohisa, M., Rolls, E. T. & Verhagen, J. V.
2005 Neuronal representations of stimuli in the mouth: the primate insular taste cortex, orbitofrontal cortex, and amygdala, *Chemical Senses*, 30, 401–419.

Koch, C.
2004 *The Quest for Consciousness*, Englewood, CO: Roberts.

Kohonen, T.
1989 *Self–Organization and Associative Memory*, Berlin: Springer-Verlag, 3rd ed.

Lange, C.
1885 The emotions, in *The Emotions*, edited by Dunlap, E., Baltimore: Williams and Wilkins, 1922 edn.

LeDoux, J. E.
1992 Emotion and the amygdala, in *The Amygdala*, edited by Aggleton, J. P., New York: Wiley-Liss, chap. 12, 339–351.

1995 Emotion: clues from the brain, *Annual Review of Psychology*, 46, 209–235.

1996 *The Emotional Brain*, New York: Simon and Schuster.

Lycan, W. G.
1997 Consciousness as internal monitoring, in *The Nature of Consciousness: Philosophical Debates*, edited by Block, N., Flanagan, O. & Guzeldere, G., Cambridge, MA: MIT Press, 755–771.

Malsburg, C. v. d.
1990 A neural architecture for the representation of scenes, in *Brain Organization and Memory: Cells, Systems and Circuits*, edited by McGaugh, J. L., Weinberger, N. M. & Lynch, G., New York: Oxford University Press, chap. 19, 356–372.

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

Panksepp, J.
1998 *Affective Neuroscience: The Foundations of Human and Animal Emotions*, New York: Oxford University Press.

Prabhakaran, R. & Gray, J. R.
2012 The pervasive nature of unconscious social information processing in executive control, *Frontiers in Human Neuroscience*, 6, 105.

Reisenzein, R.
1983 The Schachter theory of emotion: two decades later, *Psychological Bulletin*, 94, 239–264.

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

1990 A theory of emotion, and its application to understanding the neural basis of emotion, *Cognition and Emotion*, 4, 161–190.

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

1994 Brain mechanisms for invariant visual recognition and learning, *Behavioural Processes*, 33, 113–138.

1995 A theory of emotion and consciousness, and its application to understanding the neural basis of emotion, in *The Cognitive Neurosciences*, edited by Gazzaniga, M. S., Cambridge, MA: MIT Press, chap. 72, 1091–1106.

1997a Consciousness in neural networks?, *Neural Networks*, 10, 1227–1240.

1997b Taste and olfactory processing in the brain and its relation to the control of eating, *Critical Reviews in Neurobiology*, 11, 263–287.

2000 Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition, *Neuron*, 27, 205–218.

2003 Consciousness absent and present: a neurophysiological exploration, *Progress in Brain Research*, 144, 95–106.

2004 A higher order syntactic thought (HOST) theory of consciousness, in *Higher Order Theories of Consciousness*, edited by Gennaro, R. J., Amsterdam: John Benjamins, chap. 7, 137–172.

2005 *Emotion Explained*, Oxford: Oxford University Press.

2006 Consciousness absent and present: a neurophysiological exploration of masking, in *The First Half Second*, edited by Ogmen, H. & Breitmeyer, B. G., Cambridge, MA: MIT Press, chap. 6, 89–108.

2007a The affective neuroscience of consciousness: higher order syntactic thoughts, dual routes to emotion and action, and conscious-

ness, in *Cambridge Handbook of Consciousness*, edited by Zelazo, P. D., Moscovitch, M. & Thompson, E., New York: Cambridge University Press, chap. 29, 831–859.

2007b A computational neuroscience approach to consciousness, *Neural Networks*, 20, 962–982.

2008a Emotion, higher order syntactic thoughts, and consciousness, in *Frontiers of Consciousness*, edited by Weiskrantz, L. & Davies, M., Oxford: Oxford University Press, chap. 4, 131–167.

2008b *Memory, Attention, and Decision-Making. A Unifying Computational Neuroscience Approach*, Oxford: Oxford University Press.

2009 The anterior and midcingulate cortices and reward, in *Cingulate Neurobiology and Disease*, edited by Vogt, B., Oxford: Oxford University Press, chap. 8, 191–206.

2010a A computational theory of episodic memory formation in the hippocampus, *Behavioural Brain Research*, 215, 180–196.

2010b Noise in the brain, decision-making, determinism, free will, and consciousness, in *New Horizons in the Neuroscience of Consciousness*, edited by Perry, E., Collerton, D., Ashton, H. & Lebeau, F., Amsterdam: John Benjamins, 113–120.

2011 Consciousness, decision-making, and neural computation, in *Perception-Action Cycle: Models, architecture, and hardware*, edited by Cutsuridis, V., Hussain, A. & Taylor, J. G., Berlin: Springer, chap. 9, 287–333.

2012a Glutamate, obsessive-compulsive disorder, schizophrenia, and the stability of cortical attractor neuronal networks, *Pharmacology, Biochemistry and Behavior*, 100, 736–751.

2012b Invariant visual object and face recognition: neural and computational bases, and a model, VisNet, *Frontiers in Computational Neuroscience*, 6(35), 1–70.

2012c *Neuroculture: On the Implications of Brain Science*, Oxford: Oxford University Press.

2013 Central neural integration of taste, smell and other sensory modalities, in *Handbook of Olfaction and Gustation: Modern Perspectives*, edited by Doty, R. L., New York: Dekker, chap. 44, 3rd ed.

2014 *Emotion and Decision-Making Explained*, Oxford: Oxford University Press.

Rolls, E. T. & Deco, G.

2002 *Computational Neuroscience of Vision*, Oxford: Oxford University Press.

2010 *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function*, Oxford: Oxford University Press.

Rolls, E. T. & Grabenhorst, F.
2008 The orbitofrontal cortex and beyond: from affect to decision-making, *Progress in Neurobiology*, 86, 216–244.

Rolls, E. T. & Milward, T.
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. & Stringer, S. M.
2001 Invariant object recognition in the visual system with error correction and temporal difference learning, *Network: Computation in Neural Systems*, 12, 111–129.

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

Rolls, E. T. & Treves, A.
1998 *Neural Networks and Brain Function*, Oxford: Oxford University Press.
2011 The neuronal encoding of information in the brain, *Progress in Neurobiology*, 95, 448–490.

Rolls, E. T., Critchley, H. D. & Treves, A.
1996 The representation of olfactory information in the primate orbitofrontal cortex, *Journal of Neurophysiology*, 75, 1982–1996.

Rolls, E. T., Tovee, M. J. & Panzeri, S.
1999 The neurophysiology of backward visual masking: information analysis, *Journal of Cognitive Neuroscience*, 11, 335–346.

Rolls, E. T., Treves, A. & Tovee, M. J.
1997 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., Aggelopoulos, N. C., Franco, L. & Treves, A.
2004 Information encoding in the inferior temporal visual cortex: contributions of the firing rates and the correlations between the firing of neurons, *Biological Cybernetics*, 90, 19–32.

Rolls, E. T., Franco, L., Aggelopoulos, N. C. & Reece, S.
2003 An information theoretic approach to the contributions of the firing rates and the correlations between the firing of neurons, *Journal of Neurophysiology*, 89, 2810–2822.

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., Treves, A., Robertson, R. G., Georges-François, P. & Panzeri, S.
1998 Information about spatial view in an ensemble of primate hippocampal cells, *Journal of Neurophysiology*, 79, 1797–1813.

Rosenthal, D.
1990 A theory of consciousness, ZIF Report 40/1990, Zentrum für Interdisziplinaire Forschung, Bielefeld, Reprinted in Block, N., Flanagan, O. and Guzeldere, G. (eds.) (1997) *The Nature of Consciousness: Philosophical Debates*. MIT Press, Cambridge MA, 729–853.

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

1993 Thinking that one thinks, in *Consciousness*, edited by Davies, M. & Humphreys, G. W., Oxford: Blackwell, chap. 10, 197–223.

2004 Varieties of higher order theory, in *Higher Order Theories of Consciousness*, edited by Gennaro, R. J., Amsterdam: John Benjamins, 17–44.

2005 *Consciousness and Mind*, Oxford: Oxford University Press.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J.
1986 Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by Rumelhart, D. E., McClelland, J. L. & the PDP Research Group, Cambridge, MA: MIT Press, vol. 1, chap. 8, 318–362.

Schachter, S. & Singer, J.
1962 Cognitive, social and physiological determinants of emotional state, *Psychological Review*, 69, 378–399.

Singer, W.
1999 Neuronal synchrony: A versatile code for the definition of relations?, *Neuron*, 24, 49–65.

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

Squire, L. R., Stark, C. E. L. & Clark, R. E.
2004 The medial temporal lobe, *Annual Review of Neuroscience*, 27, 279–306.

Stringer, S. M. & Rolls, E. T.
2000 Position invariant recognition in the visual system with cluttered environments, *Neural Networks*, 13, 305–315.

2002 Invariant object recognition in the visual system with novel views of 3D objects, *Neural Computation*, 14, 2585–2596.

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.

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. & 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.
1994 A computational analysis of the role of the hippocampus in memory, *Hippocampus*, 4, 374–391.

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

Wang, X.-J.
2002 Probabilistic decision making by slow reverberation in cortical circuits, *Neuron*, 36, 955–968.

2008 Decision making in recurrent neuronal circuits, *Neuron*, 60, 215–234.

Weiskrantz, L.
1997 *Consciousness Lost and Found*, Oxford: Oxford University Press.

Whiten, A. & Byrne, R. W.
1997 *Machiavellian Intelligence II: Extensions and Evaluations*, Cambridge: Cambridge University Press.

Winkielman, P. & Berridge, K. C.
2003 What is an unconscious emotion?, *Cognition and Emotion*, 17, 181–211.

2005 Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value, *Personality and Social Psychology Bulletin*, 31, 111–135.